

**QUANTITY ANALYSIS01: PRODUCTIVITY**  
**NUS: LECTURE NOTES**

ED&CS

May 31, 2015



# Contents

Chapter 1 Inequalities	1
1.1 Introduction . . . . .	1
1.2 The Cauchy-Schwarz Inequality . . . . .	1
1.3 The Theorem of the Arithmetic and Geometric Mean . . . . .	4
1.4 Means of Order $r$ . . . . .	8
1.5 Schlömilch's Inequality . . . . .	10
1.6 L'Hospital's Rule and Logarithmic Means . . . . .	15
1.7 Additional Properties of Means of Order $r$ . . . . .	16
1.8 Summary of Methods used to Establish Inequalities . . . . .	18
1.9 References . . . . .	19
Chapter 2 Convex Sets and Concave Functions	21
2.1 Introduction . . . . .	21
2.2 Convex Sets . . . . .	21
2.3 The Supporting Hyperplane Theorem for Closed Convex Sets . . . . .	24
2.4 Concave Functions . . . . .	28
2.5 Convex Functions . . . . .	36
2.6 Quasiconcave Functions . . . . .	38
2.7 Quasiconvex Functions . . . . .	45
2.8 References . . . . .	46
Chapter 3 Microeconomic Theory: A Dual Approach	47
3.1 Introduction . . . . .	47
3.2 Properties of Cost Functions . . . . .	47
3.3 The Determination of the Production Function from the Cost Function . . . . .	51
3.4 The Derivative Property of the Cost Function . . . . .	55
3.5 The Comparative Statics Properties of Input Demand Functions . . . . .	58
3.6 The Application of Cost Functions to Consumer Theory . . . . .	63
3.7 Flexible Functional Forms and Nonunitary Income Elasticities of Demand . . . . .	65
3.8 Money Metric Utility Scaling and Other Methods of Cardinalizing Utility . . . . .	68
3.9 Variable Profit Functions . . . . .	75
3.10 The Comparative Statics Properties of Net Supply and Fixed Input Demand Functions	82
3.11 Flexible Functional Forms for a Variable Profit Function . . . . .	89
3.12 References . . . . .	95
Chapter 4 Notes on the Construction of a Data Set for an O.E.C.D. Country	99
4.1 Overview . . . . .	99
4.2 Basic National Accounts Data . . . . .	101
4.3 Labour and Population Statistics . . . . .	102
4.4 Capital and Interest Rate Series . . . . .	104
4.5 Taxes and Tax Rates . . . . .	106

Chapter 5	Index Number Theory: Part I: Early Approaches	109
5.1	Index Number Purpose and Overview . . . . .	109
5.2	Setting the Stage and the Levels Approach to the Index Number Problem . . . . .	110
5.3	Fixed Basket Approaches to Bilateral Index Number Theory . . . . .	112
5.4	Stochastic and Descriptive Statistics Approaches to Index Number Theory . . . . .	114
5.5	Test Approaches to Index Number Theory . . . . .	116
5.6	Fixed Base versus Chained Indexes . . . . .	125
5.7	References . . . . .	129
Chapter 6	Index Number Theory: Part II: The Economic Approach	135
6.1	Introduction . . . . .	135
6.2	Konüs True Cost of Living Indexes . . . . .	136
6.3	The True Cost of Living Index when Preferences are Homothetic . . . . .	140
6.4	Wold's Identity and Shephard's Lemma . . . . .	142
6.5	Superlative Indexes I: The Fisher Ideal Index . . . . .	143
6.6	Superlative Indexes II: Quadratic Mean of Order $r$ Indexes . . . . .	147
6.7	Superlative Indexes III: Normalized Quadratic Indexes . . . . .	150
6.8	Nonhomothetic Preferences and Cost of Living Indexes . . . . .	154
6.9	Allen Quantity Indexes . . . . .	158
6.10	Conclusion . . . . .	160
6.11	References . . . . .	160
Chapter 7	The Measurement of Productivity	165
7.1	Introduction . . . . .	165
7.2	Productivity Measurement in the Case of One Input and One Output . . . . .	166
7.3	The Determinants of Economic Growth: Primary Input Growth and Other Factors . . . . .	171
7.4	The Determinants of Economic Growth: Productivity Growth . . . . .	173
7.5	Increasing Returns to Scale . . . . .	177
7.6	Other Factors that Might Explain Growth . . . . .	182
7.7	A Summary of the Factors Explaining Productivity Growth . . . . .	188
7.8	The Role of Government in Facilitating Growth . . . . .	189
7.9	The Index Number Approach to the Measurement of Productivity . . . . .	191
7.10	The Estimation of Technical Progress and Returns to Scale . . . . .	198
7.11	Can the Use of Instrumental Variables Lead to Better Estimates of Returns to Scale? . . . . .	207
7.12	References . . . . .	213
Chapter 8	The Measurement of Capital	219
8.1	Introduction . . . . .	219
8.2	Inflation, the Length of the Accounting Period and the Measurement of Economic Activity . . . . .	220
8.3	The Fundamental Equations Relating Stocks and Flows of Capital . . . . .	224
8.4	Cross Sectional Depreciation Profiles . . . . .	230
8.5	The Empirical Determination of Interest Rates and Asset Inflation Rates . . . . .	234
8.6	Aggregation over Vintages of a Capital Good . . . . .	237
8.7	The One Hoss Shay Model of Efficiency and Depreciation . . . . .	239
8.8	The Declining Balance or Geometric Depreciation Model . . . . .	240
8.9	The Straight Line Method of Depreciation . . . . .	245
8.10	The Linear Efficiency Decline Model . . . . .	246
8.11	Appendix 1: A Theoretical Treatment of Inventory Change . . . . .	247
8.12	Appendix 2: The Underlying Model of Production . . . . .	250
8.13	References . . . . .	252

Chapter 9	The Measurement of Income and the Determinants of Income Growth	257
9.1	Introduction . . . . .	257
9.2	Measuring National Product: Gross versus Net . . . . .	258
9.3	Measuring Income: Hicks versus Samuelson . . . . .	260
9.4	The Theory of the Output Index . . . . .	263
9.5	Maintaining Capital Again: the Physical versus Real Financial Perspectives . . . . .	265
9.6	Measuring Business Income: the End of the Period Perspective . . . . .	267
9.7	Approximations to the Income Concept . . . . .	273
9.8	Choosing an Income Concept: A Summary . . . . .	276
9.9	Productivity and Real Income Growth: A Theoretical Framework . . . . .	277
9.10	The Translog GDP Function Approach . . . . .	282
9.11	The Translog GDP Function Approach and Changes in the Terms of Trade . . . . .	284
9.12	References . . . . .	286
Chapter 10	Flexible Functional Forms	291
10.1	Introduction . . . . .	291
10.2	The Definition of a Flexible Functional Form . . . . .	292
10.3	The Generalized Leontief Cost Function . . . . .	296
10.4	The Translog Unit Cost Function . . . . .	297
10.5	The Normalized Quadratic Unit Cost Function . . . . .	302
10.6	The Estimation of Consumer Preferences: The General Framework . . . . .	306
10.7	The Generalized Leontief Cost Function for Homothetic Preferences . . . . .	307
10.8	The Normalized Quadratic Cost Function for Homothetic Preferences . . . . .	308
10.9	The Problem of Cardinalizing Utility . . . . .	309
10.10	Modeling Nonhomothetic Preferences . . . . .	310
10.11	The Use of Linear Spline Functions to Achieve Greater Flexibility . . . . .	311
10.12	The Estimation of Unit Profit Functions: The General Framework . . . . .	313
10.13	The Translog Variable Profit Function with Constant Returns to Scale . . . . .	313
10.14	The Translog Variable Profit Function with Nonconstant Returns to Scale . . . . .	316
10.15	The Normalized Quadratic Unit Profit Function Model . . . . .	319
10.16	The Normalized Quadratic Unit Profit Function Model with Curvature Imposed . . . . .	320
10.17	Use of Splines for Modeling Technical Progress . . . . .	321
10.18	Allowing for Flexibility at Two Sample Points . . . . .	323
10.19	Semiflexible Functional Forms . . . . .	324
10.20	References . . . . .	326
Chapter 11	Linear Programming	329
11.1	Introduction . . . . .	329
11.2	The Geometric Interpretation of a Linear Program in Activities Space . . . . .	330
11.3	The Simplex Algorithm for Solving Linear Programs . . . . .	331
11.4	An Example of the Simplex Algorithm . . . . .	335
11.5	Finding a Starting Basic Feasible Solution . . . . .	336
11.6	Nonsingularity of the Basis Matrix in the Simplex Algorithm . . . . .	339
11.7	The Degeneracy Problem . . . . .	340
11.8	The Dual Linear Program . . . . .	341
11.9	The Geometric Interpretation of a Linear Program in Requirements Space . . . . .	346
11.10	The Saddlepoint Criterion for Solving a Linear Program . . . . .	350
11.11	Programming with Variable Coefficients . . . . .	352
11.12	References . . . . .	362
Bibliography		365



# Chapter 1

## Inequalities

### 1.1 Introduction

Inequalities play an important role in many areas of economics. Unfortunately, this topic is not usually covered in the typical Mathematics for Economists course so we will give an introduction to this topic in this chapter, deriving the most important inequalities that are used in applied economics.

In section 1.2, we provide some proofs of the *Cauchy Schwarz Inequality* while section 1.3 provides a proof of the *Theorem of the Arithmetic and Geometric Means*.

Section 1.4 introduces the *mean of order  $r$* , which is a special case of the Constant Elasticity of Substitution (or CES) functional form for a utility or production function. Means of order  $r$  are required in order to state *Schlömilch's Inequality*, which is a generalization of the Theorem of the Arithmetic and Geometric Means. Schlömilch's Inequality will be proven in section 1.5.

Section 1.6 introduces a type of mean or average that plays a prominent role in index number theory: the *logarithmic mean* of two positive numbers.

Section 1.7 establishes a few more properties of the means of order  $r$ . In particular, we look at limiting cases as  $r$  tends to plus or minus infinity.

Finally, section 1.8 concludes with a brief summary of methods that are used to establish inequalities.

### 1.2 The Cauchy-Schwarz Inequality

**Proposition 1** *Cauchy* (1821; 373) - *Schwarz* (1885) *Inequality*

Let  $\mathbf{x}$  and  $\mathbf{y}$  be  $N$  dimensional vectors. Then<sup>\*1</sup>

$$(\mathbf{x}^T \mathbf{y})^2 \leq (\mathbf{x}^T \mathbf{x})(\mathbf{y}^T \mathbf{y}). \quad (1.1)$$

**Proof.** Define the  $N \times 2$  matrix  $\mathbf{A}$  as follows:

$$\mathbf{A} \equiv [\mathbf{x}, \mathbf{y}]. \quad (1.2)$$

Define the  $2 \times 2$  matrix  $\mathbf{B}$  as follows:

$$\mathbf{B} \equiv \mathbf{A}^T \mathbf{A} = [\mathbf{x}, \mathbf{y}]^T [\mathbf{x}, \mathbf{y}] = \begin{bmatrix} \mathbf{x}^T \mathbf{x} & \mathbf{x}^T \mathbf{y} \\ \mathbf{y}^T \mathbf{x} & \mathbf{y}^T \mathbf{y} \end{bmatrix}, \quad (1.3)$$

It is easily seen that  $\mathbf{B}$  is a positive semidefinite matrix, since

$$\mathbf{z}^T \mathbf{B} \mathbf{z} = \mathbf{z}^T \mathbf{A}^T \mathbf{A} \mathbf{z} = (\mathbf{A} \mathbf{z})^T (\mathbf{A} \mathbf{z}) = \mathbf{u}^T \mathbf{u} \geq 0 \quad (1.4)$$

---

<sup>\*1</sup> This proof may be found in Hardy, Littlewood and Polya (1934; 16)[213].

where  $\mathbf{z}^T \equiv [z_1, z_2]$  and the  $N$  dimensional vector  $\mathbf{u}$  is defined as  $\mathbf{A}\mathbf{z} = \mathbf{x}z_1 + \mathbf{y}z_2$ . The determinantal conditions for  $\mathbf{B}$  to be positive semidefinite imply:

$$0 \leq |\mathbf{B}| = \begin{vmatrix} \mathbf{x}^T \mathbf{x} & \mathbf{x}^T \mathbf{y} \\ \mathbf{y}^T \mathbf{x} & \mathbf{y}^T \mathbf{y} \end{vmatrix} = \mathbf{x}^T \mathbf{x} \mathbf{y}^T \mathbf{y} - (\mathbf{x}^T \mathbf{y})^2, \quad (1.5)$$

and (1.5) simplifies to (1.1). ■

Note that for (1.1) to be a strict inequality, (1.5) must be a strict inequality and hence  $\mathbf{B}$  must be positive definite. This in turn implies that we must have:

$$\mathbf{0}_N \neq \mathbf{u} = \mathbf{x}z_1 + \mathbf{y}z_2 \text{ for } (z_1, z_2) \neq (0, 0), \quad (1.6)$$

and (1.6) in turn implies that both  $\mathbf{x}$  and  $\mathbf{y}$  must be nonzero and nonproportional. Thus to obtain a strict inequality in (1.1), we cannot have  $\mathbf{x} = k\mathbf{y}$  or  $\mathbf{y} = k\mathbf{x}$  for any scalar  $k$ .

The problem below provides an alternative proof for (1.1). The problems below and the material in the following sections will provide many applications of the Cauchy-Schwarz inequality.

**Problem 1** Assume  $\mathbf{x} \neq \mathbf{0}_N$  and  $\mathbf{y} \neq \mathbf{0}_N$  (the inequality (1.1) is trivially true if either  $\mathbf{x}$  or  $\mathbf{y}$  equals  $\mathbf{0}_N$ ), and for each real number  $\lambda$ , define  $f(\lambda)$  as

$$f(\lambda) \equiv (\mathbf{x} + \lambda\mathbf{y})^T (\mathbf{x} + \lambda\mathbf{y}) = \lambda^2 \mathbf{y}^T \mathbf{y} + 2\lambda \mathbf{x}^T \mathbf{y} + \mathbf{x}^T \mathbf{x} \geq 0. \quad (i)$$

The inequality in (i) is true because  $(\mathbf{x} + \lambda\mathbf{y})^T (\mathbf{x} + \lambda\mathbf{y}) = \sum_{i=1}^N (x_i + \lambda y_i)^2$  is a sum of squares. Now use standard calculus techniques and minimize  $f(\lambda)$  with respect to  $\lambda$ . Let the minimizing  $\lambda$  be denoted as  $\lambda^*$ . Now calculate  $f(\lambda^*)$  and it will turn out that the inequality

$$f(\lambda^*) \geq 0 \quad (ii)$$

is equivalent to the Cauchy-Schwarz Inequality (1.1) above. (It is not necessary to check the second order conditions for the minimization problem associated with minimizing  $f(\lambda)$ .)

**Problem 2** Let  $\mathbf{Y}$  and  $\mathbf{X}$  be two  $N$  dimensional vectors; i.e., define  $\mathbf{Y}^T \equiv [Y_1, \dots, Y_N]$ ;  $\mathbf{X}^T \equiv [X_1, \dots, X_N]$ . Define the arithmetic mean of the  $Y_n$  and  $X_n$  as  $Y^* \equiv (1/N) \sum_{i=1}^N Y_n$  and  $X^* \equiv (1/N) \sum_{i=1}^N X_n$  respectively. Now define the vectors  $\mathbf{y}$  and  $\mathbf{x}$  as  $\mathbf{Y}$  and  $\mathbf{X}$  except we subtract the respective means from each vector; i.e., define:

$$\mathbf{y} \equiv \mathbf{Y} - Y^* \mathbf{1}_N; \quad \mathbf{x} \equiv \mathbf{X} - X^* \mathbf{1}_N \quad (i)$$

where  $\mathbf{1}_N$  is a vector of ones of dimension  $N$ . Now consider the following regression models of  $\mathbf{y}$  on  $\mathbf{x}$  and  $\mathbf{x}$  on  $\mathbf{y}$ :

$$\mathbf{y} = \alpha \mathbf{x} + \mathbf{u}; \quad (ii)$$

$$\mathbf{x} = \beta \mathbf{y} + \mathbf{v} \quad (iii)$$

where  $\mathbf{u}$  and  $\mathbf{v}$  are error vectors and  $\alpha$  and  $\beta$  are unknown parameters. We assume that  $\mathbf{x} \neq \mathbf{0}_N$  and  $\mathbf{y} \neq \mathbf{0}_N$ . The *least squares estimator* for  $\alpha$  is the  $\alpha^*$  which solves the unconstrained minimization problem:

$$\min_{\alpha} f(\alpha) \quad (iv)$$

where  $f$  is defined as

$$f(\alpha) \equiv \mathbf{u}^T \mathbf{u} = (\mathbf{y} - \alpha \mathbf{x})^T (\mathbf{y} - \alpha \mathbf{x}). \quad (v)$$

The *least squares estimator* for  $\beta$  is the  $\beta^*$  which solves the unconstrained minimization problem:

$$\min_{\beta} g(\beta) \quad (\text{iv})$$

where  $g$  is defined as

$$g(\beta) \equiv \mathbf{v}^T \mathbf{v} = (\mathbf{x} - \beta \mathbf{y})^T (\mathbf{x} - \beta \mathbf{y}). \quad (\text{v})$$

- (a) Find the least squares estimators for  $\alpha$  and  $\beta$ ,  $\alpha^*$  and  $\beta^*$ . Check the second order conditions for your solutions.

The *variances* for  $\mathbf{Y}$  and  $\mathbf{X}$  and the *covariance* between  $\mathbf{Y}$  and  $\mathbf{X}$  are defined as follows:

$$\text{Var}(\mathbf{Y}) \equiv \mathbf{y}^T \mathbf{y} / N; \quad \text{Var}(\mathbf{X}) \equiv \mathbf{x}^T \mathbf{x} / N; \quad \text{Cov}(\mathbf{Y}, \mathbf{X}) \equiv \mathbf{x}^T \mathbf{y} / N. \quad (\text{vi})$$

The *correlation coefficient*  $\rho$  between  $\mathbf{Y}$  and  $\mathbf{X}$  is defined as follows:

$$\rho \equiv \text{Cov}(\mathbf{Y}, \mathbf{X}) / [\text{Var}(\mathbf{Y}) \text{Var}(\mathbf{X})]^{1/2} = \mathbf{x}^T \mathbf{y} / (\mathbf{x}^T \mathbf{x})^{1/2} (\mathbf{y}^T \mathbf{y})^{1/2}. \quad (\text{vii})$$

Note that  $\rho$  is well defined since we have assumed that  $\mathbf{x} \neq \mathbf{0}_N$  and  $\mathbf{y} \neq \mathbf{0}_N$  and hence  $(\mathbf{x}^T \mathbf{x}) > 0$  and  $(\mathbf{y}^T \mathbf{y}) > 0$  and so the positive square roots,  $(\mathbf{x}^T \mathbf{x})^{1/2}$  and  $(\mathbf{y}^T \mathbf{y})^{1/2}$  are well defined positive numbers.

- (b) Prove that the correlation coefficient is bounded from below by minus one and from above by plus one; i.e., show that:

$$-1 \leq \rho \leq 1. \quad (\text{viii})$$

- (c) Assume that the correlation coefficient between  $\mathbf{Y}$  and  $\mathbf{X}$  is positive; i.e., assume that  $\rho > 0$ . Prove that:

$$\alpha^* \leq 1/\beta^*. \quad (\text{ix})$$

- (d) Under what conditions will (ix) hold as an equality?

- (e) Assume that the correlation coefficient between  $\mathbf{Y}$  and  $\mathbf{X}$  is negative and derive a counterpart inequality involving  $\alpha^*$  and  $\beta^*$  to (ix) above.

*Comment:* The result (ix) is reasonably well known in the literature; e.g., see Kendall and Stuart (1967; 380)[266] or Bartelsman (1995; 60)[22]. However, the implications of the inequality are rather important for applied economists. In many applications, the magnitude of  $\alpha$  or  $\beta$  is very important. Hence if  $\rho$  is positive and a client wants an applied economist to obtain a small estimate for the parameter  $\alpha$ , then the applied economist will be tempted to run a regression of  $\mathbf{Y}$  on  $\mathbf{X}$  but if the client wants a large estimate for  $\alpha$  and hence a small estimate for  $\beta$ , then the applied economist will be tempted to run a regression of  $\mathbf{X}$  on  $\mathbf{Y}$  in order to please the client.

**Problem 3** The *Triangle Inequality*. The (Euclidean) *distance* (or norm) of an  $N$  dimensional vector  $\mathbf{x}$  from the origin is defined as

$$d(\mathbf{x}) \equiv (\mathbf{x}^T \mathbf{x})^{1/2} \quad (\text{i})$$

Let  $\mathbf{x}$  and  $\mathbf{y}$  be two  $N$  dimensional vectors. Show that the following inequality is satisfied:

$$d(\mathbf{x} + \mathbf{y}) \leq d(\mathbf{x}) + d(\mathbf{y}). \quad (\text{ii})$$

*Comment:* This inequality dates back to Euclid.

**Problem 4** Let  $\mathbf{A}$  be an  $N \times N$  positive semidefinite symmetric matrix and let  $\mathbf{x}$  and  $\mathbf{y}$  be two  $N$  dimensional vectors. Show that the following inequality is true.

$$(\mathbf{x}^T \mathbf{A} \mathbf{y})^2 \leq (\mathbf{x}^T \mathbf{A} \mathbf{x})(\mathbf{y}^T \mathbf{A} \mathbf{y}). \quad (\text{i})$$

*Hint:* Since  $\mathbf{A}$  is symmetric, there exists an orthonormal matrix  $\mathbf{U}$  such that:

$$\mathbf{U}^T \mathbf{A} \mathbf{U} = \mathbf{\Lambda}; \quad (\text{ii})$$

$$\mathbf{U}^T \mathbf{U} = \mathbf{I}_N \quad (\text{iii})$$

where  $\mathbf{\Lambda}$  is a diagonal matrix which has the nonnegative eigenvalues of  $\mathbf{A}$ ,  $\lambda_1, \dots, \lambda_N$ , running down its main diagonal and  $\mathbf{I}_N$  is the  $N \times N$  identity matrix. Thus  $\mathbf{A}$  can be written as:

$$\mathbf{A} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T = \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{\Lambda}^{1/2} \mathbf{U}^T = \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{U}^T \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{U}^T = \mathbf{S} \mathbf{S} \quad (\text{iv})$$

where  $\mathbf{\Lambda}^{1/2}$  is a diagonal matrix with the positive square roots of the eigenvalues  $\lambda_1, \dots, \lambda_N$  running down its main diagonal and  $\mathbf{S}$  is the *symmetric square root matrix* for  $\mathbf{A}$ .

### 1.3 The Theorem of the Arithmetic and Geometric Mean

Let  $\mathbf{x} \equiv [x_1, \dots, x_N]$  be a vector of nonnegative numbers.\*<sup>2</sup> The *ordinary geometric mean* of the  $N$  numbers contained in the vector  $\mathbf{x}$  is defined as  $(x_1 x_2 \cdots x_N)^{1/N}$  and the *ordinary arithmetic mean* of these numbers is defined as  $(x_1 + x_2 + \cdots + x_N)/N$ .

In this section, we will deal with *generalized or weighted geometric and arithmetic means* of the  $N$  nonnegative numbers  $x_n$ ;  $n = 1, \dots, N$ . In order to define these weighted means, we first define a vector of *positive weights*  $\boldsymbol{\alpha} \equiv [\alpha_1, \dots, \alpha_N]$ ; i.e. define the components of the  $\boldsymbol{\alpha}$  vector to satisfy the following restrictions:\*<sup>3</sup>

$$\boldsymbol{\alpha} \gg \mathbf{0}_N; \quad \mathbf{1}_N^T \boldsymbol{\alpha} \equiv \sum_{n=1}^N \alpha_n = 1. \quad (1.7)$$

Now we are ready to define the *weighted geometric mean*  $M_0(\mathbf{x})$  as follows:

$$M_0(\mathbf{x}) \equiv \prod_{n=1}^N x_n^{\alpha_n}. \quad (1.8)$$

In a similar fashion, we define the *weighted arithmetic mean*  $M_1(\mathbf{x})$  as follows:\*<sup>4</sup>

$$M_1(\mathbf{x}) \equiv \sum_{n=1}^N \alpha_n x_n. \quad (1.9)$$

\*<sup>2</sup> For consistency, we should define the column vector  $\mathbf{x}$  as  $[x_1, \dots, x_N]^T$ . When it is important to be precise, we will consider all vectors to be column vectors and transpose them when required but when casually defining vectors of variables, we will often define the vector as a row vector.

\*<sup>3</sup> Notation:  $\boldsymbol{\alpha} \gg \mathbf{0}_N$  means that each component of the vector  $\boldsymbol{\alpha}$  is positive,  $\boldsymbol{\alpha} \geq \mathbf{0}_N$  means that each component of  $\boldsymbol{\alpha}$  is nonnegative and  $\boldsymbol{\alpha} > \mathbf{0}_N$  means  $\boldsymbol{\alpha} \geq \mathbf{0}_N$  but  $\boldsymbol{\alpha} \neq \mathbf{0}_N$ .

\*<sup>4</sup> The functions  $M_0(\mathbf{x})$  and  $M_1(\mathbf{x})$  should be defined as  $M_0(\mathbf{x}, \boldsymbol{\alpha})$  and  $M_1(\mathbf{x}, \boldsymbol{\alpha})$  since these means depend on the vector of weights  $\boldsymbol{\alpha}$  as well as the vector of nonnegative variables  $\mathbf{x}$  that is being averaged. However, in all of our applications in this chapter, we will hold the weighting vector  $\boldsymbol{\alpha}$  constant when comparing various means and so for simplicity, we have followed the example of Hardy, Littlewood and Polya (1934; 12)[213] and suppressed the vector  $\boldsymbol{\alpha}$  from the notation. The subscripts 0 and 1 that appear in  $M_0(\mathbf{x})$  and  $M_1(\mathbf{x})$  will be explained later: it will turn out that  $M_0(\mathbf{x})$  and  $M_1(\mathbf{x})$  are special cases of the *means of order  $r$* ,  $M_r(\mathbf{x})$ , where  $r$  is equal to 0 and 1 respectively.

Of course, if each  $\alpha_n$  equals  $1/N$ , then the weighted means defined by (1.8) and (1.9) reduce to the ordinary geometric and arithmetic means of the  $x_n$ .

Before we prove the main result in this section, we require a preliminary result.

**Proposition 2** Let the vector  $\boldsymbol{\alpha}$  satisfy the restrictions (1.7). Define the  $N \times N$  matrix  $\mathbf{A}$  by

$$\mathbf{A} \equiv -\hat{\alpha} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T \quad (1.10)$$

where  $\hat{\alpha}$  is an  $N \times N$  diagonal matrix with  $n$ th element  $\alpha_n$  for  $n = 1, 2, \dots, N$ . Then  $\mathbf{A}$  is a negative semidefinite matrix.

**Proof.** It can readily be verified that  $\mathbf{A}$  is symmetric. We need to show that for all  $\mathbf{z} \neq \mathbf{0}_N$ , we have:

$$\mathbf{z}^T \mathbf{A} \mathbf{z} = \mathbf{z}^T [-\hat{\alpha} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T] \mathbf{z} \leq 0 \quad \text{or} \quad (1.11)$$

$$\mathbf{z}^T \boldsymbol{\alpha}\boldsymbol{\alpha}^T \mathbf{z} \leq \mathbf{z}^T \hat{\alpha} \mathbf{z} \quad \text{or}$$

$$(\boldsymbol{\alpha}^T \mathbf{z})^2 \leq \mathbf{z}^T \hat{\alpha} \mathbf{z} \quad \text{since } \mathbf{z}^T \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \mathbf{z}. \quad (1.12)$$

Since  $\boldsymbol{\alpha} \gg \mathbf{0}_N$ , we can take the positive square root of each  $\alpha_n$ . Let  $\hat{\alpha}^{1/2}$  denote the diagonal  $N \times N$  matrix which has  $n$ th element  $\alpha_n^{1/2}$  for  $n = 1, 2, \dots, N$ . Now define the  $N$  dimensional vectors  $\mathbf{x}$  and  $\mathbf{y}$  as follows:

$$\mathbf{x} \equiv \hat{\alpha}^{1/2} \mathbf{1}_N; \quad \mathbf{y} \equiv \hat{\alpha}^{1/2} \mathbf{z} \quad (1.13)$$

where  $\mathbf{1}_N$  is an  $N$  dimensional vector of ones. Recall the Cauchy-Schwarz inequality (1.1). Substituting (1.13) into (1.1) yields:

$$(\mathbf{1}_N^T \hat{\alpha}^{1/2} \hat{\alpha}^{1/2} \mathbf{z})^2 \leq (\mathbf{1}_N^T \hat{\alpha}^{1/2} \hat{\alpha}^{1/2} \mathbf{1}_N) (\mathbf{z}^T \hat{\alpha}^{1/2} \hat{\alpha}^{1/2} \mathbf{z}) \quad \text{or}$$

$$(\mathbf{1}_N^T \hat{\alpha} \mathbf{z})^2 \leq (\mathbf{1}_N^T \hat{\alpha} \mathbf{1}_N) (\mathbf{z}^T \hat{\alpha} \mathbf{z}) \quad \text{or}$$

$$(\boldsymbol{\alpha}^T \mathbf{z})^2 \leq (\mathbf{1}_N^T \boldsymbol{\alpha}) (\mathbf{z}^T \hat{\alpha} \mathbf{z}) \quad \text{or}$$

$$(\boldsymbol{\alpha}^T \mathbf{z})^2 \leq (\mathbf{z}^T \hat{\alpha} \mathbf{z}) \quad \text{using (1.7)} \quad (1.14)$$

which is (1.12). ■

We note that to get a strict inequality in (1.12), we require  $\mathbf{z} \neq \mathbf{0}_N$  and  $\mathbf{x}$  not proportional to  $\mathbf{y}$  or using (1.13), we require  $\mathbf{z} \neq k \mathbf{1}_N$  for any scalar  $k$ .

**Proposition 3** *Theorem of the Arithmetic and Geometric Means:*<sup>\*5</sup>

For every  $\mathbf{x} \gg \mathbf{0}_N$  and positive vector of weights  $\boldsymbol{\alpha}$  which satisfies (1.7), we have:

$$M_0(\mathbf{x}) \leq M_1(\mathbf{x}). \quad (1.15)$$

The strict inequality in (1.15) holds unless  $\mathbf{x} = k \mathbf{1}_N$  for some  $k > 0$  in which case (1.15) becomes:

$$M_0(k \mathbf{1}_N) = M_1(k \mathbf{1}_N) = k; \quad (1.16)$$

i.e., the weighted geometric mean of  $N$  positive numbers is always less than the corresponding weighted arithmetic mean, unless all of the numbers are equal, in which case the means are equal.

---

<sup>\*5</sup> The equal weights case of this Theorem, where  $\boldsymbol{\alpha}$  is equal to  $(1/N) \mathbf{1}_N$ , can be traced back to Euclid and Cauchy (1821; 375)[48] according to Hardy, Littlewood and Polya (1934; 17)[213]. For alternative proofs of the general Theorem, see Hardy, Littlewood and Polya (1934; 17-21)[213].

**Proof.** Define the function of  $N$  variables  $f(\mathbf{x})$  for  $\mathbf{x} \geq \mathbf{0}_N$  as follows:

$$f(\mathbf{x}) \equiv M_0(\mathbf{x}) - M_1(\mathbf{x}) = \prod_{n=1}^N x_n^{\alpha_n} - \sum_{n=1}^N \alpha_n x_n. \quad (1.17)$$

We wish to show that for every  $\mathbf{x} \geq \mathbf{0}_N$ ,

$$f(\mathbf{x}) \leq 0. \quad (1.18)$$

One way to establish (1.15) or (1.18) is to solve the following maximization problem and show that maximizing values of the objective function are equal to or less than 0:

$$\max_{\mathbf{x}} \{f(\mathbf{x}) : \mathbf{x} \geq \mathbf{0}_N\}. \quad (1.19)$$

To begin our proof, we show that points  $\mathbf{x}^0$  which satisfy the first order necessary conditions for maximizing the  $f(\mathbf{x})$  defined by (1.17) (ignoring for now the nonnegativity restrictions  $\mathbf{x} \geq \mathbf{0}_N$ ) are such that  $f(\mathbf{x}^0) = 0$ .

Partially differentiating  $f$  defined by (1.17) and setting the resulting partial derivatives equal to zero yields the following system of equations:

$$\frac{\partial f(\mathbf{x})}{\partial x_n} = \alpha_n x_n^{-1} M_0(\mathbf{x}) - \alpha_n = 0; \quad n = 1, \dots, N \quad \text{or} \quad (1.20)$$

$$x_n = M_0(\mathbf{x}); \quad n = 1, \dots, N. \quad (1.21)$$

Thus if each  $x_n^0$  equals a positive constant,  $k > 0$  say, we will satisfy the first order necessary conditions (1.20) for maximizing  $f(\mathbf{x})$  in the interior of the feasible region. Thus  $\mathbf{x}^0$  of the form:

$$\mathbf{x}^0 \equiv k \mathbf{1}_N; \quad k > 0 \quad (1.22)$$

are such that:

$$\nabla_{\mathbf{x}} f(\mathbf{x}^0) = \mathbf{0}_N \quad \text{and} \quad (1.23)$$

$$f(\mathbf{x}^0) = M_0(k \mathbf{1}_N) - M_1(k \mathbf{1}_N) = k - k = 0 \quad (1.24)$$

where we have used the restrictions in (1.7) to derive (1.24).

We now calculate the matrix of second order partial derivatives of  $f$  defined by (1.17):

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} = -\alpha_n x_n^{-2} M_0(\mathbf{x}) + \alpha_n^2 x_n^{-2} M_0(\mathbf{x}); \quad n = 1, \dots, N; \quad (1.25)$$

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} = \alpha_i \alpha_j x_i^{-1} x_j^{-1} M_0(\mathbf{x}); \quad i \neq j. \quad (1.26)$$

Thus the matrix of second order partial derivatives of  $f$  evaluated at  $\mathbf{x} \gg \mathbf{0}_N$  can be written as follows:

$$\nabla_{xx}^2 f(\mathbf{x}) = \hat{\mathbf{x}}^{-1} [-\hat{\alpha} + \boldsymbol{\alpha} \boldsymbol{\alpha}^T] \hat{\mathbf{x}}^{-1} M_0(\mathbf{x}) \quad (1.27)$$

where  $\hat{\mathbf{x}}$  and  $\hat{\alpha}$  are the vectors  $\mathbf{x}$  and  $\boldsymbol{\alpha}$  diagonalized into matrices. Note also that  $M_0(\mathbf{x}) > 0$  for any  $\mathbf{x} \gg \mathbf{0}_N$ .

To determine the definiteness properties of the  $\nabla_{xx}^2 f(\mathbf{x})$  defined by (1.27), look at:

$$\begin{aligned} M_0(\mathbf{x})^{-1} \mathbf{z}^T \nabla_{xx}^2 f(\mathbf{x}) \mathbf{z} &= \mathbf{z}^T \hat{\mathbf{x}}^{-1} [-\hat{\alpha} + \boldsymbol{\alpha} \boldsymbol{\alpha}^T] \hat{\mathbf{x}}^{-1} \mathbf{z} \\ &= \mathbf{y}^T [-\hat{\alpha} + \boldsymbol{\alpha} \boldsymbol{\alpha}^T] \mathbf{y} \quad \text{where } \mathbf{y} \equiv \hat{\mathbf{x}}^{-1} \mathbf{z} \\ &\leq 0 \end{aligned} \quad (1.28)$$

where the inequality follows using Proposition 2. The inequality in (1.28) will be strict provided that  $\mathbf{y} \neq \mathbf{0}_N$  and  $\mathbf{y} \neq k\mathbf{1}_N$  for any  $k$ .

Now let  $\mathbf{x}^1 \gg \mathbf{0}_N$  be an arbitrary positive vector which is not on the equal component ray; i.e.,

$$\mathbf{x}^1 \gg \mathbf{0}_N \quad \text{but} \quad \mathbf{x}^1 \neq k\mathbf{1}_N \quad \text{for any } k. \tag{1.29}$$

Recall that if  $\mathbf{x}^0 = k\mathbf{1}_N$  for  $k > 0$ , then (1.24) implies  $f(\mathbf{x}^0) = 0$ . Hence to establish our result, we need only show  $f(\mathbf{x}^1) < 0$ .

Recall Taylor's Theorem for  $n = 2$ . The multivariate version of this Theorem yields the following relationship between  $f(\mathbf{x}^0)$  and  $f(\mathbf{x}^1)$  where  $\mathbf{x}^0$  is defined by (1.22) and  $\mathbf{x}^1$  is defined by (1.29): there exists a  $t$  such that  $0 < t < 1$  and

$$\begin{aligned} f(\mathbf{x}^1) &= f(\mathbf{x}^0) + \nabla_x f(\mathbf{x}^0)^T (\mathbf{x}^1 - \mathbf{x}^0) + \frac{1}{2} (\mathbf{x}^1 - \mathbf{x}^0)^T \nabla_{xx}^2 f((1-t)\mathbf{x}^0 + t\mathbf{x}^1) (\mathbf{x}^1 - \mathbf{x}^0) \\ &= 0 + \mathbf{0}^T (\mathbf{x}^1 - \mathbf{x}^0) + \frac{1}{2} (\mathbf{x}^1 - \mathbf{x}^0)^T \nabla_{xx}^2 f((1-t)\mathbf{x}^0 + t\mathbf{x}^1) (\mathbf{x}^1 - \mathbf{x}^0) \quad \text{using (1.23) and (1.24)} \\ &\leq 0 \quad \text{using } M_0(\mathbf{x}) > 0 \text{ and (1.28) for } \mathbf{x} = \mathbf{x}^1 \text{ and } \mathbf{z} = \mathbf{x}^1 - \mathbf{x}^0. \end{aligned} \tag{1.30}$$

In order for the inequality (1.30) to be strict, we require that:

$$\hat{\mathbf{x}}^{-1}(\mathbf{x}^1 - \mathbf{x}^0) \neq k\mathbf{1}_N \quad \text{for any } k \text{ where } \mathbf{x} \equiv (1-t)\mathbf{x}^0 + t\mathbf{x}^1 \tag{1.31}$$

or equivalently, that

$$\mathbf{x}^1 - \mathbf{x}^0 \neq k[(1-t)\mathbf{x}^0 + t\mathbf{x}^1] \quad \text{for any } k. \tag{1.32}$$

Using the facts that  $\mathbf{x}^0 \neq \mathbf{x}^1$  and  $0 < t < 1$ , it can be verified that (1.31) is true and hence the inequality in (1.30) is strict. Thus we have proven (1.15). ■

The geometry associated with the inequalities in (1.32) is illustrated in Figure 1.1 below.

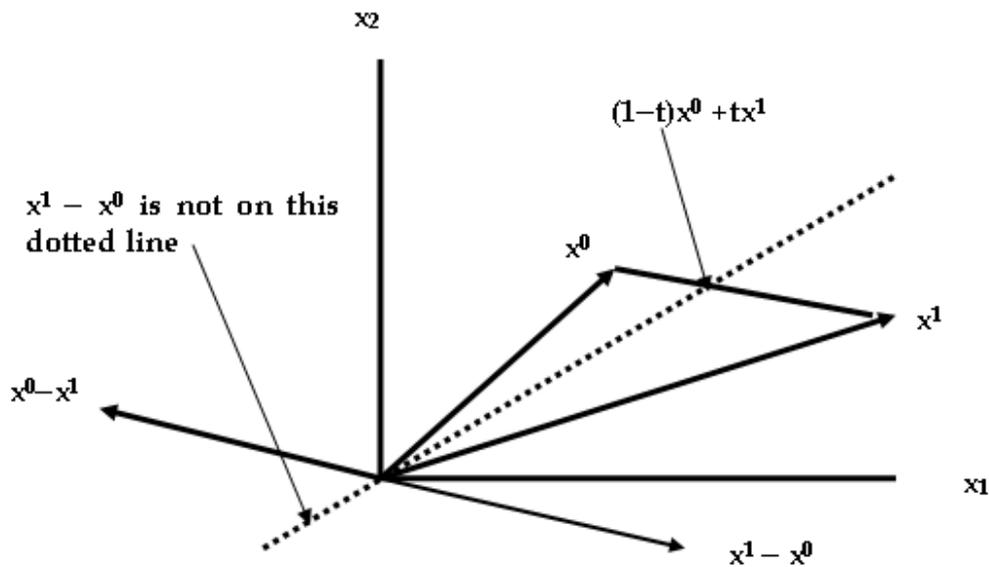


Fig. 1.1

We have established that the weighted geometric mean  $M_0(\mathbf{x})$  is strictly less than the corresponding weighted arithmetic mean  $M_1(\mathbf{x})$  for *strictly positive*  $\mathbf{x} \gg \mathbf{0}_N$ , unless  $\mathbf{x}$  has all components equal, in

which case the two means coincide and are equal to the common component. It is useful to extend the Theorem to cover the case where  $\mathbf{x}$  is *nonnegative*; i.e., to cover the case where one or more components of the  $\mathbf{x}$  vector are equal to zero. But this is easily done. In this case,  $M_0(\mathbf{x})$  is equal to zero and  $M_1(\mathbf{x})$  is equal to or greater than 0 (and strictly greater than 0 if  $\mathbf{x} > \mathbf{0}_N$ ). Thus we have:

$$0 = M_0(\mathbf{x}) < M_1(\mathbf{x}) \quad \text{if } \mathbf{x} > \mathbf{0}_N \text{ and one or more components of } \mathbf{x} \text{ are equal to 0.} \quad (1.33)$$

## 1.4 Means of Order $r$

As in the previous section, we again assume that the vector of weights  $\alpha$  has positive components which sum to one; i.e., we assume  $\alpha$  satisfies conditions (1.7). We assume initially that the number  $r$  is not equal to zero and the vector  $\mathbf{x}$  has positive components and define the *weighted mean of order  $r$*  of the  $N$  numbers in  $\mathbf{x}$  as follows:<sup>\*6</sup>

$$M_r(\mathbf{x}) \equiv \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{1/r}. \quad (1.34)$$

It can be seen that the *mean of order 1* is the weighted arithmetic mean defined earlier by (1.9). It is easy to verify that the means of order  $r$  are (positively) linearly homogeneous in the  $\mathbf{x}$  variables; i.e.,<sup>\*7</sup>

$$M_r(\lambda \mathbf{x}) = \lambda M_r(\mathbf{x}) \quad \text{for every } \mathbf{x} \gg \mathbf{0}_N \text{ and scalar } \lambda > 0. \quad (1.35)$$

The functional form defined by (1.34) occurs frequently in the economics literature. If we multiply  $M_r(\mathbf{x})$  by a constant, then we obtain the CES (constant elasticity of substitution) functional form popularized by Arrow, Chenery, Minhas and Solow (1961)[13] in the context of production theory. This functional form is also widely used as a utility function and it also used extensively when measures of income inequality are constructed.

Three other properties of the means of order  $r$  which are useful are the following ones (we assume  $\mathbf{x} \gg \mathbf{0}_N$  and  $r \neq 0$ ):<sup>\*8</sup>

$$M_r(x_1, \dots, x_N) = [M_1(x_1^r, \dots, x_N^r)]^{1/r}; \quad (1.36)$$

$$M_0(x_1, \dots, x_N) = \exp[M_1(\ln x_1, \dots, \ln x_N)]; \quad (1.37)$$

$$M_{-r}(x_1, \dots, x_N) = 1/M_r(x_1^{-1}, \dots, x_N^{-1}). \quad (1.38)$$

**Problem 5** Prove (1.36), (1.37) and (1.38).

We now consider the problems associated with extending the definition of  $M_r(\mathbf{x})$  from the positive orthant (the set of  $\mathbf{x}$  such that  $\mathbf{x} \gg \mathbf{0}_N$ ) to the nonnegative orthant (the set of  $\mathbf{x}$  such that  $\mathbf{x} \geq \mathbf{0}_N$ ). If  $r \geq 0$ , there is no problem with making this extension since in this case,  $x_n^r$  tends to 0 as  $x_n$  tends to zero and  $M_r(\mathbf{x})$  turns out to be a nice continuous function over the nonnegative orthant. But if  $r < 0$ , there is a problem since  $x_n^r$  tends to  $+\infty$  as  $x_n$  tends to zero in this case. However, in this case, we define  $M_r(\mathbf{x})$  to equal zero:

$$M_r(\mathbf{x}) \equiv 0 \quad \text{if } r < 0 \text{ and any component of } \mathbf{x} \text{ is 0.} \quad (1.39)$$

<sup>\*6</sup> Hardy, Littlewood and Polya (1934; 12-13)[213] refer to this family of means or averages as elementary weighted mean values and study their properties in great detail. When they consider the case where the weights are equal, they refer to the family of means as ordinary mean values.

<sup>\*7</sup> This is property (2.2.13) noted in Hardy, Littlewood and Polya (1934; 14)[213].

<sup>\*8</sup> These properties may be found in Hardy, Littlewood and Polya (1934; 14)[213].

It turns out that with definition (1.39), the means of order  $r$  are continuous functions over the nonnegative orthant even if  $r$  is less than 0. To see why this is the case, consider the case where  $r = -1$ ,  $N = 2$ ,  $\alpha_1 = \frac{1}{2}$ ,  $\alpha_2 = \frac{1}{2}$  and  $x_1$  tends to 0 with  $x_2 > 0$ . In this case, we have for  $x_1 > 0$ :

$$\begin{aligned} M_{-1}(x_1, x_2) &= \left[ \frac{1}{2}x_1^{-1} + \frac{1}{2}x_2^{-1} \right]^{-1} \\ &= 1 / \left[ \frac{1}{2} \left( \frac{1}{x_1} \right) + \frac{1}{2} \left( \frac{1}{x_2} \right) \right] \\ &= x_1 / \left[ \frac{1}{2} + \frac{1}{2} \left( \frac{x_1}{x_2} \right) \right]. \end{aligned} \quad (1.40)$$

Taking the limit of the right hand side of (1.40) as  $x_1$  approaches 0 gives us the limiting value of 0. We will now calculate the vector of first order derivatives of  $M_r(\mathbf{x})$  and the matrix of second order derivatives of  $M_r(\mathbf{x})$  for  $r \neq 0$  and  $\mathbf{x} \gg \mathbf{0}_N$ .<sup>\*9</sup>

**Proposition 4** The matrix of second order partial derivatives of  $M_r(\mathbf{x})$  with respect to the components of the vector  $\mathbf{x}$ ,  $\nabla_{xx}^2 M_r(\mathbf{x})$ , is negative semidefinite for  $r \leq 1$  and positive semidefinite for  $r \geq 1$  for  $\mathbf{x} \gg \mathbf{0}_N$  and  $r \neq 0$ .

**Proof.** Differentiating  $M_r(\mathbf{x})$  with respect to  $x_i$  yields:

$$\begin{aligned} \frac{\partial M_r(\mathbf{x})}{\partial x_i} &= (1/r) \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{(1/r)-1} \alpha_i r x_i^{r-1} \\ &= \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{(1/r)-1} \alpha_i x_i^{r-1}; \quad i = 1, \dots, N. \end{aligned} \quad (1.41)$$

Differentiating (1.41) again with respect to  $x_i$  yields:

$$\begin{aligned} \frac{\partial^2 M_r(\mathbf{x})}{\partial x_i^2} &= [(1/r) - 1] \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{(1/r)-2} \alpha_i r x_i^{r-1} \alpha_i x_i^{r-1} \\ &\quad + \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{(1/r)-1} \alpha_i (r-1) x_i^{r-2}; \quad i = 1, \dots, N \\ &= [r-1] \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{(1/r)-2} \left\{ \left[ \sum_{n=1}^N \alpha_n x_n^r \right] \alpha_i x_i^{r-2} - \alpha_i^2 x_i^{2r-2} \right\}. \end{aligned} \quad (1.42)$$

Differentiating (1.41) with respect to  $x_j$  for  $j \neq i$  yields:

$$\begin{aligned} \frac{\partial^2 M_r(\mathbf{x})}{\partial x_i \partial x_j} &= [(1/r) - 1] \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{(1/r)-2} \alpha_j r x_j^{r-1} \alpha_i x_i^{r-1} \\ &= -(r-1) \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{(1/r)-2} \alpha_i \alpha_j x_i^{r-1} x_j^{r-1}. \end{aligned} \quad (1.43)$$

Using (1.42) and (1.43), we can write the matrix of second order partial derivatives of  $M_r(\mathbf{x})$  as follows:

$$\nabla_{xx}^2 M_r(\mathbf{x}) = (r-1) \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{(1/r)-2} \left\{ \left[ \sum_{n=1}^N \alpha_n x_n^r \right] \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\alpha} \widehat{\mathbf{x}}^{(r/2)-1} - \widehat{\mathbf{x}}^{r-1} \alpha \alpha^T \widehat{\mathbf{x}}^{r-1} \right\} \quad (1.44)$$

<sup>\*9</sup> We have already calculated these derivatives for  $M_0(\mathbf{x})$  in Proposition 3.

where  $\widehat{\mathbf{x}}^{r-1}$  is a diagonal matrix which has  $n$ th element equal to  $x_n^{r-1}$  and  $\widehat{\mathbf{x}}^{(r/2)-1}$  is a diagonal matrix which has diagonal elements equal to  $x_n^{(r/2)-1}$  for  $n = 1, \dots, N$ .

We now want to show that the matrix  $\mathbf{A}$  defined as

$$\mathbf{A} \equiv \left[ \sum_{n=1}^N \alpha_n x_n^r \right] \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\alpha} \widehat{\mathbf{x}}^{(r/2)-1} - \widehat{\mathbf{x}}^{r-1} \alpha \alpha^T \widehat{\mathbf{x}}^{r-1} \quad (1.45)$$

is positive semidefinite.  $\mathbf{A}$  will be positive semidefinite if for every vector  $\mathbf{z}$ , we have  $\mathbf{z}^T \mathbf{A} \mathbf{z} \geq 0$  or

$$\begin{aligned} & \left[ \sum_{n=1}^N \alpha_n x_n^r \right] \mathbf{z}^T \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\alpha} \widehat{\mathbf{x}}^{(r/2)-1} \mathbf{z} \geq \mathbf{z}^T \widehat{\mathbf{x}}^{r-1} \alpha \alpha^T \widehat{\mathbf{x}}^{r-1} \mathbf{z} \quad \text{or} \\ & (\alpha^T \widehat{\mathbf{x}}^{r-1} \mathbf{z})^2 \leq \left[ \sum_{n=1}^N \alpha_n x_n^r \right] \mathbf{z}^T \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\alpha} \widehat{\mathbf{x}}^{(r/2)-1} \mathbf{z}. \end{aligned} \quad (1.46)$$

In order to establish (1.46), note that:

$$\begin{aligned} (\alpha^T \widehat{\mathbf{x}}^{r-1} \mathbf{z})^2 &= (\mathbf{1}_N^T \widehat{\alpha} \widehat{\mathbf{x}}^{r-1} \mathbf{z})^2 \quad \text{using } \alpha = \widehat{\alpha} \mathbf{1}_N \\ &= (\mathbf{1}_N^T \widehat{\alpha}^{1/2} \widehat{\alpha}^{1/2} \widehat{\mathbf{x}}^{r/2} \widehat{\mathbf{x}}^{(r/2)-1} \mathbf{z})^2 \\ &= (\mathbf{1}_N^T \widehat{\alpha}^{1/2} \widehat{\mathbf{x}}^{r/2} \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\alpha}^{1/2} \mathbf{z})^2 \quad \text{since diagonal matrices commute} \\ &= (\mathbf{u}^T \mathbf{v})^2 \quad \text{with } \mathbf{u}^T \equiv \mathbf{1}_N^T \widehat{\alpha}^{1/2} \widehat{\mathbf{x}}^{r/2} \text{ and } \mathbf{v} \equiv \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\alpha}^{1/2} \mathbf{z} \\ &\leq (\mathbf{u}^T \mathbf{u})(\mathbf{v}^T \mathbf{v}) \quad \text{using the Cauchy-Schwarz inequality} \\ &= (\mathbf{1}_N^T \widehat{\alpha}^{1/2} \widehat{\mathbf{x}}^{r/2} \widehat{\mathbf{x}}^{r/2} \widehat{\alpha}^{1/2} \mathbf{1}_N) (\mathbf{z}^T \widehat{\alpha}^{1/2} \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\alpha}^{1/2} \mathbf{z}) \\ &= (\mathbf{1}_N^T \widehat{\alpha} \widehat{\mathbf{x}}^r \mathbf{1}_N) (\mathbf{z}^T \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\alpha} \widehat{\mathbf{x}}^{(r/2)-1} \mathbf{z}) \quad \text{since diagonal matrices commute} \\ &= (\alpha^T \widehat{\mathbf{x}}^r \mathbf{1}_N) (\mathbf{z}^T \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\alpha} \widehat{\mathbf{x}}^{(r/2)-1} \mathbf{z}) \quad \text{since } \mathbf{1}_N^T \widehat{\alpha} = \alpha^T \\ &= \left[ \sum_{n=1}^N \alpha_n x_n^r \right] \mathbf{z}^T \widehat{\mathbf{x}}^{(r/2)-1} \widehat{\alpha} \widehat{\mathbf{x}}^{(r/2)-1} \mathbf{z} \end{aligned} \quad (1.47)$$

which establishes (1.45); i.e.,  $\mathbf{A}$  is positive semidefinite. Returning to (1.44), we have:

$$\nabla_{xx}^2 M_r(\mathbf{x}) = (r-1) \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{(1/r)-2} \mathbf{A}. \quad (1.48)$$

Since  $\mathbf{A}$  is positive semidefinite and  $\left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{(1/r)-2}$  is positive since we have assumed that  $\mathbf{x} \gg \mathbf{0}_N$ , we see that  $\nabla_{xx}^2 M_r(\mathbf{x})$  is positive semidefinite if  $r \geq 1$  and is negative semidefinite if  $r \leq 1$ .  
■

The above Proposition shows that  $M_r(\mathbf{x})$  is a *concave function* of  $\mathbf{x}$  over the positive orthant if  $r \leq 1$  and a *convex function* of  $\mathbf{x}$  if  $r \geq 1$ .

## 1.5 Schlömilch's Inequality

In this section, we show that if  $\mathbf{x} \neq k\mathbf{1}_N$ , then  $M_r(\mathbf{x})$  increases as the parameter  $r$  increases. In order to do this, we require a preliminary inequality.

**Proposition 5** Let  $\alpha \gg \mathbf{0}_N$ ,  $\alpha^T \mathbf{1}_N = 1$  and  $\mathbf{y} \gg \mathbf{0}_N$ . Then

$$f(\mathbf{y}) \equiv \alpha^T \mathbf{y} \ln(\alpha^T \mathbf{y}) - \sum_{n=1}^N \alpha_n y_n \ln y_n \leq 0 \quad (1.49)$$

and the inequality is strict if  $\mathbf{y} \neq k\mathbf{1}_N$ .

**Proof.** We use the same technique of proof that we used in proving the Theorem of the Arithmetic and Geometric Mean. We start out by attempting to maximize  $f(\mathbf{y})$  over the positive orthant. The first order necessary conditions for solving this maximization problem are:

$$\begin{aligned} \frac{\partial f(\mathbf{y})}{\partial y_n} &= \alpha_n \ln(\boldsymbol{\alpha}^T \mathbf{y}) + (\boldsymbol{\alpha}^T \mathbf{y})(\boldsymbol{\alpha}^T \mathbf{y})^{-1} \alpha_n - \alpha_n \ln y_n - \alpha_n y_n / y_n; \quad n = 1, \dots, N \\ &= \alpha_n \ln(\boldsymbol{\alpha}^T \mathbf{y}) - \alpha_n \ln y_n \\ &= 0. \end{aligned} \tag{1.50}$$

Equations (1.50) imply that  $\ln y_n = \ln(\boldsymbol{\alpha}^T \mathbf{y})$  for  $n = 1, \dots, N$ . Thus solutions to (1.50) have the form:

$$\mathbf{y}^0 = k\mathbf{1}_N; \quad k > 0. \tag{1.51}$$

Note that

$$\nabla_{\mathbf{y}} f(\mathbf{y}^0) = \mathbf{0}_N \quad \text{and} \tag{1.52}$$

$$f(\mathbf{y}^0) = \boldsymbol{\alpha}^T k\mathbf{1}_N \ln(\boldsymbol{\alpha}^T k\mathbf{1}_N) - \sum_{n=1}^N \alpha_n k \ln k = k \ln k - k \ln k = 0 \tag{1.53}$$

where we have used  $\boldsymbol{\alpha}^T \mathbf{1}_N = 1$ . Now differentiate equations (1.50) again in order to obtain the following second order partial derivatives of  $f$ :

$$f_{ii}(\mathbf{y}) = \alpha_i (\boldsymbol{\alpha}^T \mathbf{y})^{-1} \alpha_i - \alpha_i y_i^{-1}; \quad i = 1, \dots, N; \tag{1.54}$$

$$f_{ij}(\mathbf{y}) = \alpha_i (\boldsymbol{\alpha}^T \mathbf{y})^{-1} \alpha_j; \quad i \neq j. \tag{1.55}$$

Equations (1.54) and (1.55) can be rewritten in matrix form as follows:

$$\nabla^2 f(\mathbf{y}) = -\hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}} \hat{\mathbf{y}}^{-1/2} + (\boldsymbol{\alpha}^T \mathbf{y})^{-1} \boldsymbol{\alpha} \boldsymbol{\alpha}^T \tag{1.56}$$

where  $\hat{\mathbf{y}}^{-1/2}$  is a diagonal matrix with  $i$ th element equal to  $y_i^{-1/2}$  for  $i = 1, 2, \dots, N$ . We now show that  $\nabla^2 f(\mathbf{y})$  is negative semidefinite; i.e., we want to show that for all  $\mathbf{z}$ :

$$-\mathbf{z}^T \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}} \hat{\mathbf{y}}^{-1/2} \mathbf{z} + \mathbf{z}^T (\boldsymbol{\alpha}^T \mathbf{y})^{-1} \boldsymbol{\alpha} \boldsymbol{\alpha}^T \mathbf{z} \leq 0 \quad \text{or} \tag{1.57}$$

$$(\boldsymbol{\alpha}^T \mathbf{z})^2 \leq (\boldsymbol{\alpha}^T \mathbf{y}) \mathbf{z}^T \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}} \hat{\mathbf{y}}^{-1/2} \mathbf{z} \quad \text{for all } \mathbf{z} \neq \mathbf{0}_N. \tag{1.58}$$

To prove (1.58), we will use the Cauchy-Schwarz inequality:

$$\begin{aligned} (\boldsymbol{\alpha}^T \mathbf{z})^2 &= (\mathbf{z}^T \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}} \hat{\mathbf{y}}^{1/2} \hat{\boldsymbol{\alpha}} \mathbf{1}_N)^2 \quad \text{since } \boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}} \mathbf{1}_N \\ &= (\mathbf{z}^T \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}}^{1/2} \hat{\boldsymbol{\alpha}}^{1/2} \hat{\mathbf{y}}^{1/2} \mathbf{1}_N)^2 \\ &= (\mathbf{z}^T \hat{\boldsymbol{\alpha}}^{1/2} \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}}^{1/2} \hat{\mathbf{y}}^{1/2} \mathbf{1}_N)^2 \quad \text{since diagonal matrices commute} \\ &\leq (\mathbf{z}^T \hat{\boldsymbol{\alpha}}^{1/2} \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}}^{1/2} \mathbf{z})(\mathbf{1}_N^T \hat{\boldsymbol{\alpha}}^{1/2} \hat{\mathbf{y}}^{1/2} \hat{\boldsymbol{\alpha}}^{1/2} \mathbf{1}_N) \\ &\text{using the Cauchy-Schwarz inequality with } \mathbf{x} \equiv \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}}^{1/2} \mathbf{z} \text{ and } \mathbf{y} \equiv \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}}^{1/2} \mathbf{1}_N \\ &= (\mathbf{z}^T \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}}^{1/2} \hat{\boldsymbol{\alpha}}^{1/2} \hat{\mathbf{y}}^{-1/2} \mathbf{z})(\mathbf{1}_N^T \hat{\boldsymbol{\alpha}}^{1/2} \hat{\boldsymbol{\alpha}}^{1/2} \hat{\mathbf{y}}^{1/2} \hat{\mathbf{y}}^{1/2} \mathbf{1}_N) \\ &= (\mathbf{z}^T \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}} \hat{\mathbf{y}}^{-1/2} \mathbf{z})(\mathbf{1}_N^T \hat{\boldsymbol{\alpha}} \mathbf{1}_N) \\ &= (\mathbf{z}^T \hat{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\alpha}} \hat{\mathbf{y}}^{-1/2} \mathbf{z})(\boldsymbol{\alpha}^T \mathbf{y}) \quad \text{since } \boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}} \mathbf{1}_N \text{ and } \mathbf{y} = \hat{\mathbf{y}} \mathbf{1}_N \end{aligned} \tag{1.59}$$

which is (1.58). The inequality (1.59) will be strict, provided that  $\mathbf{x} \equiv \widehat{\mathbf{y}}^{-1/2}\widehat{\alpha}^{1/2}\mathbf{z}$  and  $\mathbf{y} \equiv \widehat{\mathbf{y}}^{-1/2}\widehat{\alpha}^{1/2}\mathbf{1}_N$  are not proportional, or  $\widehat{\alpha}^{1/2}\mathbf{z}$  and  $\widehat{\mathbf{y}}\widehat{\alpha}^{1/2}\mathbf{1}_N$  are not proportional, or  $\widehat{\alpha}^{1/2}\mathbf{z}$  and  $\widehat{\alpha}^{1/2}\widehat{\mathbf{y}}\mathbf{1}_N$  are not proportional, or provided that  $\mathbf{z}$  is not proportional to  $\mathbf{y}$ . Now let the  $\mathbf{y}$  vector in (1.59) be a  $\mathbf{y}^0 = k\mathbf{1}_N$  for  $k > 0$  that satisfies the first order conditions (1.51) above. Then the strict inequality in (1.59) will hold provided that  $\mathbf{z}$  is not proportional to  $k\mathbf{1}_N$ . We use this result that  $\mathbf{z}^T\nabla^2 f(k\mathbf{1}_N)\mathbf{z} < 0$  provided that  $\mathbf{z}$  is not equal to  $\lambda\mathbf{1}_N$  for any scalar  $\lambda$  in the last part of the proof below.

Now let  $\mathbf{y}^1$  be a positive vector that does not have all components equal; i.e.,

$$\mathbf{y}^1 \gg \mathbf{0}_N \quad \text{but} \quad \mathbf{y}^1 \neq k\mathbf{1}_N \text{ for any } k. \quad (1.60)$$

We need only show  $f(\mathbf{y}^1) < 0$  to complete the proof. Apply Taylor's Theorem to the  $f$  defined by (1.49) and the  $\mathbf{y}^0$  and  $\mathbf{y}^1$  defined by (1.51) and (1.60). Thus there exists a  $t$  such that  $0 < t < 1$  and

$$\begin{aligned} f(\mathbf{y}^1) &= f(\mathbf{y}^0) + \nabla_{\mathbf{y}} f(\mathbf{y}^0)^T(\mathbf{y}^1 - \mathbf{y}^0) + \frac{1}{2}(\mathbf{y}^1 - \mathbf{y}^0)^T \nabla_{\mathbf{y}\mathbf{y}}^2 f((1-t)\mathbf{y}^0 + t\mathbf{y}^1)(\mathbf{y}^1 - \mathbf{y}^0) \\ &= 0 + \mathbf{0}^T(\mathbf{y}^1 - \mathbf{y}^0) + \frac{1}{2}(\mathbf{y}^1 - \mathbf{y}^0)^T \nabla_{\mathbf{y}\mathbf{y}}^2 f((1-t)\mathbf{y}^0 + t\mathbf{y}^1)(\mathbf{y}^1 - \mathbf{y}^0) \quad \text{using (1.52) and (1.53)} \\ &\leq 0 \end{aligned} \quad (1.61)$$

where the inequality follows using (1.59) which implies that  $\nabla_{\mathbf{y}\mathbf{y}}^2 f((1-t)\mathbf{y}^0 + t\mathbf{y}^1)$  is negative semidefinite.

In order for the inequality (1.61) to be strict, consider the behavior of the function of one variable  $t$ ,  $g(t) \equiv f[\mathbf{y}^0 + t(\mathbf{y}^1 - \mathbf{y}^0)]$ , defined for  $0 \leq t \leq 1$ . Note that the first and second derivatives of  $g(t)$  are given by  $g'(t) = (\mathbf{y}^1 - \mathbf{y}^0)^T \nabla f[\mathbf{y}^0 + t(\mathbf{y}^1 - \mathbf{y}^0)]$  and  $g''(t) = (\mathbf{y}^1 - \mathbf{y}^0)^T \nabla^2 f[\mathbf{y}^0 + t(\mathbf{y}^1 - \mathbf{y}^0)](\mathbf{y}^1 - \mathbf{y}^0) \leq 0$  where the inequality follows using (1.56)-(1.59). Since  $\mathbf{y}^0 = k\mathbf{1}_N$  by (1.51) and  $\nabla_{\mathbf{y}} f(\mathbf{y}^0) = \mathbf{0}_N$  by (1.52), we see that  $g'(0) = 0$  and  $g''(0) = (\mathbf{y}^1 - \mathbf{y}^0)^T \nabla^2 f(\mathbf{y}^0)(\mathbf{y}^1 - \mathbf{y}^0) < 0$  since  $\mathbf{y}^0 = k\mathbf{1}_N$  and  $\mathbf{y}^1 - \mathbf{y}^0$  is not proportional to  $\mathbf{y}^0$ ; i.e., to obtain the strict inequality, we have used the inequality that we established in the paragraph above (1.60). The equality  $g'(0) = 0$  and the inequalities  $g''(0) < 0$  and  $g''(t) \leq 0$  for  $0 \leq t \leq 1$  (along with the continuity of  $f$  and hence  $g$ ) are sufficient to imply that  $g(t)$  is a nonincreasing function for  $0 \leq t \leq 1$  that is initially strictly decreasing for small  $t$ . Hence  $g(0) = f(\mathbf{y}^0) > f(\mathbf{y}^1) = g(1)$ , which completes the proof. ■

Now we are ready for the main result in this section.

**Proposition 6** *Schlömilch's (1858) Inequality:*<sup>\*10</sup> Let  $\mathbf{x} \gg \mathbf{0}_N$  but  $\mathbf{x} \neq k\mathbf{1}_N$  for any  $k > 0$  and let  $r < s$ . As usual, we assume the weighting vector  $\alpha$  satisfies (1.7). Then

$$M_r(\mathbf{x}) < M_s(\mathbf{x}). \quad (1.62)$$

If  $\mathbf{x} = k\mathbf{1}_N$  for some  $k > 0$ , then  $M_r(\mathbf{x}) = M_s(\mathbf{x}) = k$ .

**Proof.** The second part of the theorem is easily verified. The first part of the theorem, (1.62), will be true if we can show that  $M_r(\mathbf{x})$  is a monotonically increasing function of  $r$  or equivalently if we can show that:<sup>\*11</sup>

$$\frac{\partial \ln M_r(\mathbf{x})}{\partial r} > 0 \text{ for all } r \neq 0, \mathbf{x} \gg \mathbf{0}_N, \mathbf{x} \neq k\mathbf{1}_N \text{ for any } k > 0, \alpha \gg \mathbf{0}_N \text{ and } \alpha^T \mathbf{1}_N = 1. \quad (1.63)$$

<sup>\*10</sup> See Hardy, Littlewood and Polya (1934; 26)[213] for alternative proofs of this result.

<sup>\*11</sup> This is not quite equivalent to the desired result: we still have to deal with the cases where  $r$  or  $s$  are equal to 0; i.e., we cannot differentiate  $M_r(\mathbf{x})$  defined by (1.34) with respect to  $r$  when  $r = 0$ .

Recall that  $\partial c^r / \partial r = \partial e^{r \ln c} / \partial r = e^{r \ln c} \ln c = c^r \ln c$  so that using definition (1.34), it can be verified that the inequality (1.63) is equivalent to:

$$\frac{\partial \ln M_r(\mathbf{x})}{\partial r} = -r^{-2} \ln \left[ \sum_{n=1}^N \alpha_n x_n^r \right] + r^{-1} \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{-1} \left[ \sum_{n=1}^N \alpha_n x_n^r \ln x_n \right] > 0 \quad \text{or} \quad (1.64)$$

$$r^{-1} \left[ \sum_{n=1}^N \alpha_n x_n^r \ln x_n \right] > r^{-2} \left[ \sum_{n=1}^N \alpha_n x_n^r \right] \ln \left[ \sum_{n=1}^N \alpha_n x_n^r \right]. \quad (1.65)$$

Since  $r \neq 0$ ,  $r^2 > 0$  and  $r^{-2} > 0$ . Thus

$$\begin{aligned} r^{-2} \left[ \sum_{n=1}^N \alpha_n x_n^r \right] \ln \left[ \sum_{n=1}^N \alpha_n x_n^r \right] &\leq r^{-2} \left[ \sum_{n=1}^N \alpha_n x_n^r \ln x_n^r \right] && \text{using (1.49) with } y_n \equiv x_n^r \\ &= r^{-2} \left[ \sum_{n=1}^N \alpha_n x_n^r r \ln x_n \right] && \text{using } \ln x_n^r = r \ln x_n \\ &= r^{-1} \left[ \sum_{n=1}^N \alpha_n x_n^r \ln x_n \right] \end{aligned} \quad (1.66)$$

and (1.66) is a weak version of (1.65). But the inequality (1.66) is strict provided that the  $y_n = x_n^r$  are not all equal. This is the case since we have assumed  $\mathbf{x} \neq k\mathbf{1}_N$  and thus (1.66) is a strict inequality.

We still need to establish (1.62) when  $r$  or  $s$  equal 0. We first consider the case where  $s > r = 0$ . Let  $\mathbf{x} \gg \mathbf{0}_N$  with  $\mathbf{x} \neq k\mathbf{1}_N$ . Then we have:

$$\begin{aligned} [M_0(\mathbf{x})]^s &= \left[ \prod_{n=1}^N x_n^{\alpha_n} \right]^s && \text{using definition (1.8)} \\ &= M_0(x_1^s, \dots, x_N^s) \\ &< M_1(x_1^s, \dots, x_N^s) && \text{by Proposition 3 since not all of the } x_n^s \text{ are equal} \\ &= M_s(x_1, \dots, x_N)^s && \text{using result (1.36)}. \end{aligned} \quad (1.67)$$

Since  $s > 0$ , taking the  $1/s$  root of both sides of (1.67) will preserve the inequality which establishes (1.62) for  $r = 0 < s$ .

We now consider (1.62) when  $-r < 0 = s$ . Again, let  $\mathbf{x} \gg \mathbf{0}_N$  with  $\mathbf{x} \neq k\mathbf{1}_N$ . Then we have:

$$\begin{aligned} M_{-r}(\mathbf{x}) &= 1/M_r(x_1^{-1}, \dots, x_N^{-1}) && \text{using property (1.38)} \\ &< 1/M_0(x_1^{-1}, \dots, x_N^{-1}) && \text{using } r > 0, \mathbf{x} \neq k\mathbf{1}_N \text{ and (1.67)} \\ &\text{which implies that } M_r(x_1^{-1}, \dots, x_N^{-1}) > M_0(x_1^{-1}, \dots, x_N^{-1}) \\ &= M_0(\mathbf{x}) && \text{using definition (1.8)}. \end{aligned} \quad (1.68)$$

■

The above Theorem shows that the *weighted harmonic mean* of  $N$  positive numbers,  $x_1, \dots, x_N$ , will always be equal to or less than the corresponding weighted arithmetic mean; i.e., we have for  $\mathbf{x} \gg \mathbf{0}_N$ :

$$M_{-1}(\mathbf{x}) \leq M_1(\mathbf{x}) \quad \text{or} \quad (1.69)$$

$$\left[ \sum_{n=1}^N \alpha_n x_n^{-1} \right]^{-1} \leq \sum_{n=1}^N \alpha_n x_n \quad (1.70)$$

and the inequality (1.70) is strict provided that the  $x_n$  are not all equal to the same positive number. What happens if one or more of the components of the  $\mathbf{x}$  vector are equal to 0? Using the continuity of the functions  $M_r(\mathbf{x})$  over the nonnegative orthant, it can be seen that (1.62) will still hold as a weak inequality. It should be kept in mind that if  $r \leq 0$  and any component of  $\mathbf{x}$  is 0, then (1.39) implies that  $M_r(\mathbf{x})$  is equal to 0.

**Problem 6** Suppose a Statistical Agency collects price quotes on a “homogeneous” commodity (e.g. red potatoes) from  $N$  outlets during periods 0 and 1. Denote the vector of period  $t$  price quotes by  $\mathbf{p}^t \equiv [p_1^t, \dots, p_N^t]$  for  $t = 0, 1$ . An *elementary price index*  $P(\mathbf{p}^0, \mathbf{p}^1)$  is a function of  $2N$  variables that aggregates this micro information on potatoes into an aggregate price index for potatoes that will be a component of the overall consumer price index (CPI). Examples of widely used functional forms for  $P$  are the *Carli* (1764) and *Jevons* (1865) formulae defined by (i) and (ii) below:

$$P_C(\mathbf{p}^0, \mathbf{p}^1) \equiv \sum_{n=1}^N (1/N)(p_n^1/p_n^0) \quad (\text{i})$$

which is the equally weighted *arithmetic* mean of the  $N$  price ratios;

$$P_J(\mathbf{p}^0, \mathbf{p}^1) \equiv \left[ \prod_{n=1}^N (p_n^1/p_n^0) \right]^{1/N} \quad (\text{ii})$$

which is the equally weighted *geometric* mean of the  $N$  price ratios.

A very useful property for an elementary price index to satisfy is the *time reversal test*:

$$P(\mathbf{p}^0, \mathbf{p}^1)P(\mathbf{p}^1, \mathbf{p}^0) = 1; \quad (\text{iii})$$

i.e., suppose prices in period 1 reverted back to the base period prices  $\mathbf{p}^0$ . Under these conditions, we should end up at our starting point.

- (a) Show that  $P_J(\mathbf{p}^0, \mathbf{p}^1)$  satisfies the time reversal test.
- (b) Show that  $P_C(\mathbf{p}^0, \mathbf{p}^1)$  has an upward bias; i.e., show that if  $\mathbf{p}^1 \neq k\mathbf{p}^0$ , then

$$P_C(\mathbf{p}^0, \mathbf{p}^1)P_C(\mathbf{p}^1, \mathbf{p}^0) > 1. \quad (\text{iv})$$

*Hint:* You may find (1.70) useful.

*Comment:* Many Statistical Agencies are still using the biased Carli formula to aggregate their price quotes at the lowest level of aggregation. However, in the past decade, several countries (Canada, the U.S. and the member countries of the EU for their harmonized indexes) have switched to the Jevons formula. The use of  $P_C$  rather than  $P_J$  is thought to have generated an upward bias in the CPI in the 0.1- 0.4% per year range. Fisher (1922; 66 and 383)[187] seems to have been the first to establish the upward bias of the Carli index and he made the following observations on its use by statistical agencies:

“In fields other than index numbers it is often the best form of average to use. But we shall see that the simple arithmetic average produces one of the very worst of index numbers. And if this book has no other effect than to lead to the total abandonment of the simple arithmetic type of index number, it will have served a useful purpose.” Irving Fisher (1922; 29-30)[187].

**Problem 7** A *general mean function*,  $M(\mathbf{x})$ , is a function of  $N$  variables, defined for  $\mathbf{x} \gg \mathbf{0}_N$  that has the following three properties:

- (i)  $M(k\mathbf{1}_N) = k$  for  $k > 0$  (*mean value property*);
- (ii)  $M(\mathbf{x})$  is a *continuous* function; and

(iii)  $M(\mathbf{x})$  is increasing in its components; i.e., if  $\mathbf{x}^1 < \mathbf{x}^2$ , then  $M(\mathbf{x}^1) < M(\mathbf{x}^2)$ .

It is easy to see that the weighted means of order  $r$ ,  $M_r(\mathbf{x})$  defined by (1.34), satisfy properties (i) and (ii). Show that they also satisfy property (iii).

*Hint:* Show that  $\partial M_r(\mathbf{x})/\partial x_n > 0$  for  $r \neq 0$ .

**Problem 8**  $M(\mathbf{x})$  is a *symmetric mean* if  $M$  is a mean and has the following property:

(iv)  $M(P\mathbf{x}) = M(\mathbf{x})$  where  $P\mathbf{x}$  is a permutation of the components of  $\mathbf{x}$ . Are the means of order  $r$  symmetric means? If not, what conditions on  $\alpha$  will make  $M_r(\mathbf{x})$  a symmetric mean?

**Problem 9**  $M(\mathbf{x})$  is a *homogeneous mean* if it is a mean and satisfies the following additional property:

(v)  $M(\lambda\mathbf{x}) = \lambda M(\mathbf{x})$  for all  $\lambda > 0$ ,  $\mathbf{x} \gg \mathbf{0}_N$

If  $M(\mathbf{x})$  is a homogeneous mean, show that it also satisfies the following property:

(vi)  $\alpha \equiv \min_n \{x_n : n = 1, \dots, N\} \leq M(\mathbf{x}) \leq \max_n \{x_n : n = 1, \dots, N\} \equiv \beta$ .

This result is due to Eichhorn and Voeller (1976; 10)[169].

*Hint:*  $\alpha \mathbf{1}_N \leq \mathbf{x} \leq \beta \mathbf{1}_N$ . Note that properties (ii) and (iii) for a mean  $M(\mathbf{x})$  imply that the following property also holds for  $M$ :

(vii)  $M(\mathbf{x}^1) \leq M(\mathbf{x}^2)$  if  $\mathbf{x}^1 \leq \mathbf{x}^2$ .

## 1.6 L'Hospital's Rule and Logarithmic Means

In this section, we show that the weighted geometric mean,  $M_0(\mathbf{x})$ , is a limiting case of the corresponding weighted mean of order  $r$ ,  $M_r(\mathbf{x})$ , as  $r$  tends to zero. Before we do this, we require a preliminary result.

**Proposition 7** *L'Hospital's (1696) Rule:*<sup>\*12</sup> Suppose  $f(z)$  and  $g(z)$  are once continuously differentiable functions of one variable  $z$  around an interval including  $z = b$ . In addition, suppose  $f(b) = g(b) = 0$  but  $g'(b) \neq 0$ . Then

$$\lim_{z \rightarrow b} \frac{f(z)}{g(z)} = \frac{f'(b)}{g'(b)}. \quad (1.71)$$

**Proof.** Let  $z$  be close to  $b$  but  $z \neq b$ . Then by the Mean Value Theorem, there exist  $z^*$  and  $z^{**}$  between  $z$  and  $b$  such that:

$$f(z) = f(b) + f'(z^*)(z - b) = f'(z^*)(z - b) \quad \text{since } f(b) = 0; \quad (1.72)$$

$$g(z) = g(b) + g'(z^{**})(z - b) = g'(z^{**})(z - b) \quad \text{since } g(b) = 0. \quad (1.73)$$

Taking the ratio of (1.72) to (1.73) and using the assumptions that  $g'(b) \neq 0$  and that the derivative function  $g'(z)$  is continuous, we can deduce that  $g'(z^{**}) \neq 0$  using if  $z$  is close enough to  $b$  and hence for  $z - b \neq 0$  and  $z$  close to  $b$ , we get:

$$\frac{f(z)}{g(z)} = \frac{f'(z^*)}{g'(z^{**})}. \quad (1.74)$$

<sup>\*12</sup> See Rudin (1953; 82)[337] for a proof of this result.

Now take limits on both sides of (1.74) as  $z$  approaches  $b$ . Since  $z^*$  and  $z^{**}$  are between  $z$  and  $b$ ,  $z^*$  and  $z^{**}$  will tend to  $b$  and thus (1.71) follows, since both  $f'$  and  $g'$  are assumed to be continuous functions. ■

The following problems illustrate a few of the uses of L'Hospital's Rule.

**Problem 10** If  $x > 0$ , show that  $\lim_{r \rightarrow 0} (x^r - 1)/r = \ln x$ .

*Hint:* Use L'Hospital's Rule with  $f(r) \equiv x^r - 1$  and  $g(r) \equiv r$ . Note that if  $h(r) = x^r = e^{r \ln x}$ , then  $h'(r) = e^{r \ln x} \ln x = x^r \ln x$ .

*Comment:* The function  $(x^r - 1)/r$  is known as the *Box-Cox transformation* and it is widely used in statistics and econometrics as well as in the study of choice under uncertainty.

**Problem 11** The *logarithmic mean*,  $L(x_1, x_2)$  of two positive numbers  $x_1 > 0$  and  $x_2 > 0$ , is defined as follows:

$$L(x_1, x_2) \equiv \begin{cases} [x_1 - x_2]/[\ln x_1 - \ln x_2] & \text{if } x_1 \neq x_2 \\ x_2 & \text{if } x_1 = x_2 \end{cases} \quad (\text{i})$$

Show that if  $0 < x_1 < x_2$ , then

$$\lim_{x_1 \rightarrow x_2} L(x_1, x_2) = x_2. \quad (\text{ii})$$

*Hint:* Define  $f(x_1) \equiv x_1 - x_2$  and  $g(x_1) \equiv \ln x_1 - \ln x_2$  and apply L'Hospital's Rule.

*Comment:* This result establishes the continuity of  $L(x_1, x_2)$  over the positive orthant.

**Problem 12** Refer to problems 7-11 above and show that  $L(x_1, x_2)$  defined in Problem 11 above is a homogeneous symmetric mean.

*Hint:* The definition of  $L(x_1, x_2)$  in Problem 11 establishes property (i) in Problem 7. Problem 11 establishes the validity of property (ii) in Problem 7. To prove property (iii), just show  $\partial L(x_1, x_2)/\partial x_n > 0$  for  $n = 1, 2$  (you can assume  $x_1 \neq x_2$ ). In the case of only two variables, the symmetry property (iv) is just  $L(x_1, x_2) = L(x_2, x_1)$  which you can verify. Finally, verify the homogeneity property, (v), that was defined in problem 9.

*Comment:* The logarithmic mean (sometimes called the Vartia mean) plays a key role in index number theory; see Vartia (1976)[382] and Diewert (1978)[85].

## 1.7 Additional Properties of Means of Order $r$

Proposition 8 below justifies our notation,  $M_0(\mathbf{x})$ , for the weighted geometric mean since this Proposition shows that  $M_0(\mathbf{x})$  is a limiting case of  $M_r(\mathbf{x})$  as  $r$  tends to 0.

**Proposition 8** The limiting case of the weighted mean of order  $r$ ,  $M_r(\mathbf{x})$ , as  $r$  tends to 0 is the weighted geometric mean,  $M_0(\mathbf{x})$ \*<sup>13</sup>; i.e., for  $\mathbf{x} \gg \mathbf{0}_N$ ,  $\boldsymbol{\alpha} \gg \mathbf{0}_N$ ,  $\boldsymbol{\alpha}^T \mathbf{1}_N = 1$ :

$$\lim_{r \rightarrow 0} M_r(\mathbf{x}) = M_0(\mathbf{x}). \quad (1.75)$$

**Proof.** Proposition 6 above showed that  $M_r(\mathbf{x})$  is a nondecreasing function of  $r$ . Since  $M_r(\mathbf{x})$  is a homogeneous mean, Problem 9 above shows that  $M_r(\mathbf{x})$  is bounded from above and below; i.e., for all  $r \neq 0$ ;

$$\min_n \{x_n : n = 1, \dots, N\} \leq M_r(\mathbf{x}) \leq \max_n \{x_n : n = 1, \dots, N\}. \quad (1.76)$$

\*<sup>13</sup> See Hardy, Littlewood and Polya (1934; 15)[213] for a proof of this result.

The fact that  $M_r(\mathbf{x})$  is a nondecreasing function of  $r$  and is also bounded from above and below is sufficient to imply the existence of  $\lim_{r \rightarrow 0} M_r(\mathbf{x})$  and also that

$$\lim_{r \rightarrow 0} \ln M_r(\mathbf{x}) = \ln[\lim_{r \rightarrow 0} M_r(\mathbf{x})]. \quad (1.77)$$

We now compute  $\ln M_r(\mathbf{x})$  for  $r \neq 0$ :

$$\ln M_r(\mathbf{x}) = \frac{1}{r} \ln \left[ \sum_{n=1}^N \alpha_n x_n^r \right] \equiv \frac{f(r)}{g(r)} \quad (1.78)$$

where  $g(r) \equiv r$  and  $f(r) \equiv \ln \left[ \sum_{n=1}^N \alpha_n x_n^r \right]$ . Note that:

$$g(0) = 0; \quad (1.79)$$

$$f(0) = \ln \left[ \sum_{n=1}^N \alpha_n (x_n)^0 \right] = \ln \left[ \sum_{n=1}^N \alpha_n 1 \right] = \ln 1 = 0 \quad \text{using } \sum_{n=1}^N \alpha_n = 1. \quad (1.80)$$

Now calculate the derivatives of  $f(r)$  and  $g(r)$  and evaluate them at  $r = 0$ :

$$g'(r) = 1 \text{ and hence} \quad (1.81)$$

$$g'(0) = 1. \quad (1.82)$$

$$f'(r) = \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{-1} \sum_{n=1}^N \alpha_n x_n^r \ln x_n \text{ and hence} \quad (1.83)$$

$$f'(0) = \left[ \sum_{n=1}^N \alpha_n \right]^{-1} \sum_{n=1}^N \alpha_n \ln x_n = \sum_{n=1}^N \alpha_n \ln x_n \quad \text{using } \sum_{n=1}^N \alpha_n = 1. \quad (1.84)$$

Now apply L'Hospital's Rule to (1.78) when  $r = 0$ . The resulting equation is:

$$\lim_{r \rightarrow 0} \ln M_r(\mathbf{x}) = \frac{f'(0)}{g'(0)} = \sum_{n=1}^N \alpha_n \ln x_n \quad \text{using (1.82) and (1.84)}. \quad (1.85)$$

We can exponentiate both sides of (1.85) and deduce that (1.75) holds. ■

When  $N = 2$  and  $\alpha_1 = \alpha_2 = 1/2$ , we can graph the level curves  $\{(x_1, x_2) : M_r(x_1, x_2) = 1\}$  for various values of  $r$ ; see Figure 1.2 below.

We conclude with some results on limiting cases of  $M_r(\mathbf{x})$  as  $r$  tends to  $+\infty$  or  $-\infty$ . The results in Proposition 9 are used in Figure 1.2.

**Proposition 9** Hardy, Littlewood and Polya (1934; 15)[213]: The limits of  $M_r(\mathbf{x})$  as  $r$  tends to  $+\infty$  or  $-\infty$  are as follows:

$$\lim_{r \rightarrow \infty} M_r(\mathbf{x}) = \max_n \{x_n : n = 1, \dots, N\}; \quad (1.86)$$

$$\lim_{r \rightarrow -\infty} M_r(\mathbf{x}) = \min_n \{x_n : n = 1, \dots, N\}. \quad (1.87)$$

**Proof.** Let  $\mathbf{x} > \mathbf{0}_N$  and let  $x_k = \max_n \{x_n : n = 1, \dots, N\}$ . Then using the results in Problem 9, we have:

$$M_r(\mathbf{x}) \leq x_k. \quad (1.88)$$

Since the  $x_n$  are nonnegative and the  $\alpha_n$  are positive, we have:

$$\alpha_k x_k^r \leq \sum_{n=1}^N \alpha_n x_n^r. \quad (1.89)$$

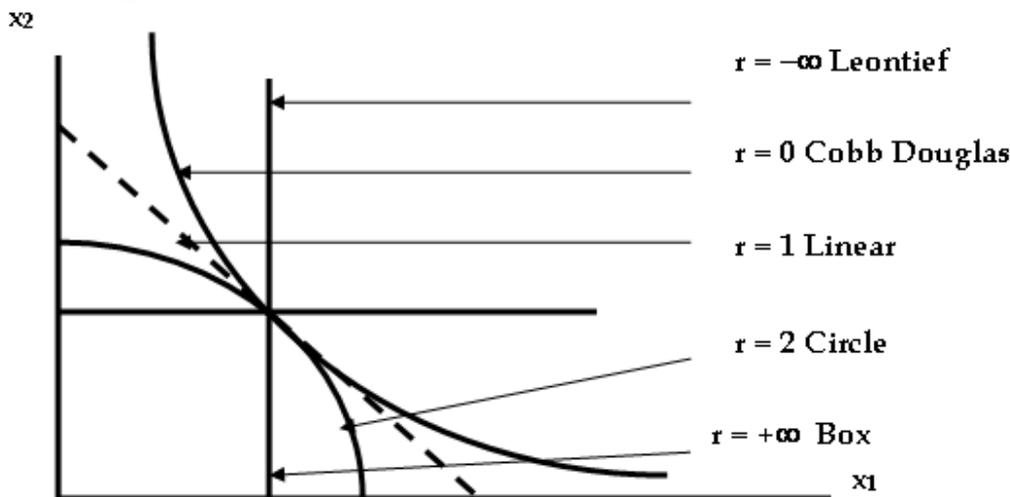


Fig. 1.2 Level Curves for the Symmetric Mean of Order  $r$

Now take the  $r$ th root of both sides of (1.89). If  $r > 0$ , the inequality is preserved and so we have in this case:

$$(\alpha_k)^{1/r} x_k \leq \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{1/r} = M_r(\mathbf{x}). \quad (1.90)$$

Now take the limit of both sides of (1.90) as  $r$  tends to  $+\infty$  and since  $(\alpha_k)^{1/r}$  tends to  $(\alpha_k)^0 = 1$ , we find that

$$x_k \leq \lim_{r \rightarrow \infty} M_r(\mathbf{x}). \quad (1.91)$$

It can be seen that (1.88) and (1.91) imply (1.86).

Now consider (1.87). If one or more of the  $x_n$  are zero, then  $M_r(\mathbf{x})$  equals 0 for all  $r < 0$ ; recall (1.39) above. Hence if one or more of the  $x_n$  are 0, then it is easy to verify that (1.87) holds. Thus we consider the case where  $\mathbf{x} \gg \mathbf{0}_N$  and let  $x_k = \max_n \{x_n : n = 1, \dots, N\}$ . By Problem 9, we have:

$$x_k \leq M_r(\mathbf{x}). \quad (1.92)$$

Since the  $x_n$  are positive and the  $\alpha_n$  are positive, we again have (1.89) but now we assume that  $r < 0$ , so that when we take the  $r$ th root of each side of (1.89), the inequality is reversed and so we have:

$$(\alpha_k)^{1/r} x_k \geq \left[ \sum_{n=1}^N \alpha_n x_n^r \right]^{1/r} = M_r(\mathbf{x}). \quad (1.93)$$

Now take the limit of both sides of (1.93) as  $r$  tends to  $-\infty$  and since  $(\alpha_k)^{1/r}$  tends to  $(\alpha_k)^0 = 1$ , we find that

$$x_k \geq \lim_{r \rightarrow -\infty} M_r(\mathbf{x}) \geq x_k \quad (1.94)$$

where the last inequality follows using (1.92). It can be seen that (1.92) and (1.94) imply (1.87). ■

## 1.8 Summary of Methods used to Establish Inequalities

A careful look at the methods of proof that we have used to establish the validity of various inequalities will show that we have basically used 3 methods:

- Transform the given inequality into a known inequality using ordinary algebra.
- Transform the given inequality into the form  $f(\mathbf{x}) \leq 0$  for the domain of definition for the inequality, say  $\mathbf{x} \in S$ , and show that  $\mathbf{x}^*$  which solve  $\max_{\mathbf{x}}\{f(\mathbf{x}) : \mathbf{x} \in S\}$  are such that  $f(\mathbf{x}^*) \leq 0$ .
- Consider the case where the last method leads to a twice continuously differentiable objective function  $f(\mathbf{x})$  which has the following properties: (a) there exist points  $\mathbf{x}^*$  such that  $f(\mathbf{x}^*) = 0$  and  $\nabla f(\mathbf{x}^*) = \mathbf{0}_N$ ; (b) the domain of definition set  $S$  is convex and (c)  $\nabla^2 f(\mathbf{x})$  is negative semidefinite for each  $\mathbf{x} \in S$ . Then in this case, we can use Taylor's Theorem for  $n = 2$  and establish the desired result,  $f(\mathbf{x}) \leq f(\mathbf{x}^*) = 0$  for all  $\mathbf{x} \in S$ .

It turns out that the three main inequalities that were established in this chapter (the Cauchy Schwarz Inequality, the Theorem of the Arithmetic and Geometric Means and Schlömilch's Inequality) have many applications in all branches of applied economics.

**Problem 13** Let  $\phi(z)$  be a monotonically increasing, continuous function of one variable that is defined for  $z > 0$  so that the inverse function for  $\phi$ ,  $\phi^{-1}(y)$ , is also a monotonically increasing, continuous function of  $y$  for all  $y$ 's belonging to the range of  $\phi$ . As usual, define the vector of weights  $\boldsymbol{\alpha} \equiv [\alpha_1, \dots, \alpha_N]$  which satisfies:

- (i)  $\boldsymbol{\alpha} \gg \mathbf{0}_N$  and
- (ii)  $\mathbf{1}^T \boldsymbol{\alpha} = 1$ .

We use the function  $\phi$  in order to define the following *quasilinear mean* for all  $\mathbf{x} \gg \mathbf{0}_N$ :<sup>\*14</sup>

$$(iii) M_\phi(\mathbf{x}) \equiv \phi^{-1} \left[ \sum_{n=1}^N \alpha_n \phi(x_n) \right].$$

Show that  $M_\phi(\mathbf{x})$  defined by (iii) is a general mean; i.e., it satisfies properties (i)-(iii) listed in Problem 7 above.

*Hint:* You do not have to prove part (ii), continuity, which is obvious.

*Comment:* Note that if  $\phi(z) \equiv z^r$  for  $r > 0$ , then  $M_\phi(\mathbf{x})$  reduces to the weighted mean of order  $r$ ,  $M_r(\mathbf{x})$  and if  $\phi(z) \equiv \ln z$ , then  $M_\phi(\mathbf{x})$  reduces to the weighted geometric mean,  $M_0(\mathbf{x})$ . Hardy, Littlewood and Polya (1934; 68)[213] show that if we require  $M_\phi(\mathbf{x})$  to be a homogeneous mean<sup>\*15</sup>, then essentially,  $M_\phi(\mathbf{x})$  must be a mean of order  $r$ .<sup>\*16</sup>

**Problem 14** Find a general mean function,  $M(x_1, x_2)$ , which is *not* a quasilinear mean of the type defined in Problem 13.

## 1.9 References

- Balk, B.M. (1995), "Axiomatic Price Index Theory: A Survey", *International Statistical Review* 63, 69-93.
- Arrow, K.J., H.B. Chenery, B.S. Minhas and R.M. Solow, (1961), "Capital-Labour Substitution and Economic Efficiency", *Review of Economics and Statistics* 63, 225-250.

<sup>\*14</sup> Eichhorn (1978; 32)[171] used this terminology. The axiomatic properties of this type of mean were first explored by Nagumo (1930)[316], Kolmogoroff (1930)[279] and Hardy, Littlewood and Polya (1934; 65-69)[213]. Kolmogoroff used the term "regular mean" while Hardy, Littlewood and Polya (1934; 65) used the awkward term "mean value with an arbitrary function". Diewert (1993; 358-359)[106] used the term "separable mean". Kolmogoroff, Nagumo and Diewert studied only the equally weighted case.

<sup>\*15</sup> Recall Property (v) in Problem 9.

<sup>\*16</sup> For proofs of this result in the case of equally weighted or symmetric separable means, see Nagumo (1930)[316] and Diewert (1993; 381)[106].

- Bartelsman, E. J. (1995), "Of Empty Boxes: Returns to Scale Revisited," *Economics Letters* 49, 59-67.
- Cauchy, A.L., (1821), *Cours d'analyse de l'École Royal Polytechnique: Analyse algébrique*: Paris.
- Diewert, W.E. (1978), "Superlative Index Numbers and Consistency in Aggregation", *Econometrica* 46, 883-900.
- Diewert, W.E. (1993), "Symmetric Means and Choice under Uncertainty", pp. 355-433 in *Essays in Index Number Theory*, Volume 1 (W.E. Diewert and A.O. Nakamura editors), Amsterdam: North-Holland.
- Diewert, W.E. (1998), "Index Number Issues in the Consumer Price Index", *Journal of Economic Perspectives* 12:1 (Winter), 47-58.
- Eichhorn, W. (1978), *Functional Equations in Economics*, Reading, MA: Addison-Wesley Publishing Company.
- Eichhorn, W. and J. Voeller (1976), *Theory of the Price Index*, Lecture Notes in Economics and Mathematical Systems, Vol. 140, Berlin: Springer-Verlag.
- Fisher, I. (1922), *The Making of Index Numbers*, Houghton-Mifflin, Boston.
- Hardy, G.H., J.E. Littlewood and G. Polya, (1934), *Inequalities*, Cambridge, England: Cambridge University Press.
- Kendall, M.G. and A.S. Stuart (1967), *The Advanced Theory of Statistics: Volume 2: Inference and Relationship*, Second Edition, New York: Hafner Publishing Co.
- Kolmogoroff, A. (1930), "Sur la notion de la moyenne", *Atti della Reale Accademia nazionale dei Lincei* 12(6), 388-391.
- L'Hospital (1696), *L'analyse des infiniment petits pour l'intelligence des lignes courbes*, Paris.
- Nagumo, M. (1930), "Über eine Klasse der Mittelwerte", *Japanese Journal of Mathematics* 7, 71-79.
- Rudin, W., (1953), *Principles of Mathematical Analysis*, New York: McGraw-Hill Book Co.
- Schlömilch, O., (1858), "Über Mittelgrößen verschiedener Ordnungen", *Zeitschrift für Mathematik und Physik* 3, 308-310.
- Schwarz, H.A., (1885), "Über ein die Flächen Kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung", *Acta Societatis scientiarum Fennicae* 15, 315-362.
- Vartia, Y.O. (1976), "Ideal Log-Change Index Numbers", *Scandinavian Journal of Statistics* 3, 121-126.

## Chapter 2

# Convex Sets and Concave Functions

### 2.1 Introduction

Many economic problems have the following structure: (i) a linear function is minimized subject to a nonlinear constraint; (ii) a linear function is maximized subject to a nonlinear constraint or (iii) a nonlinear function is maximized subject to a linear constraint. Examples of these problems are: (i) the producer's cost minimization problem (or the consumer's expenditure minimization problem); (ii) the producer's profit maximization problem and (iii) the consumer's utility maximization problem. These three constrained optimization problems play a key role in economic theory.

In each of the above 3 problems, linear functions appear in either the objective function (the function being maximized or minimized) or the constraint function. If we are maximizing or minimizing the linear function of  $\mathbf{x}$ , say  $\sum_{n=1}^N p_n x_n \equiv \mathbf{p}^T \mathbf{x}$ , where  $\mathbf{p} \equiv (p_1, \dots, p_N)^{*1}$  is a vector of prices and  $\mathbf{x} \equiv (x_1, \dots, x_N)$  is a vector of decision variables, then after the optimization problem is solved, the optimized objective function can be regarded as a function of the price vector, say  $G(\mathbf{p})$ , and perhaps other variables that appear in the constraint function. Let  $F(\mathbf{x})$  be the nonlinear function which appears in the constraint in problems (i) and (ii). Then under certain conditions, the optimized objective function  $G(\mathbf{p})$  can be used to reconstruct the nonlinear constraint function  $F(\mathbf{x})$ . This correspondence between  $F(\mathbf{x})$  and  $G(\mathbf{p})$  is known as *duality theory*. In the following chapter, we will see how the use of dual functions can greatly simplify economic modeling.

However, the mathematical foundations of duality theory rest on the theory of convex sets and concave (and convex) functions. Hence, we will study a few aspects of this theory in the present chapter before studying duality theory in the following chapter.

### 2.2 Convex Sets

**Definition** A set  $S$  in  $\mathbb{R}^N$  (Euclidean  $N$  dimensional space) is *convex* iff (if and only if):

$$\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, 0 < \lambda < 1 \text{ implies } \lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2 \in S. \quad (2.1)$$

Thus a set  $S$  is convex if the line segment joining any two points belonging to  $S$  also belongs to  $S$ . Some examples of convex sets are given below.

---

<sup>\*1</sup> Our convention is that in equations, vectors like  $\mathbf{p}$  and  $\mathbf{x}$  are regarded as column vectors and  $\mathbf{p}^T$  and  $\mathbf{x}^T$  denote their transposes, which are row vectors. However, when defining the components of a vector in the text, we will usually define  $\mathbf{p}$  more casually as  $\mathbf{p} \equiv (p_1, \dots, p_N)$ .

**Example 1** The *ball of radius 1* in  $\mathbb{R}^N$  is a convex set; i.e., the following set  $B$  is convex:

$$B \equiv \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^N; (\mathbf{x}^T \mathbf{x})^{1/2} \leq 1\}. \quad (2.2)$$

To show that a set is convex, we need only take two arbitrary points that belong to the set, pick an arbitrary number  $\lambda$  between 0 and 1, and show that the point  $\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2$  also belongs to the set. Thus let  $\mathbf{x}^1 \in B, \mathbf{x}^2 \in B$  and let  $\lambda$  be such that  $0 < \lambda < 1$ . Since  $\mathbf{x}^1 \in B, \mathbf{x}^2 \in B$ , we have upon squaring, that  $\mathbf{x}^1$  and  $\mathbf{x}^2$  satisfy the following inequalities:

$$\mathbf{x}^{1T} \mathbf{x}^1 \leq 1; \quad \mathbf{x}^{2T} \mathbf{x}^2 \leq 1. \quad (2.3)$$

We need to show that:

$$(\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2)^T (\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) \leq 1. \quad (2.4)$$

Start off with the left hand side of (2.4) and expand out the terms:

$$\begin{aligned} (\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2)^T (\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) &= \lambda^2 \mathbf{x}^{1T} \mathbf{x}^1 + 2\lambda(1 - \lambda) \mathbf{x}^{1T} \mathbf{x}^2 + (1 - \lambda)^2 \mathbf{x}^{2T} \mathbf{x}^2 \\ &\leq \lambda^2 + 2\lambda(1 - \lambda) \mathbf{x}^{1T} \mathbf{x}^2 + (1 - \lambda)^2 \\ &\quad \text{where we have used (2.3) and } \lambda^2 > 0 \text{ and } (1 - \lambda)^2 > 0 \\ &\leq \lambda^2 + 2\lambda(1 - \lambda) (\mathbf{x}^{1T} \mathbf{x}^1)^{1/2} (\mathbf{x}^{2T} \mathbf{x}^2)^{1/2} + (1 - \lambda)^2 \\ &\quad \text{using } \lambda(1 - \lambda) > 0 \text{ and the Cauchy Schwarz inequality} \\ &\leq \lambda^2 + 2\lambda(1 - \lambda) + (1 - \lambda)^2 \quad \text{using (2.3)} \\ &= [\lambda + (1 - \lambda)]^2 \\ &= 1 \end{aligned} \quad (2.5)$$

which establishes the desired inequality (2.4).

The above example illustrates an important point: in order to prove that a set is convex, it is often necessary to verify that a certain *inequality* is true. Thus when studying convexity, it is useful to know some of the most frequently occurring inequalities, such as the Cauchy Schwarz inequality and the Theorem of the Arithmetic and Geometric Mean.

**Example 2** Let  $\mathbf{b} \in \mathbb{R}^N$  (i.e., let  $\mathbf{b}$  be an  $N$  dimensional vector) and let  $b_0$  be a scalar. Define the set

$$S \equiv \{\mathbf{x} : \mathbf{b}^T \mathbf{x} = b_0\}. \quad (2.6)$$

If  $N = 2$ , the set  $S$  is *straight line*, if  $N = 3$ , the set  $S$  is a *plane* and for  $N > 3$ , the set  $S$  is called a *hyperplane*. Show that  $S$  is a convex set.

Let  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S$  and let  $\lambda$  be such that  $0 < \lambda < 1$ . Since  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S$ , we have using definition (2.6) that

$$\mathbf{b}^T \mathbf{x}^1 = b_0; \quad \mathbf{b}^T \mathbf{x}^2 = b_0. \quad (2.7)$$

We use the relations (2.7) in (2.8) below. Thus

$$\begin{aligned} b_0 &= \lambda b_0 + (1 - \lambda) b_0 \\ &= \lambda \mathbf{b}^T \mathbf{x}^1 + (1 - \lambda) \mathbf{b}^T \mathbf{x}^2 \quad \text{using (2.7)} \\ &= \mathbf{b}^T [\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2] \quad \text{rearranging terms} \end{aligned} \quad (2.8)$$

and thus using definition (2.6),  $[\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2] \in S$ . Thus  $S$  is a convex set.

**Example 3** Let  $\mathbf{b} \in \mathbb{R}^N$  and let  $b_0$  be a scalar. Define a *halfspace*  $H$  as follows:

$$H \equiv \{\mathbf{x} : \mathbf{b}^T \mathbf{x} \leq b_0\}. \quad (2.9)$$

A halfspace is equal to a hyperplane plus the set which lies on one side of the hyperplane. It is easy to prove that a halfspace is a convex set: the proof is analogous to the proof used in Example 2 above except that now we also have to use the fact that  $\lambda > 0$  and  $1 - \lambda > 0$ .

**Example 4** Let  $S^j$  be a convex set in  $\mathbb{R}^N$  for  $j = 1, \dots, J$ . Then assuming that the intersection of the  $S^j$  is a nonempty set, this *intersection set* is also a convex set; i.e.,

$$S \equiv \bigcap_{j=1}^J S^j \quad (2.10)$$

is a convex set.

To prove this, let  $\mathbf{x}^1 \in \bigcap_{j=1}^J S^j$  and let  $\mathbf{x}^2 \in \bigcap_{j=1}^J S^j$  and let  $0 < \lambda < 1$ . Since  $S^j$  is convex for each  $j$ ,  $[\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2] \in S^j$  for each  $j$ . Therefore  $[\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2] \in \bigcap_{j=1}^J S^j$ . Hence  $S$  is also a convex set.

**Example 5** The *feasible region* for a linear programming problem is the following set  $S$ :

$$S \equiv \{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N\} \quad (2.11)$$

where  $\mathbf{A}$  is an  $M \times N$  matrix of constants and  $\mathbf{b}$  is an  $M$  dimensional vector of constants.

It is easy to show that the set  $S$  defined by (2.11) is a convex set, since the set  $S$  can be written as follows:

$$S = \left[ \bigcap_{m=1}^M H^m \right] \cap \Omega \quad (2.12)$$

where  $H^m \equiv \{\mathbf{x} : \mathbf{A}_m \cdot \mathbf{x} \leq b_m\}$  for  $m = 1, \dots, M$  is a halfspace ( $\mathbf{A}_m$  is the  $m$ th row of the matrix  $\mathbf{A}$  and  $b_m$  is the  $m$ th component of the column vector  $\mathbf{b}$ ) and  $\Omega \equiv \{\mathbf{x} : \mathbf{x} \geq \mathbf{0}_N\}$  is the nonnegative orthant in  $N$  dimensional space. Thus  $S$  is equal to the intersection of  $M + 1$  convex sets and hence using the result in Example 4, is a convex set.

**Example 6** Convex sets occur in economics frequently. For example, if a consumer is maximizing a utility function  $f(\mathbf{x})$  subject to a budget constraint, then we usually assume that the upper level sets of the function are convex sets; i.e., for every utility level  $\mathbf{u}$  in the range of  $f$ , we usually assume that the *upper level set*  $L(\mathbf{u}) \equiv \{\mathbf{x} : f(\mathbf{x}) \geq \mathbf{u}\}$  is convex. This upper level set consists of the consumer's indifference curve (or surface if  $N > 2$ ) and the set of  $\mathbf{x}$ 's lying above it.

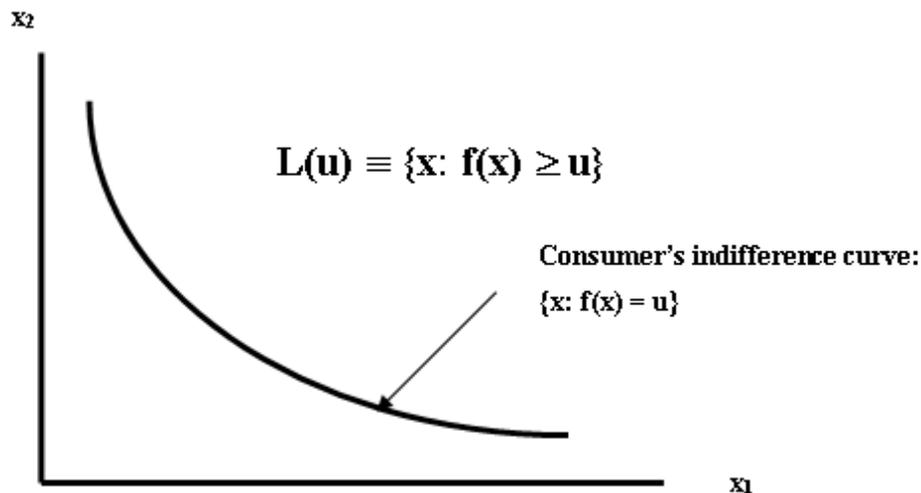


Fig. 2.1 A Consumer's Upper Level Set  $L(\mathbf{u})$

It is also generally assumed that production functions  $f(\mathbf{x})$  have the property that the set of inputs  $\mathbf{x}$  that can produce at least the output level  $\mathbf{y}$  (this is the upper level set  $L(\mathbf{y}) \equiv \{\mathbf{x} : f(\mathbf{x}) \geq \mathbf{y}\}$ ) is a convex set for every output level  $\mathbf{y}$  that belongs to the range of  $f$ . Functions which have this property are called *quasiconcave*. We will study these functions later in this chapter.

### 2.3 The Supporting Hyperplane Theorem for Closed Convex Sets

The results in this section are the key to duality theory in economics as we shall see later.

**Definition** A point  $\mathbf{x}^0 \in S$  is an *interior point* of  $S$  iff there exists  $\delta > 0$  such that the open ball of radius  $\delta$  around the point  $\mathbf{x}^0$ ,  $B_\delta(\mathbf{x}^0) \equiv \{\mathbf{x} : (\mathbf{x} - \mathbf{x}^0)^T(\mathbf{x} - \mathbf{x}^0) < \delta^2\}$ , also belongs to  $S$ ; i.e.,  $B_\delta(\mathbf{x}^0) \subset S$ .

**Definition** A set  $S$  is *open* iff it consists entirely of interior points.

**Definition** A set  $S$  is *closed* iff for every sequence of points,  $\{\mathbf{x}^n : n = 1, 2, \dots\}$  such that  $\mathbf{x}^n \in S$  for every  $n$  and  $\lim_{n \rightarrow \infty} \mathbf{x}^n \equiv \mathbf{x}^0$  exists, then  $\mathbf{x}^0 \in S$  as well.

**Definition** The *closure* of a set  $S$ ,  $\text{Clo } S$ , is defined as the set  $\{\mathbf{x} : \mathbf{x} = \lim_{n \rightarrow \infty} \mathbf{x}^n, \mathbf{x}^n \in S, n = 1, 2, \dots\}$ . Thus the closure of  $S$  is the set of points belonging to  $S$  plus any additional points that are limiting points of a sequence of points that belong to  $S$ . Note that for every set  $S$ ,  $\text{Clo } S$  is a closed set and  $S \subset \text{Clo } S$ .

**Definition**  $\mathbf{x}^0$  is a *boundary point* of  $S$  iff  $\mathbf{x}^0 \in \text{Clo } S$  but  $\mathbf{x}^0$  is not an interior point of  $S$ .

**Definition**  $\text{Int } S$  is defined to be the *set of interior points* of  $S$ . If there are no interior points of  $S$ , then  $\text{Int } S \equiv \emptyset$ , the empty set.

**Theorem 1** *Minkowski's (1911)[312] Theorem:* Let  $S$  be a closed convex set in  $\mathbb{R}^N$  and let  $\mathbf{b}$  be a point which does not belong to  $S$ . Then there exists a nonzero vector  $\mathbf{c}$  such that

$$\mathbf{c}^T \mathbf{b} < \min_{\mathbf{x}} \{\mathbf{c}^T \mathbf{x} : \mathbf{x} \in S\}; \quad (2.13)$$

i.e., there exists a hyperplane passing through the point  $\mathbf{b}$  which lies entirely below the convex set  $S$ .

**Proof.** Since  $(\mathbf{x} - \mathbf{b})^T(\mathbf{x} - \mathbf{b}) \geq 0$  for all vectors  $\mathbf{x}$ , it can be seen that  $\min_{\mathbf{x}}\{(\mathbf{x} - \mathbf{b})^T(\mathbf{x} - \mathbf{b}) : \mathbf{x} \in S\}$  exists. Let  $\mathbf{x}^0$  be a boundary point of  $S$  which attains this minimum; i.e.,

$$(\mathbf{x}^0 - \mathbf{b})^T(\mathbf{x}^0 - \mathbf{b}) = \min_{\mathbf{x}}\{(\mathbf{x} - \mathbf{b})^T(\mathbf{x} - \mathbf{b}) : \mathbf{x} \in S\}. \quad (2.14)$$

Now pick an arbitrary  $\mathbf{x} \in S$  and let  $0 < \lambda < 1$ . Since both  $\mathbf{x}$  and  $\mathbf{x}^0$  belong to  $S$ , by the convexity of  $S$ ,  $\lambda\mathbf{x} + (1 - \lambda)\mathbf{x}^0 \in S$ . Now use (2.14) to conclude that

$$\begin{aligned} (\mathbf{x}^0 - \mathbf{b})^T(\mathbf{x}^0 - \mathbf{b}) &\leq ([\lambda\mathbf{x} + (1 - \lambda)\mathbf{x}^0] - \mathbf{b})^T([\lambda\mathbf{x} + (1 - \lambda)\mathbf{x}^0] - \mathbf{b}) \\ &= (\mathbf{x}^0 - \mathbf{b} + \lambda[\mathbf{x} - \mathbf{x}^0])^T(\mathbf{x}^0 - \mathbf{b} + \lambda[\mathbf{x} - \mathbf{x}^0]) \\ &= (\mathbf{x}^0 - \mathbf{b})^T(\mathbf{x}^0 - \mathbf{b}) + 2\lambda(\mathbf{x}^0 - \mathbf{b})^T(\mathbf{x} - \mathbf{x}^0) + \lambda^2(\mathbf{x} - \mathbf{x}^0)^T(\mathbf{x} - \mathbf{x}^0). \end{aligned} \quad (2.15)$$

Define the vector  $\mathbf{c}$  as

$$\mathbf{c} \equiv \mathbf{x}^0 - \mathbf{b}. \quad (2.16)$$

If  $\mathbf{c}$  were equal to  $\mathbf{0}_N$ , then we would have  $\mathbf{b} = \mathbf{x}^0$ . But this is impossible because  $\mathbf{x}^0 \in S$  and  $\mathbf{b}$  was assumed to be exterior to  $S$ . Hence

$$\mathbf{c} \neq \mathbf{0}_N. \quad (2.17)$$

The inequality (2.17) in turn implies that

$$0 < \mathbf{c}^T \mathbf{c} = (\mathbf{x}^0 - \mathbf{b})^T(\mathbf{x}^0 - \mathbf{b}) = \mathbf{c}^T(\mathbf{x}^0 - \mathbf{b}) \quad \text{or} \quad (2.18)$$

$$\mathbf{c}^T \mathbf{b} < \mathbf{c}^T \mathbf{x}^0. \quad (2.19)$$

Rearranging terms in the inequality (2.15) leads to the following inequality that holds for all  $\mathbf{x} \in S$  and  $0 < \lambda < 1$ :

$$0 \leq 2\lambda(\mathbf{x}^0 - \mathbf{b})^T(\mathbf{x} - \mathbf{x}^0) + \lambda^2(\mathbf{x} - \mathbf{x}^0)^T(\mathbf{x} - \mathbf{x}^0). \quad (2.20)$$

Divide both sides of (2.20) by  $2\lambda > 0$  and take the limit of the resulting inequality as  $\lambda$  tends to 0 in order to obtain the following inequality:

$$\begin{aligned} 0 &\leq (\mathbf{x}^0 - \mathbf{b})^T(\mathbf{x} - \mathbf{x}^0) \quad \text{for all } \mathbf{x} \in S \\ &= \mathbf{c}^T(\mathbf{x} - \mathbf{x}^0) \quad \text{for all } \mathbf{x} \in S \text{ using definition (2.16) or} \end{aligned} \quad (2.21)$$

$$\mathbf{c}^T \mathbf{x}^0 \leq \mathbf{c}^T \mathbf{x} \quad \text{for all } \mathbf{x} \in S. \quad (2.22)$$

Putting (2.22) and (2.19) together, we obtain the following inequalities, which are equivalent to the desired result, (2.13):

$$\mathbf{c}^T \mathbf{b} < \mathbf{c}^T \mathbf{x}^0 \leq \mathbf{c}^T \mathbf{x} \quad \text{for all } \mathbf{x} \in S. \quad (2.23)$$

■

**Theorem 2** *Minkowski's Supporting Hyperplane Theorem:* Let  $S$  be a convex set and let  $\mathbf{b}$  be a boundary point of  $S$ . Then there exists a hyperplane through  $\mathbf{b}$  which supports  $S$ ; i.e., there exists a nonzero vector  $\mathbf{c}$  such that

$$\mathbf{c}^T \mathbf{b} = \min_{\mathbf{x}}\{\mathbf{c}^T \mathbf{x} : \mathbf{x} \in \text{Clo } S\}. \quad (2.24)$$

**Proof.** Let  $\mathbf{b}^n \notin \text{Clo } S$  for  $n = 1, 2, \dots$  but let the limit  $\lim_{n \rightarrow \infty} \mathbf{b}^n = \mathbf{b}$ ; i.e., each member of the sequence of points  $\mathbf{b}^n$  is exterior to the closure of  $S$  but the limit of the points  $\mathbf{b}^n$  is the boundary point  $\mathbf{b}$ . By Minkowski's Theorem, there exists a sequence of nonzero vectors  $\mathbf{c}^n$  such that

$$\mathbf{c}^{nT} \mathbf{b}^n < \min_{\mathbf{x}}\{\mathbf{c}^{nT} \mathbf{x} : \mathbf{x} \in \text{Clo } S\}; \quad n = 1, 2, \dots \quad (2.25)$$

There is no loss of generality in normalizing the vectors  $\mathbf{c}^n$  so that they are of unit length. Thus we can assume that  $\mathbf{c}^{nT} \mathbf{c}^n = 1$  for every  $n$ . By a Theorem in analysis due to Weierstrass,<sup>\*2</sup> there

<sup>\*2</sup> See Rudin (1953; 31)[337].

exists a subsequence of the points  $\{\mathbf{c}^n\}$  which tends to a limit, which we denote by  $\mathbf{c}$ . Along this subsequence, we will have  $\mathbf{c}^{nT}\mathbf{c}^n = 1$  and so the limiting vector  $\mathbf{c}$  will also have this property so that  $\mathbf{c} \neq \mathbf{0}_N$ . For each  $\mathbf{c}^n$  in the subsequence, (2.25) will be true so that we have

$$\mathbf{c}^{nT}\mathbf{b}^n < \mathbf{c}^{nT}\mathbf{x}; \quad \text{for every } \mathbf{x} \in S. \quad (2.26)$$

Thus

$$\begin{aligned} \mathbf{c}^T\mathbf{b} &= \lim_{n \rightarrow \infty} \text{along the subsequence } \mathbf{c}^{nT}\mathbf{b}^n \\ &\leq \lim_{n \rightarrow \infty} \text{along the subsequence } \mathbf{c}^{nT}\mathbf{x} \quad \text{for every } \mathbf{x} \in S \text{ using (2.26)} \\ &= \mathbf{c}^T\mathbf{x} \quad \text{for every } \mathbf{x} \in S \end{aligned} \quad (2.27)$$

which is equivalent to the desired result (2.24). ■

The halfspace  $\{\mathbf{x} : \mathbf{c}^T\mathbf{b} \leq \mathbf{c}^T\mathbf{x}\}$  where  $\mathbf{b}$  is a boundary point of  $S$  and  $\mathbf{c}$  was defined in the above Theorem is called a *supporting halfspace* to the convex set  $S$  at the boundary point  $\mathbf{b}$ .

**Theorem 3** *Minkowski's Theorem Characterizing Closed Convex Sets:* Let  $S$  be a closed convex set that is not equal to  $\mathbb{R}^N$ . Then  $S$  is equal to the intersection of its supporting halfspaces.

**Proof.** If  $S$  is closed and convex and not the entire space  $\mathbb{R}^N$ , then by the previous result, it is clear that  $S$  is contained in each of its supporting halfspaces and hence is a subset of the intersection of its supporting halfspaces. Now let  $\mathbf{x} \notin S$  so that  $\mathbf{x}$  is exterior to  $S$ . Then using the previous two Theorems, it is easy to see that  $\mathbf{x}$  does not belong to at least one supporting halfspace to  $S$  and thus  $\mathbf{x}$  does not belong to the intersection of the supporting halfspaces to  $S$ . ■

**Problem 1** Let  $\mathbf{A} = \mathbf{A}^T$  be a positive definite  $N \times N$  symmetric matrix. Let  $\mathbf{x}$  and  $\mathbf{y}$  be  $N$  dimensional vectors. Show that the following generalization of the Cauchy Schwarz inequality holds:

$$(\mathbf{x}^T\mathbf{A}\mathbf{y})^2 \leq (\mathbf{x}^T\mathbf{A}\mathbf{x})(\mathbf{y}^T\mathbf{A}\mathbf{y}). \quad (a)$$

*Hint:* You may find the concept of a *square root matrix* for a positive definite matrix helpful. From matrix algebra, we know that every symmetric matrix has the following eigenvalue-eigenvector decomposition with the following properties: there exist  $N \times N$  matrices  $\mathbf{U}$  and  $\mathbf{\Lambda}$  such that

$$\mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{\Lambda}; \quad (b)$$

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_N \quad (c)$$

where  $\mathbf{\Lambda}$  is a diagonal matrix with the eigenvalues of  $\mathbf{A}$  on the main diagonal and  $\mathbf{U}$  is an orthonormal matrix. Note that  $\mathbf{U}$  is the inverse of  $\mathbf{U}^T$ . Hence premultiply both sides of (b) by  $\mathbf{U}$  and postmultiply both sides of (b) by  $\mathbf{U}^T$  in order to obtain the following equation:

$$\begin{aligned} \mathbf{A} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{U}^T \quad \text{where we use the assumption that } \mathbf{A} \text{ is positive definite and we} \\ &\quad \text{define } \mathbf{\Lambda}^{1/2} \text{ to be a diagonal matrix with diagonal elements equal to} \\ &\quad \text{the positive square roots of the diagonal elements of } \mathbf{\Lambda} \text{ (which are} \\ &\quad \text{the positive eigenvalues of } \mathbf{A}, \lambda_1, \dots, \lambda_N. \\ &= \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^T \quad \text{using (c)} \\ &= \mathbf{B}^T\mathbf{B} \end{aligned} \quad (d)$$

where the  $N \times N$  square root matrix  $\mathbf{B}$  is defined as

$$\mathbf{B} \equiv \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^T. \quad (\text{e})$$

Note that  $\mathbf{B}$  is symmetric so that

$$\mathbf{B} = \mathbf{B}^T \quad (\text{f})$$

and thus we can also write  $\mathbf{A}$  as

$$\mathbf{A} = \mathbf{B}\mathbf{B}. \quad (\text{g})$$

**Problem 2** Let  $\mathbf{A} = \mathbf{A}^T$  be a positive definite  $N \times N$  symmetric matrix. Let  $\mathbf{x}$  and  $\mathbf{y}$  be  $N$  dimensional vectors. Show that the following generalization of the Cauchy Schwarz inequality holds:

$$(\mathbf{x}^T \mathbf{y})^2 \leq (\mathbf{x}^T \mathbf{A} \mathbf{x})(\mathbf{y}^T \mathbf{A}^{-1} \mathbf{y}).$$

**Problem 3** Let  $\mathbf{A} = \mathbf{A}^T$  be a positive definite symmetric matrix. Define the set  $S \equiv \{\mathbf{x} : \mathbf{x}^T \mathbf{A} \mathbf{x} \leq 1\}$ . Show that  $S$  is a convex set.

**Problem 4** A set  $S$  in  $\mathbb{R}^N$  is a *cone* iff it has the following property:

$$\text{(a) } \mathbf{x} \in S, \lambda \geq 0 \text{ implies } \lambda \mathbf{x} \in S.$$

A set  $S$  in  $\mathbb{R}^N$  is a convex *cone* iff it both a convex set and a cone. Show that  $S$  is a convex cone iff it satisfies the following property:

$$\text{(b) } \mathbf{x} \in S, \mathbf{y} \in S, \alpha \geq 0 \text{ and } \beta \geq 0 \text{ implies } \alpha \mathbf{x} + \beta \mathbf{y} \in S.$$

**Problem 5** Let  $S$  be a nonempty closed convex set in  $\mathbb{R}^N$  (that is not the entire space  $\mathbb{R}^N$ ) and let  $\mathbf{b}$  be a boundary point of  $S$ . Using Minkowski's supporting hyperplane theorem, there exists at least one vector  $\mathbf{c}^* \neq \mathbf{0}_N$  such that  $\mathbf{c}^{*T} \mathbf{b} \leq \mathbf{c}^{*T} \mathbf{x}$  for all  $\mathbf{x} \in S$ . Define the *set of supporting hyperplanes*  $S(\mathbf{b})$  to the set  $S$  at the boundary point  $\mathbf{b}$  to be the following set:

$$\text{(a) } S(\mathbf{b}) \equiv \{c : \mathbf{c}^T \mathbf{b} \leq \mathbf{c}^T \mathbf{x} \text{ for all } \mathbf{x} \in S\}.$$

Show that the set  $S(\mathbf{b})$  has the following properties:

- (b)  $\mathbf{0}_N \in S(\mathbf{b})$ ;
- (c)  $S(\mathbf{b})$  is a cone;
- (d)  $S(\mathbf{b})$  is a closed set;
- (e)  $S(\mathbf{b})$  is a convex set and
- (f)  $S(\mathbf{b})$  contains at least one *ray*, a set of the form  $\{\mathbf{x} : \mathbf{x} = \lambda \mathbf{x}^*, \lambda \geq 0\}$  where  $\mathbf{x}^* \neq \mathbf{0}_N$ .

**Problem 6** If  $X$  and  $Y$  are nonempty sets in  $\mathbb{R}^N$ , the set  $X - Y$  is defined as follows:

$$X - Y \equiv \{\mathbf{x} - \mathbf{y} : \mathbf{x} \in X \text{ and } \mathbf{y} \in Y\}. \quad (\text{a})$$

If  $X$  and  $Y$  are nonempty convex sets, show that  $X - Y$  is also a convex set.

**Problem 7** *Separating hyperplane theorem between two disjoint convex sets* Fenchel (1953; 48-49)[179]: Let  $X$  and  $Y$  be two nonempty, convex sets in  $\mathbb{R}^N$  that have no points in common; i.e.,  $X \cap Y = \emptyset$  (the empty set). Assume that at least one of the two sets  $X$  or  $Y$  has a nonempty interior. Prove that there exists a hyperplane that separates  $X$  and  $Y$ ; i.e., show that there exists a nonzero vector  $\mathbf{c}$  and a scalar  $\alpha$  such that

$$\mathbf{c}^T \mathbf{y} \leq \alpha \leq \mathbf{c}^T \mathbf{x} \quad \text{for all } \mathbf{x} \in X \text{ and } \mathbf{y} \in Y. \quad (\text{a})$$

*Hint:* Consider the set  $S \equiv X - Y$  and show that  $\mathbf{0}_N$  does not belong to  $S$ . If  $\mathbf{0}_N$  does not belong to the closure of  $S$ , apply Minkowski's Theorem 1. If  $\mathbf{0}_N$  does belong to the closure of  $S$ , then since it does not belong to  $S$  and  $S$  has an interior, it must be a boundary point of  $S$  and apply Minkowski's Theorem 2.

## 2.4 Concave Functions

**Definition** A function  $f(\mathbf{x})$  of  $N$  variables  $\mathbf{x} \equiv [x_1, \dots, x_N]$  defined over a convex subset  $S$  of  $\mathbb{R}^N$  is *concave* iff for every  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S$  and  $0 < \lambda < 1$ , we have

$$f(\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) \geq \lambda f(\mathbf{x}^1) + (1 - \lambda) f(\mathbf{x}^2). \quad (2.28)$$

In the above definition,  $f$  is defined over a convex set so that we can be certain that points of the form  $\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2$  belong to  $S$  if  $\mathbf{x}^1$  and  $\mathbf{x}^2$  belong to  $S$ . Note that when  $\lambda$  equals 0 or 1, the weak inequality (2.28) is automatically valid as an equality so that in the above definition, we could replace the restriction  $0 < \lambda < 1$  by  $0 \leq \lambda \leq 1$ .

If  $N = 1$ , a geometric interpretation of a concave function is easy to obtain; see Figure 2.2 below. As the scalar  $\lambda$  travels from 1 to 0,  $f(\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2)$  traces out the value of  $f$  between  $\mathbf{x}^1$  and  $\mathbf{x}^2$ . On the other hand, as  $\lambda$  travels from 1 to 0,  $\lambda f(\mathbf{x}^1) + (1 - \lambda) f(\mathbf{x}^2)$  is a linear function of  $\lambda$  which joins up the point  $(\mathbf{x}^1, f(\mathbf{x}^1))$  to the point  $(\mathbf{x}^2, f(\mathbf{x}^2))$  on the graph of  $f(\mathbf{x})$ ; i.e.,  $\lambda f(\mathbf{x}^1) + (1 - \lambda) f(\mathbf{x}^2)$  traces out the chord between two points on the graph of  $f$ . The inequality (2.28) says that if the function is concave, then the graph of  $f$  between  $\mathbf{x}^1$  and  $\mathbf{x}^2$  will lie above (or possibly be coincident with) the chord joining these two points on the graph. This property must hold for any two points on the graph of  $f$ .

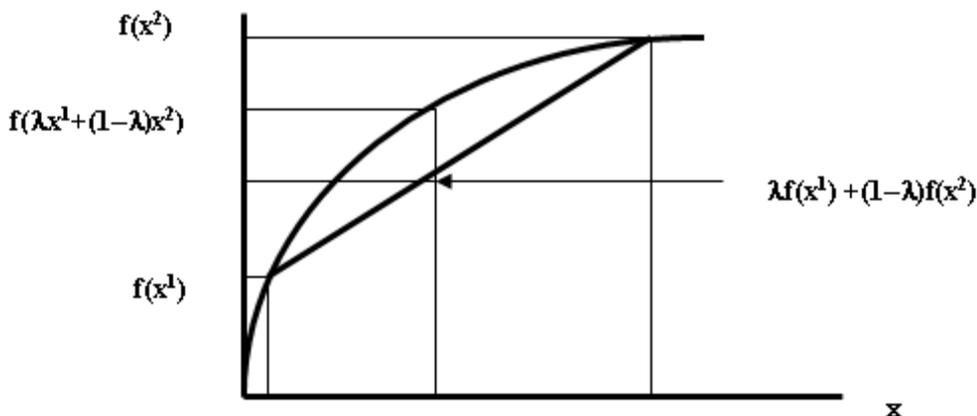


Fig. 2.2 A Concave Function of One Variable  $f(\mathbf{x})$

For a general  $N$ , the interpretation of the concavity property is the same as in the previous paragraph: we look at the behavior of the function along the line segment joining  $\mathbf{x}^1$  to  $\mathbf{x}^2$  compared to the straight line segment joining the point  $[\mathbf{x}^1, f(\mathbf{x}^1)]$  in  $\mathbb{R}^{N+1}$  to the point  $[\mathbf{x}^2, f(\mathbf{x}^2)]$ . This line segment must lie below (or be coincident with) the former curve. This property must hold for any two points in the domain of definition of  $f$ .

Concave functions occur quite frequently in economic as we shall see in subsequent chapters.

One very convenient property that a concave function possesses is given by the following result.

**Theorem 4** *Local Maximum is a Global Maximum*; Fenchel (1953; 63)[179]: Let  $f$  be a concave function defined over a convex subset  $S$  of  $\mathbb{R}^N$ . If  $f$  attains a local maximum at the point  $\mathbf{x}^0 \in S$ , then  $f$  attains a global maximum at  $\mathbf{x}^0$ ; i.e., we have

$$f(\mathbf{x}^0) \geq f(\mathbf{x}) \text{ for all } \mathbf{x} \in S. \quad (2.29)$$

**Proof.** Since  $f$  attains a local maximum at  $\mathbf{x}^0$ , there exists a  $\delta > 0$  such that

$$f(\mathbf{x}^0) \geq f(\mathbf{x}) \text{ for all } \mathbf{x} \in S \cap B(\mathbf{x}^0, \delta) \quad (2.30)$$

where  $B(\mathbf{x}^0, \delta) \equiv \{\mathbf{x} : (\mathbf{x} - \mathbf{x}^0)^T(\mathbf{x} - \mathbf{x}^0) < \delta^2\}$  is the open ball of radius  $\delta$  around the point  $\mathbf{x}^0$ . Suppose there exists an  $\mathbf{x}^1 \in S$  such that

$$f(\mathbf{x}^1) > f(\mathbf{x}^0). \quad (2.31)$$

Using the concavity of  $f$ , for  $0 < \lambda < 1$ , we have

$$\begin{aligned} f(\lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^0) &\geq \lambda f(\mathbf{x}^1) + (1-\lambda)f(\mathbf{x}^0) \\ &\geq \lambda f(\mathbf{x}^0) + (1-\lambda)f(\mathbf{x}^0) \quad \text{using } \lambda > 0 \text{ and (2.31)} \\ &= f(\mathbf{x}^0). \end{aligned} \quad (2.32)$$

But for  $\lambda$  close to 0,  $\lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^0$  will belong to  $S \cap B(\mathbf{x}^0, \delta)$  and hence for  $\lambda$  close to 0, (2.32) will contradict (2.30). Thus our *supposition* must be false and (2.29) holds. ■

It turns out to be very useful to have several alternative characterizations for the concavity property. Our first characterization is provided by the definition (2.28).

**Theorem 5** *Second Characterization of Concavity*; Fenchel (1953; 57)[179]: (a)  $f$  is a concave function defined over the convex subset  $S$  of  $\mathbb{R}^N$  iff (b) the set  $H \equiv \{(y, \mathbf{x}) : y \leq f(\mathbf{x}), \mathbf{x} \in S\}$  is a convex set in  $\mathbb{R}^{N+1}$ .<sup>\*3</sup>

**Proof.** (a) implies (b): Let  $(y^1, \mathbf{x}^1) \in H, (y^2, \mathbf{x}^2) \in H$  and  $0 < \lambda < 1$ . Thus

$$y^1 \leq f(\mathbf{x}^1) \text{ and } y^2 \leq f(\mathbf{x}^2) \quad (2.33)$$

with  $\mathbf{x}^1 \in S$  and  $\mathbf{x}^2 \in S$ . Since  $f$  is concave over  $S$ ,

$$\begin{aligned} f(\lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^2) &\geq \lambda f(\mathbf{x}^1) + (1-\lambda)f(\mathbf{x}^2) \\ &\geq \lambda y^1 + (1-\lambda)y^2 \quad \text{using } \lambda > 0, (1-\lambda) > 0 \text{ and (2.33)}. \end{aligned} \quad (2.34)$$

Using the definition of  $S$ , (2.34) shows that  $[\lambda y^1 + (1-\lambda)y^2, \lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^2] \in H$ . Thus  $H$  is a convex set.

(b) implies (a): Let  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S$  and  $0 < \lambda < 1$ . Define  $y^1$  and  $y^2$  as follows:

$$y^1 \equiv f(\mathbf{x}^1) \text{ and } y^2 \equiv f(\mathbf{x}^2). \quad (2.35)$$

The definition of  $H$  and the equalities in (2.35) show that  $(y^1, \mathbf{x}^1) \in H$  and  $(y^2, \mathbf{x}^2) \in H$ . Since  $H$  is a convex set by assumption,  $[\lambda y^1 + (1-\lambda)y^2, \lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^2] \in H$ . Hence, by the definition of  $H$ ,

$$\begin{aligned} f(\lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^2) &\geq \lambda y^1 + (1-\lambda)y^2 \\ &= \lambda f(\mathbf{x}^1) + (1-\lambda)f(\mathbf{x}^2) \quad \text{using (2.35)} \end{aligned} \quad (2.36)$$

---

<sup>\*3</sup> The set  $H$  is called the *hypograph* of the function  $f$ ; it consists of the graph of  $f$  and all of the points lying below it.

which establishes the concavity of  $f$ . ■

A geometric interpretation of the second characterization of concavity when  $N = 1$  is illustrated in Figure 2.3.

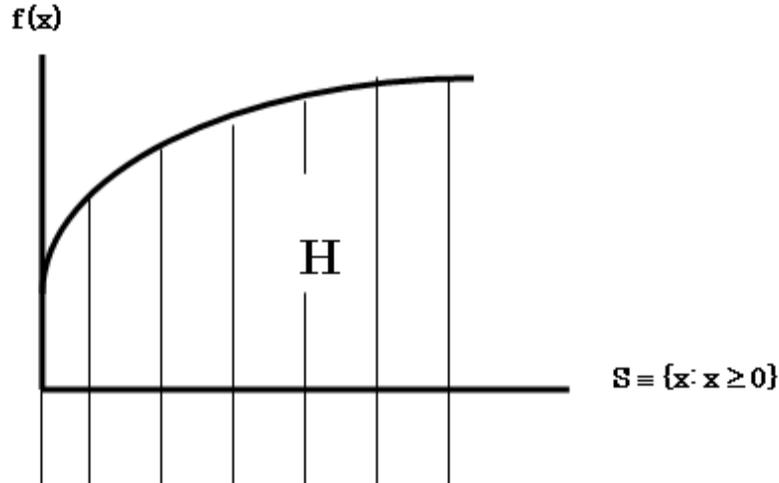


Fig. 2.3 The Second Characterization of Concavity

The second characterization of concavity is useful because it shows us why concave functions are *continuous* over the interior of their domain of definition.

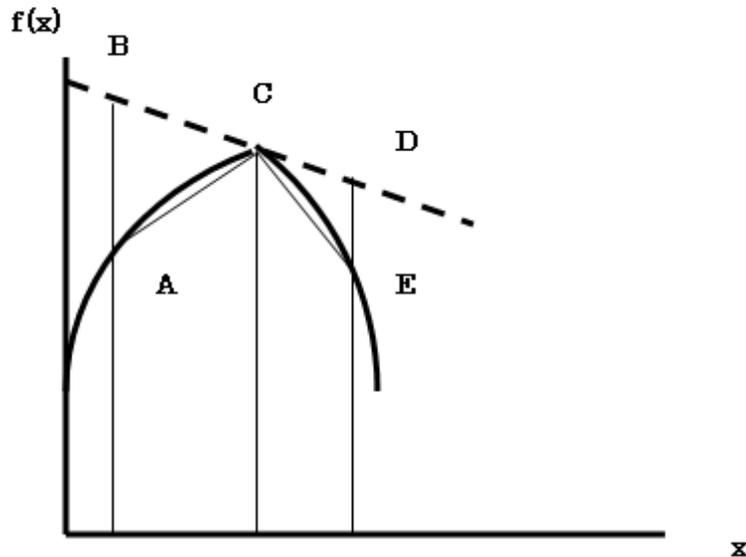


Fig. 2.4 The Continuity of a Concave Function

Let the function of one variable,  $f(\mathbf{x})$ , be concave. In Figure 2.4, the point  $\mathbf{x}^0$  is in the interior of the domain of definition set  $S$  and the point  $[f(\mathbf{x}^0), \mathbf{x}^0]$  is on the boundary of the convex set hypograph  $H$  of  $f$ . By Theorem 2, there is at least one supporting hyperplane to the boundary point  $[f(\mathbf{x}^0), \mathbf{x}^0]$  of  $H$  (this is the point  $C$  in Figure 2.4) and we have drawn one of these hyperplanes

(which are lines in this case) as the dashed line BCD.\*<sup>4</sup> Note that this dashed line serves as an *upper bounding* approximating function to  $f(\mathbf{x})$ . Now consider some point  $\mathbf{x}^1$  to the left of  $\mathbf{x}^0$ . Since  $f$  is a concave function, it can be seen that the straight line joining the points  $[f(\mathbf{x}^1), \mathbf{x}^1]$  and  $[f(\mathbf{x}^0), \mathbf{x}^0]$ , the line segment AC in Figure 2.4, is a *lower bounding* approximating function to  $f(\mathbf{x})$  to the left of  $\mathbf{x}^0$ . Now consider some point  $\mathbf{x}^2$  to the right of  $\mathbf{x}^0$ . Again using the concavity of  $f$ , the straight line joining the points  $[f(\mathbf{x}^0), \mathbf{x}^0]$  and  $[f(\mathbf{x}^2), \mathbf{x}^2]$ , the line segment CE in Figure 2.4, is a *lower bounding* approximating function to  $f(\mathbf{x})$  to the right of  $\mathbf{x}^0$ . Thus  $f(\mathbf{x})$  is sandwiched between two linear functions that meet at the point C both to the left and to the right of  $\mathbf{x}^0$  and it can be seen that  $f(\mathbf{x})$  must be continuous at the interior point  $\mathbf{x}^0$ . The same conclusion can be derived for any interior point of the domain of definition set  $S$  and hence we conclude that a concave function  $f$  is continuous over  $\text{Int } S$ .\*<sup>5</sup>

The example shown in Figure 2.4 shows that there can be more than one supporting hyperplane to a boundary point on the hypograph of a concave function. Consider the following definition and Theorem:

**Definition** A vector  $\mathbf{b}$  is a *supergradient* to the function of  $N$  variables  $f$  defined over  $S$  at the point  $\mathbf{x}^0 \in S$  iff

$$f(\mathbf{x}) \leq f(\mathbf{x}^0) + \mathbf{b}^T(\mathbf{x} - \mathbf{x}^0) \quad \text{for all } \mathbf{x} \in S. \quad (2.37)$$

Note that the function on the right hand side of (2.37) is a linear function of  $\mathbf{x}$  which takes on the value  $f(\mathbf{x}^0)$  when  $\mathbf{x} = \mathbf{x}^0$ . This linear function acts as an upper bounding function to  $f$ .

**Theorem 6** Rockafellar (1970; 217)[335]: If  $f$  is a concave function defined over an open convex subset  $S$  of  $\mathbb{R}^N$ , then for every  $\mathbf{x}^0 \in S$ ,  $f$  has at least one supergradient vector  $\mathbf{b}^0$  to  $f$  at the point  $\mathbf{x}^0$ . Denote the set of all such supergradient vectors as  $\partial f(\mathbf{x}^0)$ . Then  $\partial f(\mathbf{x}^0)$  is a nonempty, closed convex set.\*<sup>6</sup>

**Proof.** Define the hypograph of  $f$  as  $H \equiv \{(y, \mathbf{x}) : y \leq f(\mathbf{x}), \mathbf{x} \in S\}$ . Since  $f$  is concave, by Theorem 5,  $H$  is a convex set. Note also that  $[f(\mathbf{x}^0), \mathbf{x}^0]$  is a boundary point of  $H$ . By Theorem 2, there exists  $[c_0, \mathbf{c}^T] \neq \mathbf{0}_{N+1}$  such that

$$c_0 f(\mathbf{x}^0) + \mathbf{c}^T \mathbf{x}^0 \leq c_0 y + \mathbf{c}^T \mathbf{x} \quad \text{for every } [y, \mathbf{x}] \in H. \quad (2.38)$$

Suppose  $c_0$  in (2.38) were equal to 0. Then (2.38) becomes

$$\mathbf{c}^T \mathbf{x}^0 \leq \mathbf{c}^T \mathbf{x} \quad \text{for every } \mathbf{x} \in S. \quad (2.39)$$

Since  $\mathbf{x}^0 \in \text{Int } S$ , (2.39) cannot be true. Hence our *supposition* is false and we can assume  $c_0 \neq 0$ . If  $c_0 > 0$ , then (2.38) is not satisfied if  $y < f(\mathbf{x}^0)$  and  $\mathbf{x} = \mathbf{x}^0$ . Thus we must have  $c_0 < 0$ . Multiplying (2.38) through by  $1/c_0$  yields the following inequality:

$$f(\mathbf{x}^0) - \mathbf{b}^{0T} \mathbf{x}^0 \geq y - \mathbf{b}^{0T} \mathbf{x} \quad \text{for every } [y, \mathbf{x}] \in H \quad (2.40)$$

where the vector  $\mathbf{b}^0$  is defined as

$$\mathbf{b}^0 \equiv -\mathbf{c}/c_0. \quad (2.41)$$

\*<sup>4</sup> Note that the graph of  $f$  is kinked at the point C and so there is an entire set of supporting hyperplanes to the point C in this case.

\*<sup>5</sup> The argument is a bit more complex when  $N$  is greater than 1 but the same conclusion is obtained. We cannot extend the above argument to boundary points of  $S$  because the supporting hyperplane to  $H$  may be vertical. See Fenchel (1953; 74)[179] for a general proof.

\*<sup>6</sup> Rockafellar shows that  $\partial f(\mathbf{x}^0)$  is also a bounded set.

Now let  $\mathbf{x} \in S$  and  $y = f(\mathbf{x})$ . Then  $[y, \mathbf{x}] \in H$  and (2.40) becomes

$$f(\mathbf{x}) \leq f(\mathbf{x}^0) + \mathbf{b}^{0T}(\mathbf{x} - \mathbf{x}^0) \quad \text{for all } \mathbf{x} \in S. \quad (2.42)$$

Using definition (2.37), (2.42) shows that  $\mathbf{b}^0$  is a supergradient to  $f$  at  $\mathbf{x}^0$  and hence  $\partial f(\mathbf{x}^0)$  is nonempty.

To show that  $\partial f(\mathbf{x}^0)$  is a convex set, let  $\mathbf{b}^1 \in \partial f(\mathbf{x}^0)$ ,  $\mathbf{b}^2 \in \partial f(\mathbf{x}^0)$  and  $0 < \lambda < 1$ . Then

$$\begin{aligned} f(\mathbf{x}) &\leq f(\mathbf{x}^0) + \mathbf{b}^{1T}(\mathbf{x} - \mathbf{x}^0) && \text{for all } \mathbf{x} \in S; \\ f(\mathbf{x}) &\leq f(\mathbf{x}^0) + \mathbf{b}^{2T}(\mathbf{x} - \mathbf{x}^0) && \text{for all } \mathbf{x} \in S. \end{aligned} \quad (2.43)$$

Thus

$$\begin{aligned} f(\mathbf{x}) &= \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{x}) && \text{for all } \mathbf{x} \in S; \\ &\leq \lambda[f(\mathbf{x}^0) + \mathbf{b}^{1T}(\mathbf{x} - \mathbf{x}^0)] + (1 - \lambda)f(\mathbf{x}) && \text{using } \lambda > 0 \text{ and (2.43)} \\ &\leq \lambda[f(\mathbf{x}^0) + \mathbf{b}^{1T}(\mathbf{x} - \mathbf{x}^0)] + (1 - \lambda)[f(\mathbf{x}^0) + \mathbf{b}^{2T}(\mathbf{x} - \mathbf{x}^0)] && \text{using } 1 - \lambda > 0 \text{ and (2.43)} \\ &= f(\mathbf{x}^0) + [\lambda\mathbf{b}^{1T} + (1 - \lambda)\mathbf{b}^{2T}](\mathbf{x} - \mathbf{x}^0) && \text{for all } \mathbf{x} \in S \end{aligned} \quad (2.44)$$

and so  $[\lambda\mathbf{b}^1 + (1 - \lambda)\mathbf{b}^2] \in \partial f(\mathbf{x}^0)$ . Thus  $\partial f(\mathbf{x}^0)$  is a convex set.

The closedness of  $\partial f(\mathbf{x}^0)$  follows from the fact that the vector  $\mathbf{b}$  enters the inequalities (2.37) in a linear fashion. ■

**Corollary** If  $f$  is a concave function defined over a convex subset  $S$  of  $\mathbb{R}^N$ ,  $\mathbf{x}^0 \in \text{Int } S$  and the first order partial derivatives of  $f$  evaluated at  $\mathbf{x}^0$  exist,<sup>\*7</sup> then  $\partial f(\mathbf{x}^0) = \{\nabla f(\mathbf{x}^0)\}$ ; i.e., if  $f$  has first order partial derivatives, then the set of supergradients reduces to the gradient vector of  $f$  evaluated at  $\mathbf{x}^0$ .

**Proof.** Since  $\mathbf{x}^0 \in \text{Int } S$ ,  $\partial f(\mathbf{x}^0)$  is nonempty. Let  $\mathbf{b} \in \partial f(\mathbf{x}^0)$ . Using definition (2.37) of a supergradient, it can be seen that the function  $g(\mathbf{x})$  defined by (2.45) below is nonpositive:

$$g(\mathbf{x}) \equiv f(\mathbf{x}) - f(\mathbf{x}^0) - \mathbf{b}^T(\mathbf{x} - \mathbf{x}^0) \leq 0 \quad \text{for all } \mathbf{x} \in S. \quad (2.45)$$

Since  $g(\mathbf{x}^0) = 0$ , (2.45) shows that  $g(\mathbf{x})$  attains a global maximum over the set  $S$  at  $\mathbf{x} = \mathbf{x}^0$ . Hence, the following first order necessary conditions for maximizing a function of  $N$  variables will hold at  $\mathbf{x}^0$ :

$$\nabla g(\mathbf{x}^0) = \nabla f(\mathbf{x}^0) - \mathbf{b} = \mathbf{0}_N \text{ or} \quad (2.46)$$

$$\mathbf{b} = \nabla f(\mathbf{x}^0). \quad (2.47)$$

Hence we have shown that if  $\mathbf{b} \in \partial f(\mathbf{x}^0)$ , then  $\mathbf{b} = \nabla f(\mathbf{x}^0)$ . Hence  $\partial f(\mathbf{x}^0)$  is the single point,  $\nabla f(\mathbf{x}^0)$ . ■

**Theorem 7 Third Characterization of Concavity:** Roberts and Varberg (1973; 12)[334]: Let  $f$  be a function of  $N$  variables defined over an open convex subset  $S$  of  $\mathbb{R}^N$ . Then (a)  $f$  is concave over  $S$  iff (b) for every  $\mathbf{x}^0 \in S$ , there exists a  $\mathbf{b}^0$  such that

$$f(\mathbf{x}) \leq f(\mathbf{x}^0) + \mathbf{b}^{0T}(\mathbf{x} - \mathbf{x}^0) \quad \text{for all } \mathbf{x} \in S. \quad (2.48)$$

<sup>\*7</sup> Thus the vector of first order partial derivatives  $\nabla f(\mathbf{x}^0) \equiv [\partial f(\mathbf{x}^0)/\partial x_1, \dots, \partial f(\mathbf{x}^0)/\partial x_N]^T$  exists.

**Proof.** (a) implies (b). This has been done in Theorem 6 above.

(b) implies (a). Let  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S$  and  $0 < \lambda < 1$ . Let  $\mathbf{b}$  be a supergradient to  $f$  at the point  $\lambda\mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2$ . Thus we have:

$$f(\mathbf{x}) \leq f(\lambda\mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2) + \mathbf{b}^T(\mathbf{x} - [\lambda\mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2]) \quad \text{for all } \mathbf{x} \in S. \quad (2.49)$$

Now evaluate (2.49) at  $\mathbf{x} = \mathbf{x}^1$  and then multiply both sides of the resulting inequality through by  $\lambda > 0$ . Evaluate (2.49) at  $\mathbf{x} = \mathbf{x}^2$  and then multiply both sides of the resulting inequality through by  $1 - \lambda > 0$ . Add the two inequalities to obtain:

$$\begin{aligned} \lambda f(\mathbf{x}^1) + (1 - \lambda)f(\mathbf{x}^2) &\leq \lambda f(\lambda\mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2) + \lambda\mathbf{b}^T(\mathbf{x}^1 - [\lambda\mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2]) \\ &\quad + (1 - \lambda)\mathbf{b}^T(\mathbf{x}^2 - [\lambda\mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2]) \\ &= f(\lambda\mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2) \end{aligned} \quad (2.50)$$

which shows that  $f$  is concave over  $S$ . ■

**Corollary** *Third Characterization of Concavity in the Once Differentiable Case*; Mangasarian (1969; 84)[302]: Let  $f$  be a once differentiable function of  $N$  variables defined over an open convex subset  $S$  of  $\mathbb{R}^N$ . Then (a)  $f$  is concave over  $S$  iff

$$f(\mathbf{x}^1) \leq f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^T(\mathbf{x}^1 - \mathbf{x}^0) \quad \text{for all } \mathbf{x}^0 \in S \text{ and } \mathbf{x}^1 \in S. \quad (2.51)$$

**Proof.** (a) implies (2.51). This follows from Theorem 6 and its corollary. (2.51) implies (a). Use the proof of (b) implies (a) in Theorem 7. ■

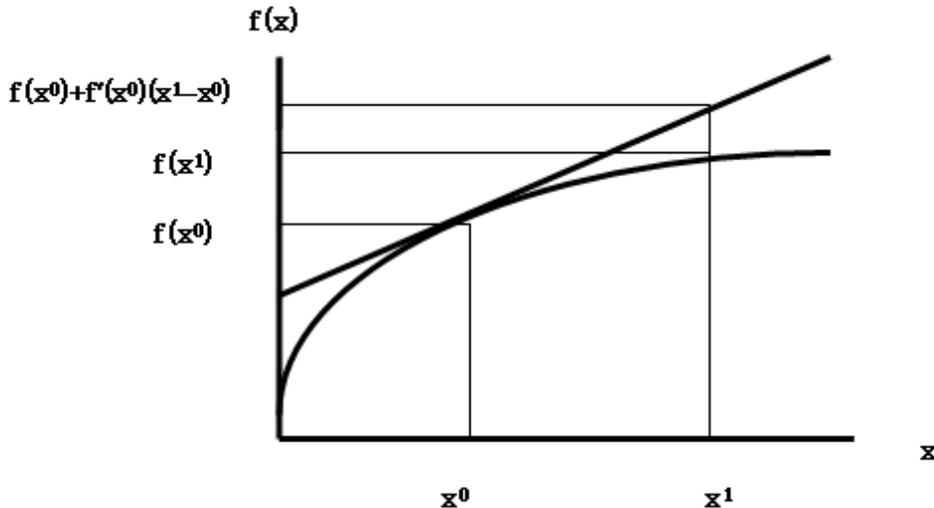


Fig. 2.5 The Third Characterization of Concavity

Thus in the case where the function is once differentiable and defined over an open convex set, then a necessary and sufficient condition for the function to be concave is that the first order linear approximation to  $f(\mathbf{x})$  around any point  $\mathbf{x}^0 \in S$ , which is  $f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^T(\mathbf{x} - \mathbf{x}^0)$ , must lie above (or be coincident with) the surface of the function.

Our final characterization of concavity is for functions of  $N$  variables,  $f(\mathbf{x})$ , that are twice continuously differentiable. This means that the first and second order partial derivative functions exist and are continuous. Note that in this case, Young's Theorem from calculus implies that

$$\frac{\partial^2 f(x_1, \dots, x_N)}{\partial x_i \partial x_k} = \frac{\partial^2 f(x_1, \dots, x_N)}{\partial x_k \partial x_i} \quad \text{for } 1 \leq i < k \leq N; \quad (2.52)$$

i.e., the matrix of second order partial derivatives,  $\nabla^2 f(\mathbf{x}) \equiv [\partial^2 f(x_1, \dots, x_N) / \partial x_i \partial x_k]$  is symmetric.

**Theorem 8** *Fourth Characterization of Concavity in the Twice Continuously Differentiable Case;* Fenchel (1953; 87-88)[179]: Let  $f$  be a twice continuously differentiable function of  $N$  variables defined over an open convex subset  $S$  of  $\mathbb{R}^N$ . Then (a)  $f$  is concave over  $S$  iff (b)  $\nabla^2 f(\mathbf{x})$  is negative semidefinite for all  $\mathbf{x} \in S$ .

**Proof.** (b) implies (a). Let  $\mathbf{x}^0$  and  $\mathbf{x}^1$  be two arbitrary points in  $S$ . Then by Taylor's Theorem for  $n = 2$ ,<sup>\*8</sup> there exists  $\theta$  such that  $0 < \theta < 1$  and

$$\begin{aligned} f(\mathbf{x}^1) &= f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^T (\mathbf{x}^1 - \mathbf{x}^0) + (1/2)(\mathbf{x}^1 - \mathbf{x}^0)^T \nabla^2 f(\theta \mathbf{x}^0 + (1 - \theta)\mathbf{x}^1)(\mathbf{x}^1 - \mathbf{x}^0) \\ &\leq f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^T (\mathbf{x}^1 - \mathbf{x}^0) \end{aligned} \quad (2.53)$$

where the inequality follows from the assumption (b) that  $\nabla^2 f(\theta \mathbf{x}^0 + (1 - \theta)\mathbf{x}^1)$  is negative semidefinite and hence

$$(1/2)(\mathbf{x}^1 - \mathbf{x}^0)^T \nabla^2 f(\theta \mathbf{x}^0 + (1 - \theta)\mathbf{x}^1)(\mathbf{x}^1 - \mathbf{x}^0) \leq 0. \quad (2.54)$$

But the inequalities in (2.53) are equivalent to (2.51) and hence  $f$  is concave.

(a) implies (b). We show that not (b) implies not (a). Not (b) means there exist  $\mathbf{x}^0 \in S$  and  $\mathbf{z} \neq \mathbf{0}_N$  such that

$$\mathbf{z}^T \nabla^2 f(\mathbf{x}^0) \mathbf{z} > 0. \quad (2.55)$$

Using (2.55) and the continuity of the second order partial derivatives of  $f$ , we can find a  $\delta > 0$  small enough so that  $\mathbf{x}^0 + t\mathbf{z} \in S$  and

$$\mathbf{z}^T \nabla^2 f(\mathbf{x}^0 + t\mathbf{z}) \mathbf{z} > 0 \quad \text{for } t \text{ such that } -\delta \leq t \leq \delta. \quad (2.56)$$

Define

$$\mathbf{x}^1 \equiv \mathbf{x}^0 + \delta \mathbf{z}. \quad (2.57)$$

By Taylor's Theorem, there exists  $\theta$  such that  $0 < \theta < 1$  and

$$\begin{aligned} f(\mathbf{x}^1) &= f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^T (\mathbf{x}^1 - \mathbf{x}^0) + (1/2)(\mathbf{x}^1 - \mathbf{x}^0)^T \nabla^2 f(\theta \mathbf{x}^0 + (1 - \theta)\mathbf{x}^1)(\mathbf{x}^1 - \mathbf{x}^0) \\ &= f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^T (\mathbf{x}^1 - \mathbf{x}^0) + (1/2)(\delta \mathbf{z})^T \nabla^2 f(\theta \mathbf{x}^0 + (1 - \theta)\mathbf{x}^1)(\delta \mathbf{z}) \quad \text{using (2.57)} \\ &= f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^T (\mathbf{x}^1 - \mathbf{x}^0) + \delta^2 (1/2) \mathbf{z}^T \nabla^2 f(\theta \mathbf{x}^0 + (1 - \theta)[\mathbf{x}^0 + \delta \mathbf{z}]) \mathbf{z} \quad \text{using (2.57)} \\ &= f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^T (\mathbf{x}^1 - \mathbf{x}^0) + \delta^2 (1/2) \mathbf{z}^T \nabla^2 f(\mathbf{x}^0 + (1 - \theta)\delta \mathbf{z}) \mathbf{z} \\ &> f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^T (\mathbf{x}^1 - \mathbf{x}^0) \end{aligned} \quad (2.58)$$

<sup>\*8</sup> For a function of one variable,  $g(t)$  say, Taylor's Theorem for  $n = 1$  is the Mean Value Theorem; i.e., if the derivative of  $g$  exists for say  $0 < t < 1$ , then there exists  $t^*$  such that  $0 < t^* < 1$  and  $g(1) = g(0) + g'(t^*)[1 - 0]$ . Taylor's Theorem for  $n = 2$  is: suppose the first and second derivatives of  $g(t)$  exist for  $0 \leq t \leq 1$ . Then there exists  $t^*$  such that  $0 < t^* < 1$  and  $g(1) = g(0) + g'(0)[1 - 0] + (1/2)g''(t^*)[1 - 0]^2$ . To see that (2.53) follows from this Theorem, define  $g(t) \equiv f(\mathbf{x}^0 + t[\mathbf{x}^1 - \mathbf{x}^0])$  for  $0 \leq t \leq 1$ . Routine calculations show that  $g'(t) = \nabla f(\mathbf{x}^0 + t[\mathbf{x}^1 - \mathbf{x}^0])^T [\mathbf{x}^1 - \mathbf{x}^0]$  and  $g''(t) = [\mathbf{x}^1 - \mathbf{x}^0]^T \nabla^2 f(\mathbf{x}^0 + t[\mathbf{x}^1 - \mathbf{x}^0]) [\mathbf{x}^1 - \mathbf{x}^0]$ . Now (2.53) follows from Taylor's Theorem for  $n = 2$  with  $\theta = 1 - t^*$ .

where the inequality in (2.58) follows from

$$\delta^2(1/2)\mathbf{z}^T\nabla^2f(\mathbf{x}^0) + (1-\theta)\delta\mathbf{z} > 0 \quad (2.59)$$

which in turn follows from  $\delta^2 > 0$ ,  $0 < (1-\theta)\delta < \delta$  and the inequality (2.56). But (2.58) contradicts (2.51) so that  $f$  cannot be concave. ■

For a twice continuously differentiable function of one variable  $f(\mathbf{x})$  defined over the open convex set  $S$ , the fourth characterization of concavity boils down to checking the following inequalities:

$$f''(\mathbf{x}) \leq 0 \quad \text{for all } \mathbf{x} \in S; \quad (2.60)$$

i.e., we need only check that the second derivative of  $f$  is negative or zero over its domain of definition. Thus as  $\mathbf{x}$  increases, we need the first derivative  $f'(\mathbf{x})$  to be decreasing or constant.

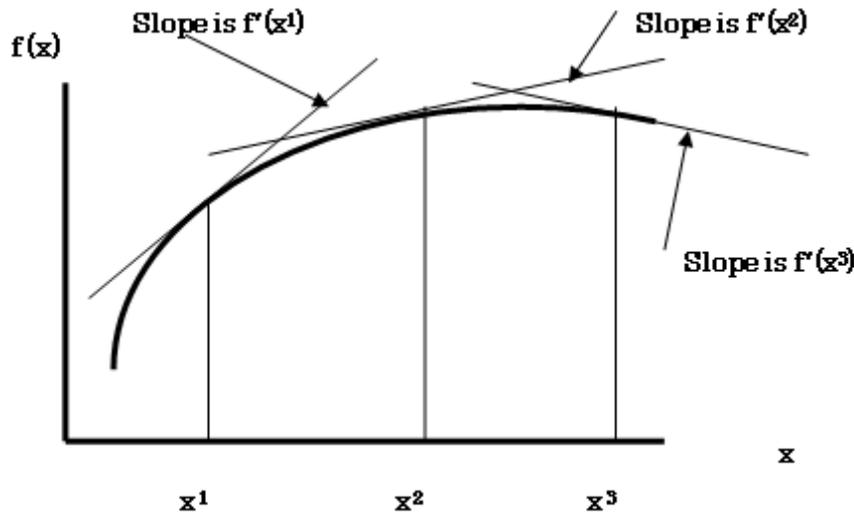


Fig. 2.6 The Fourth Characterization of Concavity

In Figure 2.6, as  $\mathbf{x}$  increases from  $\mathbf{x}^1$  to  $\mathbf{x}^2$  to  $\mathbf{x}^3$ , the slope of the tangent to the function  $f(\mathbf{x})$  decreases; i.e., we have  $f'(\mathbf{x}^1) > f'(\mathbf{x}^2) > f'(\mathbf{x}^3)$ . Thus the second derivative,  $f''(\mathbf{x})$ , decreases as  $\mathbf{x}$  increases.

**Problem 8** Define the function of one variable  $f(x) \equiv x^{1/2}$  for  $x > 0$ . Use the first, third and fourth characterizations of concavity to show that  $f$  is a concave function over the convex set  $S \equiv \{x : x > 0\}$ . Which characterization provides the easiest proof of concavity?

**Problem 9** Let  $f(\mathbf{x})$  be a concave function of  $N$  variables  $\mathbf{x} \equiv [x_1, \dots, x_N]$  defined over the open convex set  $S$ . Let  $\mathbf{x}^0 \in S$  and suppose  $\nabla f(\mathbf{x}^0) = \mathbf{0}_N$ . Then show that  $f(\mathbf{x}^0) \geq f(\mathbf{x})$  for all  $\mathbf{x} \in S$ ; i.e.,  $\mathbf{x}^0$  is a global maximizer for  $f$  over  $S$ .

*Hint:* Use one of the characterizations of concavity.

*Note:* This result can be extended to the case where  $S$  is a closed convex set with a nonempty interior. Thus the first order necessary conditions for maximizing a function of  $N$  variables are also sufficient if the function happens to be concave.

**Problem 10** Prove that: if  $f(\mathbf{x})$  and  $g(\mathbf{x})$  are concave functions of  $N$  variables  $\mathbf{x}$  defined over  $S \subset \mathbb{R}^N$  and  $\alpha \geq 0$  and  $\beta \geq 0$ , then  $\alpha f(\mathbf{x}) + \beta g(\mathbf{x})$  is concave over  $S$ .

**Problem 11** Fenchel (1953; 61)[179]: Show that: if  $f(\mathbf{x})$  is a concave function defined over the convex set  $S \subset \mathbb{R}^N$  and  $g$  is an increasing concave function of one variable defined over an interval that includes all of the numbers  $f(\mathbf{x})$  for  $\mathbf{x} \in S$ , then  $h(\mathbf{x}) \equiv g[f(\mathbf{x})]$  is a concave function over  $S$ .

**Problem 12** A function  $f(\mathbf{x})$  of  $N$  variables  $\mathbf{x} \equiv [x_1, \dots, x_N]$  defined over a convex subset of  $\mathbb{R}^N$  is *strictly concave* iff for every  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S$  and  $0 < \lambda < 1$ , we have

$$f(\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) > \lambda f(\mathbf{x}^1) + (1 - \lambda) f(\mathbf{x}^2). \quad (2.61)$$

Suppose that  $\mathbf{x}^1 \in S$  and  $\mathbf{x}^2 \in S$  and  $f(\mathbf{x}^1) = f(\mathbf{x}^2) = \max_{\mathbf{x}} \{f(\mathbf{x}) : \mathbf{x} \in S\}$ . Then show that  $\mathbf{x}^1 = \mathbf{x}^2$ .

*Note:* This shows that if the maximum of a strictly concave function over a convex set exists, then the set of maximizers is unique.

**Problem 13** Let  $f$  be a strictly concave function defined over a convex subset  $S$  of  $\mathbb{R}^N$ . If  $f$  attains a local maximum at the point  $\mathbf{x}^0 \in S$ , then show that  $f$  attains a strict global maximum at  $\mathbf{x}^0$ ; i.e., we have

$$f(\mathbf{x}^0) > f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in S \text{ where } \mathbf{x} \neq \mathbf{x}^0. \quad (a)$$

*Hint:* Modify the proof of Theorem 4.

## 2.5 Convex Functions

**Definition** A function  $f(\mathbf{x})$  of  $N$  variables  $\mathbf{x} \equiv [x_1, \dots, x_N]$  defined over a convex subset  $S$  of  $\mathbb{R}^N$  is *convex* iff for every  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S$  and  $0 < \lambda < 1$ , we have

$$f(\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) \leq \lambda f(\mathbf{x}^1) + (1 - \lambda) f(\mathbf{x}^2). \quad (2.62)$$

Comparing the definition (2.62) for a convex function with our previous definition (2.28) for a concave function, it can be seen that the inequalities and (2.28) and (2.62) are reversed. Thus an equivalent definition for a convex function  $f$  is:  $f(\mathbf{x})$  is convex over the convex set  $S$  iff  $-f$  is concave over the convex set  $S$ . This fact means that we do not have to do much work to establish the properties of convex functions: we can simply use all of the material in the previous section, replacing  $f$  in each result in the previous section by  $-f$  and then multiplying through the various inequalities by  $-1$  (thus reversing them) in order to eliminate the minus signs from the various inequalities. Following this strategy leads to (2.62) as the *first characterization of a convex function*. We list the other characterizations below.

*Second Characterization of Convexity;* Fenchel (1953; 57)[179]: (a)  $f$  is a convex function defined over the convex subset  $S$  of  $\mathbb{R}^N$  iff (b) the set  $E \equiv \{(y, \mathbf{x}) : y \geq f(\mathbf{x}), \mathbf{x} \in S\}$  is a convex set in  $\mathbb{R}^{N+1}$ .<sup>\*9</sup>

*Third Characterization of Convexity in the Once Differentiable Case;* Mangasarian (1969; 84)[302]: Let  $f$  be a once differentiable function of  $N$  variables defined over an open convex subset  $S$  of  $\mathbb{R}^N$ . Then  $f$  is convex over  $S$  iff

$$f(\mathbf{x}^1) \geq f(\mathbf{x}^0) + \nabla f(\mathbf{x}^0)^T (\mathbf{x}^1 - \mathbf{x}^0) \quad \text{for all } \mathbf{x}^0 \in S \text{ and } \mathbf{x}^1 \in S. \quad (2.63)$$

<sup>\*9</sup> The set  $E$  is called the *epigraph* of the function  $f$ ; it consists of the graph of  $f$  and all of the points lying above it.

We note that a vector  $\mathbf{b}$  is a *subgradient* to the function of  $N$  variables  $f$  defined over  $S$  at the point  $\mathbf{x}^0 \in S$  iff

$$f(\mathbf{x}) \geq f(\mathbf{x}^0) + \mathbf{b}^T(\mathbf{x} - \mathbf{x}^0) \quad \text{for all } \mathbf{x} \in S. \quad (2.64)$$

Applying this definition to (2.63) shows that  $\nabla f(\mathbf{x}^0)$  is a subgradient to  $f$  at the point  $\mathbf{x}^0$ .

*Fourth Characterization of Convexity in the Twice Continuously Differentiable Case*; Fenchel (1953; 87-88)[179]: Let  $f$  be a twice continuously differentiable function of  $N$  variables defined over an open convex subset  $S$  of  $\mathbb{R}^N$ . Then (a)  $f$  is convex over  $S$  iff (b)  $\nabla^2 f(\mathbf{x})$  is positive semidefinite for all  $\mathbf{x} \in S$ .

The counterpart to Theorem 4 about concave functions in the previous section is Theorem 9 below for convex functions.

**Theorem 9** *Local Minimum is a Global Minimum*; Fenchel (1953; 63)[179]: Let  $f$  be a convex function defined over a convex subset  $S$  of  $\mathbb{R}^N$ . If  $f$  attains a local minimum at the point  $\mathbf{x}^0 \in S$ , then  $f$  attains a global minimum at  $\mathbf{x}^0$ ; i.e., we have

$$f(\mathbf{x}^0) \leq f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in S. \quad (2.65)$$

**Problem 14** Recall that Example 5 in section 2.2 defined the *feasible region* for a linear programming problem as the following set  $S$ :

$$S \equiv \{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N\} \quad (a)$$

where  $\mathbf{A}$  is an  $M \times N$  matrix of constants and  $\mathbf{b}$  is an  $M$  dimensional vector of constants. We showed in section 2.2 that  $S$  was a closed convex set. We now assume in addition, that  $S$  is a nonempty bounded set. Now consider the following function of the vector  $\mathbf{c}$ :

$$f(\mathbf{c}) \equiv \max_{\mathbf{x}} \{\mathbf{c}^T \mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N\}. \quad (b)$$

Show that  $f(\mathbf{c})$  is a convex function for  $\mathbf{c} \in \mathbb{R}^N$ .

Now suppose that we define  $f(\mathbf{c})$  as follows:

$$f(\mathbf{c}) \equiv \min_{\mathbf{x}} \{\mathbf{c}^T \mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N\}. \quad (c)$$

Show that  $f(\mathbf{c})$  is a concave function for  $\mathbf{c} \in \mathbb{R}^N$ .

**Problem 15** Mangasarian (1969; 149)[302]: Let  $f$  be a positive concave function defined over a convex subset  $S$  of  $\mathbb{R}^N$ . Show that  $h(\mathbf{x}) \equiv 1/f(\mathbf{x})$  is a positive convex function over  $S$ .

*Hint:* You may find the fact that a weighted harmonic mean is less than or equal to the corresponding weighted arithmetic mean helpful.

**Problem 16** Let  $S$  be a closed and bounded set in  $\mathbb{R}^N$ . For  $\mathbf{p} \in \mathbb{R}^N$ , define the *support function* of  $S$  as

$$(a) \pi(\mathbf{p}) \equiv \max_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : \mathbf{x} \in S\}.$$

(b) Show that  $\pi(\mathbf{p})$  is a (positively) linearly homogeneous function over  $\mathbb{R}^N$ .<sup>\*10</sup>

(c) Show that  $\pi(\mathbf{p})$  is a convex function over  $\mathbb{R}^N$ .

(d) If  $\mathbf{0}_N \in S$ , then show  $\pi(\mathbf{p}) \geq 0$  for all  $\mathbf{p} \in \mathbb{R}^N$ .

---

<sup>\*10</sup> This means for all  $\lambda > 0$  and  $\mathbf{p} \in \mathbb{R}^N$ ,  $\pi(\lambda\mathbf{p}) = \lambda\pi(\mathbf{p})$ .

**Note:** If we changed the domain of definition for the vectors  $\mathbf{p}$  from  $\mathbb{R}^N$  to the positive orthant,  $\Omega \equiv \{\mathbf{x} : \mathbf{x} \gg \mathbf{0}_N\}$ , and defined  $\pi(\mathbf{p})$  by (a), then  $\pi(\mathbf{p})$  would satisfy the same properties (b), (c) and (d) and we could interpret  $\pi(\mathbf{p})$  as the *profit function* that corresponds to the technology set  $S$ .

**Problem 17** Define  $\Omega$  as the positive orthant in  $\mathbb{R}^N$ ; i.e.,  $\Omega \equiv \{\mathbf{x} : \mathbf{x} \gg \mathbf{0}_N\}$ . Suppose  $f(\mathbf{x})$  is a positive, positively linearly homogeneous and concave function defined over  $\Omega$ . Show that  $f$  is also increasing over  $\Omega$ ; i.e., show that  $f$  satisfies the following property:

$$\mathbf{x}^2 \gg \mathbf{x}^1 \gg \mathbf{0}_N \text{ implies } f(\mathbf{x}^2) > f(\mathbf{x}^1). \quad (\text{a})$$

## 2.6 Quasiconcave Functions

Example 6 in section 2.2 above indicated why quasiconcave functions arise in economic applications. In this section, we will study the properties of quasiconcave functions more formally.

**Definition** *First Characterization of Quasiconcavity*; Fenchel (1953; 117)[179]:  $f$  is a *quasiconcave function* defined over a convex subset  $S$  of  $\mathbb{R}^N$  iff

$$\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, 0 < \lambda < 1 \text{ implies } f(\lambda \mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2) \geq \min\{f(\mathbf{x}^1), f(\mathbf{x}^2)\}. \quad (2.66)$$

The above definition asks that the line segment joining  $\mathbf{x}^1$  to  $\mathbf{x}^2$  that has height equal to the minimum value of the function at the points  $\mathbf{x}^1$  and  $\mathbf{x}^2$  lies below (or is coincident with) the graph of  $f$  along the line segment joining  $\mathbf{x}^1$  to  $\mathbf{x}^2$ .

If  $f$  is concave over  $S$ , then

$$\begin{aligned} f(\lambda \mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2) &\geq \lambda f(\mathbf{x}^1) + (1 - \lambda)f(\mathbf{x}^2) \quad \text{using (2.28), the definition of concavity} \\ &\geq \min\{f(\mathbf{x}^1), f(\mathbf{x}^2)\} \end{aligned} \quad (2.67)$$

where the second inequality follows since  $\lambda f(\mathbf{x}^1) + (1 - \lambda)f(\mathbf{x}^2)$  is an average of  $f(\mathbf{x}^1)$  and  $f(\mathbf{x}^2)$ . Thus if  $f$  is concave, then it is also quasiconcave.

A geometric interpretation of property (2.66) can be found in Figure 2.7 for the case  $N = 1$ . Essentially, the straight line segment above  $\mathbf{x}^1$  and  $\mathbf{x}^2$  parallel to the  $\mathbf{x}$  axis that has height equal to the minimum of  $f(\mathbf{x}^1)$  and  $f(\mathbf{x}^2)$  must lie below (or be coincident with) the graph of the function between  $\mathbf{x}^1$  and  $\mathbf{x}^2$ . This property must hold for any two points in the domain of definition of  $f$ .

If  $f$  is a function of one variable, then any *monotonic* function<sup>\*11</sup> defined over a convex set will be quasiconcave. Functions of one variable that are increasing (or nondecreasing) and then are decreasing (or nonincreasing) are also quasiconcave. Such functions need not be continuous or concave and thus quasiconcavity is a genuine generalization of concavity. Note also that quasiconcave functions can have flat spots on their graphs.

The following Theorem shows that the above definition of quasiconcavity is equivalent to the definition used in Example 6 in section 2.2 above.

**Theorem 10** *Second Characterization of Quasiconcavity*; Fenchel (1953; 118)[179]:  $f$  is a quasiconcave function over the convex subset  $S$  of  $\mathbb{R}^N$  iff

$$\text{For every } u \in \text{Range } f, \text{ the upper level set } L(u) \equiv \{\mathbf{x} : f(\mathbf{x}) \geq u; \mathbf{x} \in S\} \text{ is a convex set.} \quad (2.68)$$

<sup>\*11</sup> A monotonic function of one variable is one that is either increasing, decreasing, nonincreasing or nondecreasing.

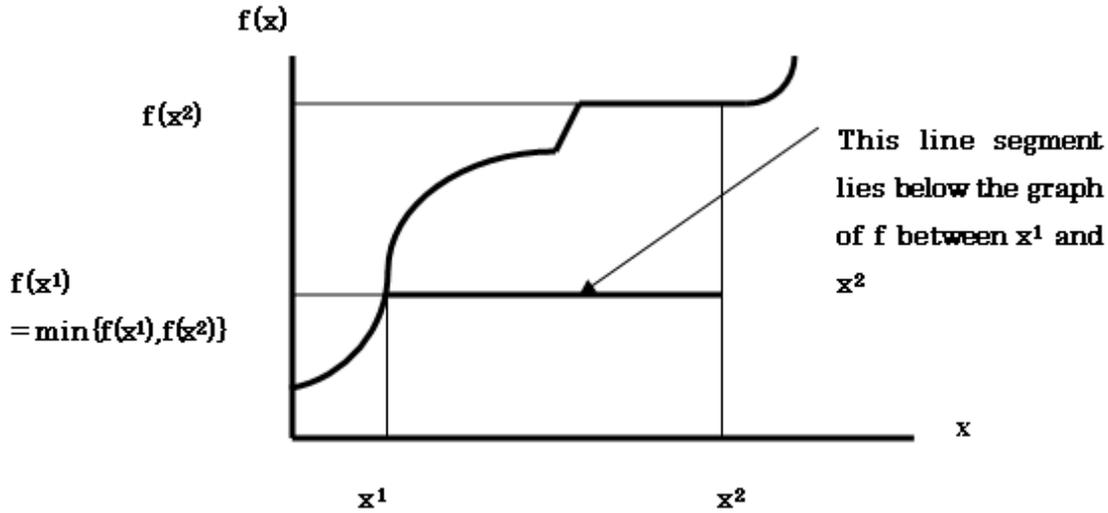


Fig. 2.7 The First Characterization of Quasiconcavity

**Proof.** (2.66) implies (2.68): Let  $u \in \text{Range } f$ ,  $\mathbf{x}^1 \in L(u)$ ,  $\mathbf{x}^2 \in L(u)$  and  $0 < \lambda < 1$ . Since  $\mathbf{x}^1$  and  $\mathbf{x}^2$  belong to  $L(u)$ ,

$$f(\mathbf{x}^1) \geq u; \quad f(\mathbf{x}^2) \geq u. \quad (2.69)$$

Using (2.66), we have:

$$\begin{aligned} f(\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) &\geq \min\{f(\mathbf{x}^1), f(\mathbf{x}^2)\} \\ &\geq u \end{aligned} \quad (2.70)$$

where the last inequality follows using (2.69). But (2.70) shows that  $\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2 \in L(u)$  and thus  $L(u)$  is a convex set.

(2.68) implies (2.66): Let  $\mathbf{x}^1 \in S$ ,  $\mathbf{x}^2 \in S$ ,  $0 < \lambda < 1$  and let  $u \equiv \min\{f(\mathbf{x}^1), f(\mathbf{x}^2)\}$ . Thus  $f(\mathbf{x}^1) \geq u$  and hence,  $\mathbf{x}^1 \in L(u)$ . Similarly,  $f(\mathbf{x}^2) \geq u$  and hence,  $\mathbf{x}^2 \in L(u)$ . Since  $L(u)$  is a convex set using (2.68),  $\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2 \in L(u)$ . Hence, using the definition of  $L(u)$ ,

$$f(\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) \geq u \equiv \min\{f(\mathbf{x}^1), f(\mathbf{x}^2)\} \quad (2.71)$$

which is (2.66). ■

Figure 2.1 in section 2.2 above suffices to give a geometric interpretation of the second characterization of quasiconcavity.

The first characterization of quasiconcavity (2.66) can be written in an equivalent form as follows:

$$\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, \mathbf{x}^1 \neq \mathbf{x}^2, 0 < \lambda < 1, f(\mathbf{x}^1) \leq f(\mathbf{x}^2) \text{ implies } f(\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) \geq f(\mathbf{x}^1). \quad (2.72)$$

We now turn to our third characterization of quasiconcavity. For this characterization, we will assume that  $f$  is defined over an *open* convex subset  $S$  of  $\mathbb{R}^N$  and that the first order partial derivatives of  $f$ ,  $\partial f(\mathbf{x})/\partial x_n$  for  $n = 1, \dots, N$ , exist and are *continuous* functions over  $S$ . In this case, the property on  $f$  that will characterize a quasiconcave function is the following one:

$$\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, \mathbf{x}^1 \neq \mathbf{x}^2, \nabla f(\mathbf{x}^1)^T (\mathbf{x}^2 - \mathbf{x}^1) < 0 \text{ implies } f(\mathbf{x}^2) < f(\mathbf{x}^1). \quad (2.73)$$

**Theorem 11** *Third Characterization of Quasiconcavity*; Mangasarian (1969; 147)[302]: Let  $f$  be a once continuously differentiable function defined over the open convex subset  $S$  of  $\mathbb{R}^N$ . Then  $f$  is quasiconcave over  $S$  iff  $f$  satisfies (2.73).

**Proof.** (2.72) implies (2.73). We show that not (2.73) implies not (2.72). Not (2.73) means that there exist  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, \mathbf{x}^1 \neq \mathbf{x}^2$  such that

$$\nabla f(\mathbf{x}^1)^T(\mathbf{x}^2 - \mathbf{x}^1) < 0 \text{ and} \quad (2.74)$$

$$f(\mathbf{x}^2) \geq f(\mathbf{x}^1). \quad (2.75)$$

Define the function of one variable  $g(t)$  for  $0 \leq t \leq 1$  as follows:

$$g(t) \equiv f(\mathbf{x}^1 + t[\mathbf{x}^2 - \mathbf{x}^1]). \quad (2.76)$$

It can be verified that

$$g(0) = f(\mathbf{x}^1) \text{ and } g(1) = f(\mathbf{x}^2). \quad (2.77)$$

It can be verified that the derivative of  $g(t)$  for  $0 \leq t \leq 1$  can be computed as follows:

$$g'(t) = \nabla f(\mathbf{x}^1 + t[\mathbf{x}^2 - \mathbf{x}^1])^T(\mathbf{x}^2 - \mathbf{x}^1). \quad (2.78)$$

Evaluating (2.78) at  $t = 0$  and using (2.74) shows that<sup>\*12</sup>

$$g'(0) < 0. \quad (2.79)$$

Using the continuity of the first order partial derivatives of  $f$ , it can be seen that (2.79) implies the existence of a  $\delta$  such that

$$0 < \delta < 1 \quad \text{and} \quad (2.80)$$

$$g'(t) < 0 \quad \text{for all } t \text{ such that } 0 \leq t \leq \delta. \quad (2.81)$$

Thus  $g(t)$  is a decreasing function over this interval of  $t$ 's and thus

$$g(\delta) \equiv f(\mathbf{x}^1 + \delta[\mathbf{x}^2 - \mathbf{x}^1]) = f([1 - \delta]\mathbf{x}^1 + \delta\mathbf{x}^2) < g(0) \equiv f(\mathbf{x}^1). \quad (2.82)$$

But (2.80) and (2.82) imply that

$$f(\lambda\mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2) < f(\mathbf{x}^1) \quad (2.83)$$

where  $\lambda \equiv 1 - \delta$ . Since (2.80) implies that  $0 < \lambda < 1$ , (2.83) contradicts (2.72) and so  $f$  is not quasiconcave.

(2.73) implies (2.72). We show not (2.72) implies not (2.73). Not (2.72) means that there exist  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, \mathbf{x}^1 \neq \mathbf{x}^2$  and  $0 < \lambda^* < 1$  such that

$$f(\mathbf{x}^1) \leq f(\mathbf{x}^2) \quad \text{and} \quad (2.84)$$

$$f(\lambda^*\mathbf{x}^1 + (1 - \lambda^*)\mathbf{x}^2) < f(\mathbf{x}^1). \quad (2.85)$$

Define the function of one variable  $g(t)$  for  $0 \leq t \leq 1$  as follows:

$$g(t) \equiv f(\mathbf{x}^1 + t[\mathbf{x}^2 - \mathbf{x}^1]). \quad (2.86)$$

<sup>\*12</sup> Note that  $g'(0)$  is the directional derivative of  $f(\mathbf{x})$  in the direction defined by  $\mathbf{x}^2 - \mathbf{x}^1$ .

Define  $t^*$  as follows:

$$t^* \equiv 1 - \lambda^* \tag{2.87}$$

and note that  $0 < t^* < 1$  and

$$\begin{aligned} g(t^*) &\equiv f(\lambda^* \mathbf{x}^1 + (1 - \lambda^*) \mathbf{x}^2) \\ &\leq f(\mathbf{x}^1) \quad \text{using (2.85)} \\ &= g(0) \quad \text{using definition (2.86)}. \end{aligned} \tag{2.88}$$

The continuity of the first order partial derivatives of  $f$  implies that  $g'(t)$  and  $g(t)$  are continuous functions of  $t$  for  $0 \leq t \leq 1$ . Now consider the behavior of  $g(t)$  along the line segment  $0 \leq t \leq 1$ . The inequality (2.88) shows that  $g(t)$  eventually decreases from  $g(0)$  to the lower number  $g(t^*)$  along this interval. Thus there must exist a  $t^{**}$  such that

$$0 \leq t^{**} < t^*; \tag{2.89}$$

$$g(t) \leq g(0) \quad \text{for all } t \text{ such that } t^{**} \leq t \leq t^* \text{ and} \tag{2.90}$$

$$g(t^{**}) = g(0). \tag{2.91}$$

Essentially, the inequalities (2.89)-(2.91) say that there exists a closed interval to the immediate left of the point  $t^*$ ,  $[t^{**}, t^*]$ , such that  $g(t)$  is less than or equal to  $g(0)$  for  $t$  in this interval and the lower boundary point of the interval,  $t^{**}$ , is such that  $g(t^{**})$  equals  $g(0)$ .

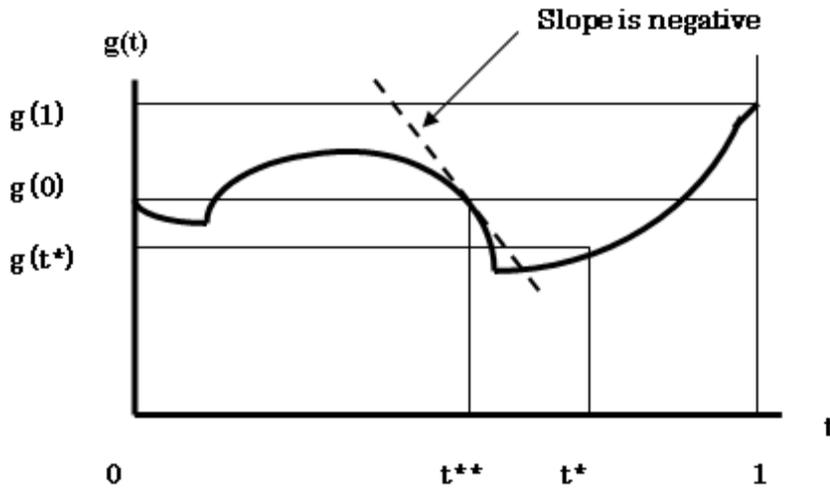


Fig. 2.8 The Geometry of Theorem 11

Now suppose that the derivative of  $g$  is nonnegative for all  $t$  in the interval  $[t^{**}, t^*]$ ; i.e.,

$$g'(t) \geq 0 \quad \text{for all } t \text{ such that } t^{**} \leq t \leq t^*. \tag{2.92}$$

Then by the Mean Value Theorem, there exists  $t^{***}$  such that  $t^{**} < t^{***} < t^*$  and

$$\begin{aligned} g(t^*) &= g(t^{**}) + g'(t^{***})(t^* - t^{**}) \\ &\geq g(t^{**}) \quad \text{using (2.92) and } t^{**} < t^* \\ &= g(0) \quad \text{using (2.91)}. \end{aligned} \tag{2.93}$$

But the inequality (2.93) contradicts  $g(t^*) < g(0)$ , which is equivalent to (2.85). Thus our *supposition* is false. Hence there exists  $t'$  such that

$$t^{**} < t' < t^* \quad \text{and} \quad (2.94)$$

$$g'(t') < 0. \quad (2.95)$$

In Figure 2.8, such a point  $t'$  is just to the right of  $t^{**}$  where the dashed line is tangent to the graph of  $g(t)$ . Using (2.90), we also have

$$g(t') \leq g(0). \quad (2.96)$$

Using definition (2.86), the inequalities (2.95) and (2.96) translate into the following inequalities:

$$g'(t') = \nabla f(\mathbf{x}^1 + t'[\mathbf{x}^2 - \mathbf{x}^1])^T(\mathbf{x}^2 - \mathbf{x}^1) < 0; \quad (2.97)$$

$$\begin{aligned} g(t') &= f(\mathbf{x}^1 + t'[\mathbf{x}^2 - \mathbf{x}^1]) \leq f(\mathbf{x}^1) \\ &\leq f(\mathbf{x}^2) \quad \text{using (2.84)}. \end{aligned} \quad (2.98)$$

Now define

$$\mathbf{x}^3 \equiv \mathbf{x}^1 + t'[\mathbf{x}^2 - \mathbf{x}^1] \quad (2.99)$$

and note that the inequalities (2.94) imply that

$$0 < t' < 1. \quad (2.100)$$

Using definition (2.99), we have

$$\begin{aligned} \mathbf{x}^2 - \mathbf{x}^3 &= \mathbf{x}^2 - \{\mathbf{x}^1 + t'[\mathbf{x}^2 - \mathbf{x}^1]\} \\ &= (1 - t')[\mathbf{x}^2 - \mathbf{x}^1] \\ &\neq \mathbf{0}_N \quad \text{using (2.100) and } \mathbf{x}^1 \neq \mathbf{x}^2. \end{aligned} \quad (2.101)$$

Note that the second equation in (2.101) implies that

$$\mathbf{x}^2 - \mathbf{x}^1 = (1 - t')^{-1}[\mathbf{x}^2 - \mathbf{x}^3]. \quad (2.102)$$

Now substitute (2.99) and (2.102) into (2.97) and we obtain the following inequality:

$$\begin{aligned} (1 - t')^{-1} \nabla f(\mathbf{x}^3)^T(\mathbf{x}^2 - \mathbf{x}^3) &< 0 \quad \text{or} \\ \nabla f(\mathbf{x}^3)^T(\mathbf{x}^2 - \mathbf{x}^3) &< 0 \quad \text{since } (1 - t')^{-1} > 0. \end{aligned} \quad (2.103)$$

$$f(\mathbf{x}^3) \leq f(\mathbf{x}^2). \quad (2.104)$$

The inequalities (2.103) and (2.104) show that (2.73) does not hold, with  $\mathbf{x}^3$  playing the role of  $\mathbf{x}^1$  in condition (2.73). ■

**Corollary** Arrow and Enthoven (1961; 780)[14]: Let  $f$  be a once continuously differentiable function defined over the open convex subset  $S$  of  $\mathbb{R}^N$ . Then  $f$  is quasiconcave over  $S$  iff  $f$  satisfies the following condition:

$$\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, \mathbf{x}^1 \neq \mathbf{x}^2, f(\mathbf{x}^2) \geq f(\mathbf{x}^1) \text{ implies } \nabla f(\mathbf{x}^1)^T(\mathbf{x}^2 - \mathbf{x}^1) \geq 0. \quad (2.105)$$

**Proof.** Condition (2.105) is the contrapositive to condition (2.73) and is logically equivalent to it. ■

We can use the third characterization of concavity to show that the third characterization of quasiconcavity holds and hence a once continuously differentiable concave function  $f$  defined over an open convex set  $S$  is also quasiconcave (a fact which we already know). Thus let  $f$  be concave over  $S$  and assume the conditions in (2.73); i.e., let  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, \mathbf{x}^1 \neq \mathbf{x}^2$  and assume

$$\nabla f(\mathbf{x}^1)^T(\mathbf{x}^2 - \mathbf{x}^1) < 0. \quad (2.106)$$

We need only show that  $f(\mathbf{x}^2) < f(\mathbf{x}^1)$ . Using the third characterization of concavity, we have:

$$\begin{aligned} f(\mathbf{x}^2) &\leq f(\mathbf{x}^1) + \nabla f(\mathbf{x}^1)^T(\mathbf{x}^2 - \mathbf{x}^1) \\ &\leq f(\mathbf{x}^1) \quad \text{using (2.106)} \end{aligned} \quad (2.107)$$

which is the desired result.

It turns out that it is quite difficult to get simple necessary and sufficient conditions for quasiconcavity in the case where  $f$  is twice continuously differentiable (although it is quite easy to get sufficient conditions). In order to get necessary and sufficient conditions, we will have to take a bit of a detour for a while.

**Definition** Let  $f(\mathbf{x})$  be a function of  $N$  variables  $\mathbf{x}$  defined for  $\mathbf{x} \in S$  where  $S$  is a convex set. Then  $f$  has the *line segment minimum property* iff

$$\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, \mathbf{x}^1 \neq \mathbf{x}^2 \text{ implies } \min_t \{f(t\mathbf{x}^1 + (1-t)\mathbf{x}^2) : 0 \leq t \leq 1\} \text{ exists;} \quad (2.108)$$

i.e., the minimum of  $f$  along any line segment in its domain of definition exists.

It is easy to verify that if  $f$  is a quasiconcave function defined over a convex set  $S$ , then it satisfies the line segment minimum property (2.108), since the minimum in (2.108) will be attained at one or both of the endpoints of the interval; i.e., the minimum in (2.108) will be attained at either  $f(\mathbf{x}^1)$  or  $f(\mathbf{x}^2)$  (or both points) since  $f(t\mathbf{x}^1 + (1-t)\mathbf{x}^2)$  for  $0 \leq t \leq 1$  is equal to or greater than  $\min\{f(\mathbf{x}^1), f(\mathbf{x}^2)\}$  and this minimum is attained at either  $f(\mathbf{x}^1)$  or  $f(\mathbf{x}^2)$  (or both points).

**Definition** Diewert, Avriel and Zang (1981; 400)[136]: The function of one variable,  $g(t)$ , defined over an interval  $S$  attains a *semistrict minimum* at  $t^0 \in \text{Int } S$  iff there exist  $\delta_1 > 0$  and  $\delta_2 > 0$  such that  $t^0 - \delta_1 \in S, t^0 + \delta_2 \in S$  and

$$g(t^0) \leq g(t) \quad \text{for all } t \text{ such that } t^0 - \delta_1 \leq t \leq t^0 + \delta_2; \quad (2.109)$$

$$g(t^0) < g(t^0 - \delta_1) \text{ and } g(t^0) < g(t^0 + \delta_2). \quad (2.110)$$

If  $g$  just satisfied (2.109) at the point  $t^0$ , then it can be seen that  $g(t)$  attains a *local minimum* at  $t^0$ . But the conditions (2.110) show that a semistrict local minimum is stronger than a local minimum: for  $g$  to attain a semistrict local minimum at  $t^0$ , we need  $g$  to attain a local minimum at  $t^0$ , but the function must eventually strictly increase at the end points of the region where the function attains the local minimum. Note that  $g(t)$  attains a *strict local minimum* at  $t^0 \in \text{Int } S$  iff there exist  $\delta > 0$  such that  $t^0 - \delta \in S, t^0 + \delta \in S$  and

$$g(t^0) < g(t) \quad \text{for all } t \text{ such that } t^0 - \delta \leq t \leq t^0 + \delta \text{ but } t \neq t^0. \quad (2.111)$$

It can be seen that if  $g$  attains a strict local minimum at  $t^0$ , then it also attains a semistrict local minimum at  $t^0$ . Hence, a semistrict local minimum is a concept that is intermediate to the concept of a local and strict local minimum.

**Theorem 12** Diewert, Avriel and Zang (1981; 400)[136]: Let  $f(\mathbf{x})$  be a function of  $N$  variables  $\mathbf{x}$  defined for  $\mathbf{x} \in S$  where  $S$  is a convex set and suppose that  $f$  has the line segment minimum property (2.108). Then  $f$  is quasiconcave over  $S$  iff  $f$  has the following property:

$$\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, \mathbf{x}^1 \neq \mathbf{x}^2 \text{ implies that } g(t) \equiv f(\mathbf{x}^1 + t[\mathbf{x}^2 - \mathbf{x}^1]) \text{ does not attain a semistrict local minimum for any } t \text{ such that } 0 < t < 1. \quad (2.112)$$

**Proof.** Quasiconcavity implies (2.112): This is equivalent to showing that not (2.112) implies not (2.66). Not (2.112) means there exists  $t^*$  such that  $0 < t^* < 1$  and  $g(t)$  attains a semistrict local minimum at  $t^*$ . This implies the existence of  $t_1$  and  $t_2$  such that

$$0 \leq t_1 < t^* < t_2 \leq 1; \quad (2.113)$$

$$g(t_1) > g(t^*); \quad g(t_2) > g(t^*). \quad (2.114)$$

Using the definition of  $g$ , (2.114) implies that

$$f(\mathbf{x}^1 + t^*[\mathbf{x}^2 - \mathbf{x}^1]) < \min\{f(\mathbf{x}^1 + t_1[\mathbf{x}^2 - \mathbf{x}^1]), f(\mathbf{x}^1 + t_2[\mathbf{x}^2 - \mathbf{x}^1])\}. \quad (2.115)$$

But (2.113) can be used to show that the point  $\mathbf{x}^1 + t^*[\mathbf{x}^2 - \mathbf{x}^1]$  is a convex combination of the points  $\mathbf{x}^1 + t_1[\mathbf{x}^2 - \mathbf{x}^1]$  and  $\mathbf{x}^1 + t_2[\mathbf{x}^2 - \mathbf{x}^1]$ ,<sup>\*13</sup> and hence (2.115) contradicts the definition of quasiconcavity, (2.66). Hence  $f$  is not quasiconcave.

(2.112) implies quasiconcavity (2.66). This is equivalent to showing not (2.66) implies not (2.112). Suppose  $f$  is not quasiconcave. Then there exist  $\mathbf{x}^1 \in S, \mathbf{x}^2 \in S$ , and  $\lambda^*$  such that  $0 < \lambda^* < 1$  and

$$f(\lambda^* \mathbf{x}^1 + (1 - \lambda^*) \mathbf{x}^2) < \min\{f(\mathbf{x}^1), f(\mathbf{x}^2)\}. \quad (2.116)$$

Define  $g(t) \equiv f(\mathbf{x}^1 + t[\mathbf{x}^2 - \mathbf{x}^1])$  for  $0 \leq t \leq 1$ . Since  $f$  is assumed to satisfy the line segment minimum property, there exists a  $t^*$  such that  $0 \leq t^* \leq 1$  and

$$g(t^*) = \min_t \{g(t) : 0 \leq t \leq 1\}. \quad (2.117)$$

The definition of  $g$  and (2.116) shows that  $t^*$  satisfies  $0 < t^* < 1$  and

$$f(\mathbf{x}^1 + t^*[\mathbf{x}^2 - \mathbf{x}^1]) = f((1 - t^*) \mathbf{x}^1 + t^* \mathbf{x}^2) < \min\{f(\mathbf{x}^1), f(\mathbf{x}^2)\}. \quad (2.118)$$

Thus  $f$  attains a semistrict local minimum, which contradicts (2.112). ■

**Theorem 13** *Fourth Characterization of Quasiconcavity*; Diewert, Avriel and Zang (1981; 401)[136]: Let  $f(\mathbf{x})$  be a twice continuously differentiable function of  $N$  variables  $\mathbf{x}$  defined over an open convex subset  $S$  of  $\mathbb{R}^N$ .<sup>\*14</sup> Then  $f$  is quasiconcave over  $S$  iff  $f$  has the following property:

$$\mathbf{x}^0 \in S, \mathbf{v} \neq \mathbf{0}_N, \mathbf{v}^T \nabla f(\mathbf{x}^0) = 0 \text{ implies (i) } \mathbf{v}^T \nabla^2 f(\mathbf{x}^0) \mathbf{v} < 0 \text{ or (ii) } \mathbf{v}^T \nabla^2 f(\mathbf{x}^0) \mathbf{v} = 0 \text{ and } g(t) \equiv f(\mathbf{x}^0 + t\mathbf{v}) \text{ does not attain a semistrict local minimum at } t = 0. \quad (2.119)$$

**Proof.** We need only show that property (2.119) is equivalent to property (2.112) in the twice continuously differentiable case. Property (2.112) is equivalent to

$$\mathbf{x}^0 \in S, \mathbf{v} \neq \mathbf{0}_N \text{ and } g(t) \equiv f(\mathbf{x}^0 + t\mathbf{v}) \text{ does not attain a semistrict local minimum at } t = 0. \quad (2.120)$$

<sup>\*13</sup> The weights are  $\lambda \equiv [t_2 - t^*]/[t_2 - t_1]$  and  $1 - \lambda \equiv [t^* - t_1]/[t_2 - t_1]$ .

<sup>\*14</sup> These conditions are strong enough to imply the continuity of  $f$  over  $S$  and hence the line segment minimum property will hold for  $f$ .

Consider case (i) in (2.119). If this case occurs, then  $g(t)$  attains a strict local maximum at  $t = 0$  and hence cannot attain a semistrict local minimum at  $t = 0$ . Hence, in the twice continuously differentiable case, (2.119) is equivalent to (2.120). ■

Note that the following condition is *sufficient* for (2.119) in the twice continuously differentiable case and hence if it holds for every  $\mathbf{x}^0 \in S$ ,  $f$  will be quasiconcave:

$$\mathbf{x}^0 \in S, \mathbf{v} \neq \mathbf{0}_N, \mathbf{v}^T \nabla f(\mathbf{x}^0) = 0 \text{ implies } \mathbf{v}^T \nabla^2 f(\mathbf{x}^0) \mathbf{v} < 0. \quad (2.121)$$

If  $f$  satisfies (2.121), then the Hessian matrix of  $f$ , the matrix of second order partial derivatives  $\nabla^2 f(\mathbf{x}^0)$ , is said to be *negative definite in the subspace orthogonal to the gradient vector*,  $\nabla f(\mathbf{x}^0)$ .

Examination of (2.119) shows that the following condition is *necessary* for quasiconcavity in the twice continuously differentiable case:

$$\mathbf{x}^0 \in S, \mathbf{v} \neq \mathbf{0}_N, \mathbf{v}^T \nabla f(\mathbf{x}^0) = 0 \text{ implies } \mathbf{v}^T \nabla^2 f(\mathbf{x}^0) \mathbf{v} \leq 0. \quad (2.122)$$

If  $f$  satisfies (2.122), then the Hessian matrix of  $f$ , the matrix of second order partial derivatives  $\nabla^2 f(\mathbf{x}^0)$ , is said to be *negative semidefinite in the subspace orthogonal to the gradient vector*,  $\nabla f(\mathbf{x}^0)$ .

Diewert, Avriel and Zang (1981)[136] show how many other generalizations of concavity can be obtained by looking at the local minimum or maximum properties of a function.

## 2.7 Quasiconvex Functions

**Definition** Fenchel (1953; 117)[179]:  $f$  is a *quasiconvex function* defined over a convex subset  $S$  of  $\mathbb{R}^N$  iff

$$\mathbf{x}^1 \in S, \mathbf{x}^2 \in S, 0 < \lambda < 1 \text{ implies } f(\lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) \leq \max\{f(\mathbf{x}^1), f(\mathbf{x}^2)\}. \quad (2.123)$$

Comparing the definition (2.123) for a quasiconvex function with our previous definition (2.66) for a quasiconcave function, it can be seen that  $f$  satisfies (2.123) if and only if  $-f$  satisfies (2.66). Thus an equivalent definition for a quasiconvex function  $f$  is:  $f(\mathbf{x})$  is quasiconvex over the convex set  $S$  iff  $-f$  is quasiconcave over the convex set  $S$ . This fact means that we do not have to do much work to establish the properties of quasiconvex functions: we can simply use all of the material in the previous section, replacing  $f$  in each result in the previous section by  $-f$  and then multiplying through the various inequalities by  $-1$  (thus reversing them) in order to eliminate the minus signs from the various inequalities.

**Problem 18** Write down counterparts to the second, third and fourth characterizations for quasiconcave functions for quasiconvex functions. Your characterizations for quasiconvexity should not contain any minus signs. (No proofs are required.)

**Problem 19** Luenberger (1968)[296]: Let  $f$  be a quasiconcave function defined over a convex subset  $S$  of  $\mathbb{R}^N$ . If  $f$  attains a strict local maximum at  $\mathbf{x}^0 \in S$ , then  $f$  attains a strict global maximum over  $S$  at  $\mathbf{x}^0$ .

**Problem 20** Let  $f$  be a quasiconcave, positive function defined over the convex subset  $S$  of  $\mathbb{R}^N$ . Show that  $g(\mathbf{x}) \equiv 1/f(\mathbf{x})$  is a quasiconvex function over  $S$ .

**Problem 21** Berge (1963; 208)[35]: Let  $f$  be a positive, linearly homogeneous and quasiconcave function defined over the positive orthant in  $\mathbb{R}^N$ ,  $\Omega \equiv \{\mathbf{x} : \mathbf{x} \gg \mathbf{0}_N\}$ . Show that  $f$  is concave over  $\Omega$ .

**Problem 22** Yaari (1977; 1184)[407]: Define  $F(x, y) \equiv f(x) + g(y)$  for  $x \in X$  and  $y \in Y$  where  $X$  and  $Y$  are convex subsets of  $\mathbb{R}^1$ . Suppose that  $f$  and  $g$  are continuous increasing functions over their domains of definition and  $F$  is quasiconcave over  $X \otimes Y$ . Show that at least one of the functions  $f$  or  $g$  must be concave.

**Problem 23** Blackorby, Davidson and Donaldson (1977; 357-358)[37]: Define  $F(\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^M) \equiv g[\sum_{m=1}^M f^m(\mathbf{x}^m)]$  for  $\mathbf{x}^m \in X^m, m = 1, 2, \dots, M$  where each  $X^m$  is an open convex subset of  $\mathbb{R}^{N^m}$  for each  $m$ . Assume that  $F$  is quasiconcave,  $g$  is a monotonically increasing twice continuously differentiable function of one variable, defined over the range of  $\sum_{m=1}^M f^m(\mathbf{x}^m)$  for  $\mathbf{x}^m \in X^m, m = 1, 2, \dots, M$ , and that the functions  $f^m$  are twice continuously differentiable. Show that each  $f^m$  is quasiconcave over  $X^m$  and at most one of the  $M$  functions,  $f^1, f^2, \dots, f^M$ , can fail to be concave.

## 2.8 References

- Arrow, K.J. and A.C. Enthoven (1961), "Quasi-Concave Programming", *Econometrica* 29, 779-800.
- Berge, C. (1963), *Topological Spaces*, New York: MacMillan.
- Blackorby, C., R. Davidson and D. Donaldson (1977), "A Homiletic Exposition of the Expected Utility Theorem", *Economica* 44, 351-358.
- Diewert, W.E., M. Avriel and I. Zang (1981), "Nine Kinds of Quasiconcavity and Concavity", *Journal of Economic Theory* 25:3, 397-420.
- Fenchel, W. (1953), "Convex Cones, Sets and Functions", Lecture Notes at Princeton University, Department of Mathematics, Princeton, N.J.
- Luenberger, D.G. (1968), "Quasi-Convex Programming", *SIAM Journal of Applied Mathematics* 16, 1090-1095.
- Mangasarian, O. (1969), *Nonlinear Programming*, New York: McGraw-Hill.
- Minkowski, H. (1911), *Theorie der konvexen Körper*, Gesammelte Abhandlungen, Zweiter Band, Berlin.
- Ponstein, J. (1967), "Seven Kinds of Convexity", *SIAM Review* 9, 115-119.
- Roberts, A.W. and D.E. Varberg (1973), *Convex Functions*, New York: Academic Press.
- Rockafellar, R.T. (1970), *Convex Analysis*, Princeton, N.J.: Princeton University Press.
- Rudin, W. (1953), *Principles of Mathematical Analysis*, New York: McGraw-Hill.
- Yaari, M.E. (1977), "A Note on Separability and Quasiconcavity", *Econometrica* 45, 1183-1186.

## Chapter 3

# Microeconomic Theory: A Dual Approach

### 3.1 Introduction

In this chapter, we will show how the theory of convex sets and concave and convex functions can be useful in deriving some theorems in microeconomics. Section 3.2 starts off by developing the properties of cost functions. It is shown that without assuming any regularity properties on an underlying production function, the corresponding function satisfies a large number of regularity properties. Section 3.3 shows how the cost function can be used to determine a production function that is consistent with a given cost function satisfying the appropriate regularity conditions. Section 3.4 establishes the derivative property of the cost function: it is shown that the first order partial derivatives of the cost function generate the firm's system of cost minimizing input demand functions. Section 3.5 shows how the material in the previous sections can be used to derive the comparative statics properties of the producer's system of cost minimizing input demand functions. Section 3.6 asks under what conditions can we assume that the technology exhibits constant returns to scale. Section 3.7 indicates that price elasticities of demand will tend to decrease in magnitude as a production model becomes more aggregated.

Section 3.8 notes that the duality between cost and production functions is isomorphic or identical to the duality between utility and expenditure functions. In this extension of the previous theory, the output level of the producer is replaced with the utility level of the consumer, the production function of the producer is replaced with the utility function of the consumer and the producer's cost minimization problem is replaced by the problem of the consumer minimizing the expenditure required to attain a target utility level. Thus the results in the first 5 sections have an immediate application to the consumer's system of Hicksian demand functions.

The final sections of the chapter return to producer theory but it is no longer assumed that only one output is produced; we extend the earlier analysis to the case of multiple output and multiple input technologies.

### 3.2 Properties of Cost Functions

The *production function* and the corresponding *cost* function play a central role in many economic applications. In this section, we will show that under certain conditions, the cost function is a sufficient statistic for the corresponding production function; i.e., if we know the cost function of a producer, then this cost function can be used to generate the underlying production function.

Let the producer's *production function*  $f(\mathbf{x})$  denote the maximum amount of output that can be produced in a given time period, given that the producer has access to the nonnegative vector of

inputs,  $\mathbf{x} \equiv [x_1, \dots, x_N] \geq \mathbf{0}_N$ . If the production function satisfies certain regularity conditions,<sup>\*1</sup> then given any positive output level  $y$  that the technology can produce and any strictly positive vector of input prices  $\mathbf{p} \equiv [p_1, \dots, p_N] \geq \mathbf{0}_N$ , we can calculate the producer's *cost function*  $C(y, \mathbf{p})$  as the solution value to the following constrained minimization problem:

$$C(y, \mathbf{p}) \equiv \min_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : f(\mathbf{x}) \geq y; \mathbf{x} \geq \mathbf{0}_N\}. \quad (3.1)$$

It turns out that the cost function  $C$  will satisfy the following 7 properties, irrespective of the properties of the production function  $f$ .

**Theorem 1** Diewert (1982; 537-543)[93]<sup>\*2</sup>: Suppose  $f$  is continuous from above. Then  $C$  defined by (3.1) has the following properties:

*Property 1:*  $C(y, \mathbf{p})$  is a *nonnegative* function.

*Property 2:*  $C(y, \mathbf{p})$  is *positively linearly homogeneous* in  $\mathbf{p}$  for each fixed  $y$ ; i.e.,

$$C(y, \lambda \mathbf{p}) = \lambda C(y, \mathbf{p}) \text{ for all } \lambda > 0, \mathbf{p} \gg \mathbf{0}_N \text{ and } y \in \text{Range } f \text{ (i.e., } y \text{ is an output level that is producible by the production function } f\text{)}. \quad (3.2)$$

*Property 3:*  $C(y, \mathbf{p})$  is *nondecreasing* in  $\mathbf{p}$  for each fixed  $y \in \text{Range } f$ ; i.e.,

$$y \in \text{Range } f, \mathbf{0}_N \ll \mathbf{p}^1 < \mathbf{p}^2 \text{ implies } C(y, \mathbf{p}^1) \leq C(y, \mathbf{p}^2). \quad (3.3)$$

*Property 4:*  $C(y, \mathbf{p})$  is a *concave* function of  $\mathbf{p}$  for each fixed  $y \in \text{Range } f$ ; i.e.,

$$y \in \text{Range } f, \mathbf{p}^1 \gg \mathbf{0}_N; \mathbf{p}^2 \gg \mathbf{0}_N; 0 < \lambda < 1 \text{ implies} \\ C(y, \lambda \mathbf{p}^1 + (1 - \lambda) \mathbf{p}^2) \geq \lambda C(y, \mathbf{p}^1) + (1 - \lambda) C(y, \mathbf{p}^2). \quad (3.4)$$

*Property 5:*  $C(y, \mathbf{p})$  is a *continuous* function of  $\mathbf{p}$  for each fixed  $y \in \text{Range } f$ .

*Property 6:*  $C(y, \mathbf{p})$  is *nondecreasing* in  $y$  for fixed  $\mathbf{p}$ ; i.e.,

$$\mathbf{p} \gg \mathbf{0}_N, y^1 \in \text{Range } f, y^2 \in \text{Range } f, y^1 < y^2 \text{ implies } C(y^1, \mathbf{p}) \leq C(y^2, \mathbf{p}). \quad (3.5)$$

*Property 7:* For every  $\mathbf{p} \gg \mathbf{0}_N$ ,  $C(y, \mathbf{p})$  is *continuous from below* in  $y$ ; i.e.,

$$y^* \in \text{Range } f, y^n \in \text{Range } f \text{ for } n = 1, 2, \dots, y^n \leq y^{n+1}, \lim_{n \rightarrow \infty} y^n = y^* \text{ implies} \\ \lim_{n \rightarrow \infty} C(y^n, \mathbf{p}) = C(y^*, \mathbf{p}). \quad (3.6)$$

**Proof of Property 1:** Let  $y \in \text{Range } f$  and  $\mathbf{p} \gg \mathbf{0}_N$ . Then

$$\begin{aligned} C(y, \mathbf{p}) &\equiv \min_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : f(\mathbf{x}) \geq y; \mathbf{x} \geq \mathbf{0}_N\} \\ &= \mathbf{p}^T \mathbf{x}^* \quad \text{where } \mathbf{x}^* \geq \mathbf{0}_N \text{ and } f(\mathbf{x}^*) \geq y \\ &\geq 0 \quad \text{since } \mathbf{p} \gg \mathbf{0}_N \text{ and } \mathbf{x}^* \geq \mathbf{0}_N. \end{aligned}$$

<sup>\*1</sup> We require that  $f$  be *continuous from above* for the minimum to the cost minimization problem to exist; i.e., for every output level  $y$  that can be produced by the technology (so that  $y \in \text{Range } f$ ), we require that the set of  $\mathbf{x}$ 's that can produce at least output level  $y$  (this is the upper level set  $L(y) \equiv \{\mathbf{x} : f(\mathbf{x}) \geq y\}$ ) is a closed set in  $\mathbb{R}^N$ .

<sup>\*2</sup> For the history of closely related results, see Diewert (1974a; 116-120)[77].

**Proof of Property 2:** Let  $y \in \text{Range } f$ ,  $\mathbf{p} \gg \mathbf{0}_N$  and  $\lambda > 0$ . Then

$$\begin{aligned} C(y, \lambda \mathbf{p}) &\equiv \min_{\mathbf{x}} \{ \lambda \mathbf{p}^T \mathbf{x} : f(\mathbf{x}) \geq y; \mathbf{x} \geq \mathbf{0}_N \} \\ &= \lambda \min_{\mathbf{x}} \{ \mathbf{p}^T \mathbf{x} : f(\mathbf{x}) \geq y; \mathbf{x} \geq \mathbf{0}_N \} \quad \text{since } \lambda > 0 \\ &= \lambda C(y, \mathbf{p}) \quad \text{using the definition of } C(y, \mathbf{p}). \end{aligned}$$

**Proof of Property 3:** Let  $y \in \text{Range } f$ ,  $\mathbf{0}_N \ll \mathbf{p}^1 < \mathbf{p}^2$ . Then

$$\begin{aligned} C(y, \mathbf{p}^2) &\equiv \min_{\mathbf{x}} \{ \mathbf{p}^{2T} \mathbf{x} : f(\mathbf{x}) \geq y; \mathbf{x} \geq \mathbf{0}_N \} \\ &= \mathbf{p}^{2T} \mathbf{x}^* \quad \text{where } f(\mathbf{x}^*) \geq y \text{ and } \mathbf{x}^* \geq \mathbf{0}_N \\ &\geq \mathbf{p}^{1T} \mathbf{x}^* \quad \text{since } \mathbf{x}^* \geq \mathbf{0}_N \text{ and } \mathbf{p}^2 > \mathbf{p}^1 \\ &\geq \min_{\mathbf{x}} \{ \mathbf{p}^{1T} \mathbf{x} : f(\mathbf{x}) \geq y; \mathbf{x} \geq \mathbf{0}_N \} \quad \text{since } \mathbf{x}^* \text{ is feasible for this problem} \\ &\equiv C(y, \mathbf{p}^1). \end{aligned}$$

**Proof of Property 4:** Let  $y \in \text{Range } f$ ,  $\mathbf{p}^1 \gg \mathbf{0}_N$ ;  $\mathbf{p}^2 \gg \mathbf{0}_N$ ;  $0 < \lambda < 1$ . Then

$$\begin{aligned} C(y, \lambda \mathbf{p}^1 + (1 - \lambda) \mathbf{p}^2) &\equiv \min_{\mathbf{x}} \{ [\lambda \mathbf{p}^1 + (1 - \lambda) \mathbf{p}^2]^T \mathbf{x} : f(\mathbf{x}) \geq y; \mathbf{x} \geq \mathbf{0}_N \} \\ &= [\lambda \mathbf{p}^1 + (1 - \lambda) \mathbf{p}^2]^T \mathbf{x}^* \quad \text{where } \mathbf{x}^* \geq \mathbf{0}_N \text{ and } f(\mathbf{x}^*) \geq y \\ &= \lambda \mathbf{p}^{1T} \mathbf{x}^* + (1 - \lambda) \mathbf{p}^{2T} \mathbf{x}^* \\ &\geq \lambda \min_{\mathbf{x}} \{ \mathbf{p}^{1T} \mathbf{x} : f(\mathbf{x}) \geq y; \mathbf{x} \geq \mathbf{0}_N \} + (1 - \lambda) \mathbf{p}^{2T} \mathbf{x}^* \\ &\quad \text{since } \mathbf{x}^* \text{ is feasible for the cost minimization problem that uses} \\ &\quad \text{the price vector } \mathbf{p}^1 \text{ and using also } \lambda > 0 \\ &= \lambda C(y, \mathbf{p}^1) + (1 - \lambda) \mathbf{p}^{2T} \mathbf{x}^* \quad \text{using the definition of } C(y, \mathbf{p}^1) \\ &\geq \lambda C(y, \mathbf{p}^1) + (1 - \lambda) \min_{\mathbf{x}} \{ \mathbf{p}^{2T} \mathbf{x} : f(\mathbf{x}) \geq y; \mathbf{x} \geq \mathbf{0}_N \} \\ &\quad \text{since } \mathbf{x}^* \text{ is feasible for the cost minimization problem that uses} \\ &\quad \text{the price vector } \mathbf{p}^2 \text{ and using also } 1 - \lambda > 0 \\ &= \lambda C(y, \mathbf{p}^1) + (1 - \lambda) C(y, \mathbf{p}^2) \quad \text{using the definition of } C(y, \mathbf{p}^2). \end{aligned}$$

Figure 3.1 below illustrates why this concavity property holds.

In Figure 3.1, the isocost line  $\{\mathbf{x} : \mathbf{p}^{1T} \mathbf{x} = C(y, \mathbf{p}^1)\}$  is tangent to the production possibilities set  $L(y) \equiv \{\mathbf{x} : f(\mathbf{x}) \geq y, \mathbf{x} \geq \mathbf{0}_N\}$  at the point  $\mathbf{x}^1$  and the isocost line  $\{\mathbf{x} : \mathbf{p}^{2T} \mathbf{x} = C(y, \mathbf{p}^2)\}$  is tangent to the production possibilities set  $L(y)$  at the point  $\mathbf{x}^2$ . Note that the point  $\mathbf{x}^{**}$  belongs to both of these isocost lines. Thus  $\mathbf{x}^{**}$  will belong to any weighted average of the two isocost lines. The  $\lambda$  and  $1 - \lambda$  weighted average isocost line is the set  $\{\mathbf{x} : [\lambda \mathbf{p}^1 + (1 - \lambda) \mathbf{p}^2]^T \mathbf{x} = \lambda C(y, \mathbf{p}^1) + (1 - \lambda) C(y, \mathbf{p}^2)\}$  and this set is the dotted line through  $\mathbf{x}^{**}$  in Figure 3.1. Note that this dotted line lies *below*<sup>\*3</sup> the parallel dotted line that is just tangent to  $L(y)$ , which is the isocost line  $\{\mathbf{x} : [\lambda \mathbf{p}^1 + (1 - \lambda) \mathbf{p}^2]^T \mathbf{x} = [\lambda \mathbf{p}^1 + (1 - \lambda) \mathbf{p}^2]^T \mathbf{x}^* = C(y, \lambda \mathbf{p}^1 + (1 - \lambda) \mathbf{p}^2)\}$  and it is this fact that gives us the concavity inequality (3.4).

**Proof of Property 5:** Since  $C(y, \mathbf{p})$  is a concave function of  $\mathbf{p}$  defined over the open set of  $\mathbf{p}$ 's,  $\Omega \equiv \{\mathbf{p} : \mathbf{p} \gg \mathbf{0}_N\}$ , it follows that  $C(y, \mathbf{p})$  is also continuous in  $\mathbf{p}$  over this domain of definition set for each fixed  $y \in \text{Range } f$ .<sup>\*4</sup>

<sup>\*3</sup> It can happen that the two dotted lines coincide.

<sup>\*4</sup> See Fenchel (1953; 75)[179] or Rockafellar (1970; 82)[335].

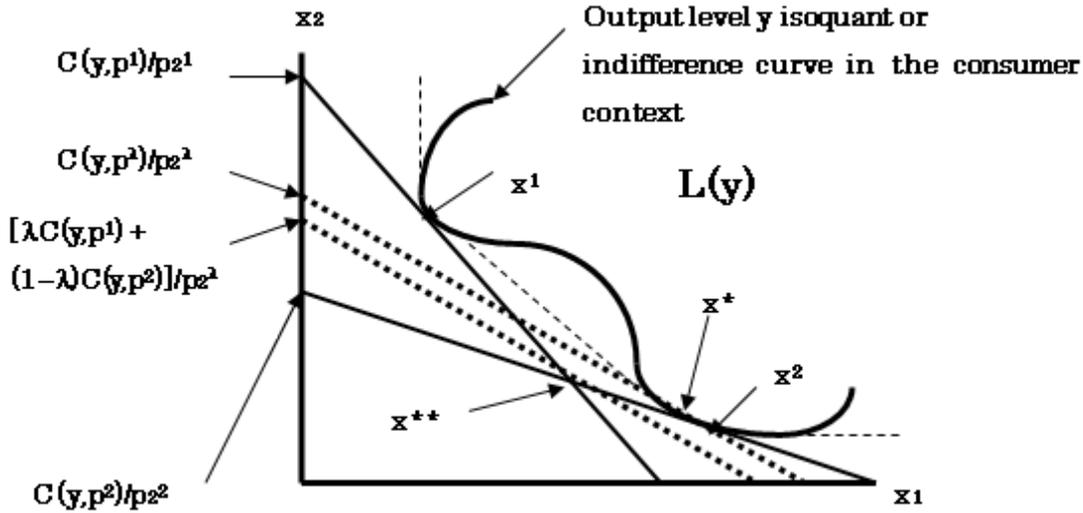


Fig. 3.1 The Concavity in Prices Property of the Cost Function

**Proof of Property 6:** Let  $\mathbf{p} \gg \mathbf{0}_N$ ,  $y^1 \in \text{Range } f$ ,  $y^2 \in \text{Range } f$ ,  $y^1 < y^2$ . Then

$$\begin{aligned} C(y^2, \mathbf{p}) &\equiv \min_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : f(\mathbf{x}) \geq y^2; \mathbf{x} \geq \mathbf{0}_N\} \\ &\geq \min_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : f(\mathbf{x}) \geq y^1; \mathbf{x} \geq \mathbf{0}_N\} \\ &\quad \text{since if } y^1 < y^2, \text{ the set } \{\mathbf{x} : f(\mathbf{x}) \geq y^2\} \text{ is a subset of the set } \{\mathbf{x} : f(\mathbf{x}) \geq y^1\} \text{ and} \\ &\quad \text{the minimum of a linear function over a bigger set cannot increase} \\ &\equiv C(y^1, \mathbf{p}). \end{aligned}$$

**Proof of Property 7:** The proof is rather technical and may be found in Diewert (1993; 113-114)[108].

■

**Problem 1** In industrial organization,<sup>\*5</sup> it used to be fairly common to assume that a firm's cost function had the following linear functional form:  $C(y, \mathbf{p}) \equiv \alpha + \boldsymbol{\beta}^T \mathbf{p} + \gamma y$  where  $\alpha$  and  $\gamma$  are scalar parameters and  $\boldsymbol{\beta}$  is a vector of parameters to be estimated econometrically. What are sufficient conditions on these  $N + 2$  parameters for this cost function to satisfy properties 1 to 7 above? Is the resulting cost function very realistic?

**Problem 2** Suppose a producer's production function,  $f(\mathbf{x})$ , defined for  $\mathbf{x} \in S$  where  $S \equiv \{\mathbf{x} : \mathbf{x} \geq \mathbf{0}_N\}$  satisfies the following conditions:

- (i)  $f$  is continuous over  $S$ ;
- (ii)  $f(\mathbf{x}) > 0$  if  $\mathbf{x} \gg \mathbf{0}_N$  and
- (iii)  $f$  is positively linearly homogeneous over  $S$ ; i.e., for every  $\mathbf{x} \geq \mathbf{0}_N$  and  $\lambda > 0$ ,  $f(\lambda \mathbf{x}) = \lambda f(\mathbf{x})$ .

Define the producer's unit cost function  $c(\mathbf{p})$  for  $\mathbf{p} \gg \mathbf{0}_N$  as follows:

$$c(\mathbf{p}) \equiv C(1, \mathbf{p}) \equiv \min_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : f(\mathbf{x}) \geq 1; \mathbf{x} \geq \mathbf{0}_N\}; \quad (\text{iv})$$

<sup>\*5</sup> For example, see Walters (1961)[394].

i.e.,  $c(\mathbf{p})$  is the minimum cost of producing one unit of output if the producer faces the positive input price vector  $\mathbf{p}$ . For  $y > 0$  and  $\mathbf{p} \gg \mathbf{0}_N$ , show that

$$C(y, \mathbf{p}) = c(\mathbf{p})y. \quad (\text{v})$$

**Note:** A production function  $f$  that satisfies property (iii) is said to exhibit *constant returns to scale*. The interpretation of (v) is that if a production function exhibits constant returns to scale, then total cost is equal to unit cost times the output level.

**Problem 3** Shephard (1953; 4)[355] defined a production function  $F$  to be *homothetic* if it could be written as

$$F(\mathbf{x}) = g[f(\mathbf{x})]; \mathbf{x} \geq \mathbf{0}_N \quad (\text{i})$$

where  $f$  satisfies conditions (i)-(iii) in Problem 2 above and  $g(z)$ , defined for all  $z \geq 0$ , satisfies the following regularity conditions:

- (ii)  $g(z)$  is positive if  $z > 0$ ;
- (iii)  $g$  is a continuous function of one variable and
- (iv)  $g$  is monotonically increasing; i.e., if  $0 \leq z^1 < z^2$ , then  $g(z^1) < g(z^2)$ .

Let  $C(y, \mathbf{p})$  be the cost function that corresponds to  $F(\mathbf{x})$ . Show that under the above assumptions, for  $y > 0$  and  $\mathbf{p} \gg \mathbf{0}_N$ , we have

$$C(y, \mathbf{p}) = g^{-1}(y)c(\mathbf{p}) \quad (\text{v})$$

where  $c(\mathbf{p})$  is the unit cost function that corresponds to the linearly homogeneous  $f$  and  $g^{-1}$  is the inverse function for  $g$ ; i.e.,  $g^{-1}[g(z)] = z$  for all  $z \geq 0$ . Note that  $g^{-1}(y)$  is a monotonically increasing continuous function of one variable.

### 3.3 The Determination of the Production Function from the Cost Function

The material in the previous section shows how the cost function can be determined from a knowledge of the production function. We now ask whether a knowledge of the cost function is sufficient to determine the underlying production function. The answer to this question is *yes*, but with some qualifications.

To see how we might use a given cost function (satisfying the 7 regularity conditions listed in the previous section) to determine the production function that generated it, pick an arbitrary feasible output level  $y > 0$  and an arbitrary vector of positive prices,  $\mathbf{p}^1 \gg \mathbf{0}_N$  and use the given cost function  $C$  to define the following isocost surface:  $\{\mathbf{x} : \mathbf{p}^{1T} \mathbf{x} = C(y, \mathbf{p}^1)\}$ . This isocost surface must be tangent to the set of feasible input combinations  $\mathbf{x}$  that can produce at least output level  $y$ , which is the upper level set,  $L(y) \equiv \{\mathbf{x} : f(\mathbf{x}) \geq y; \mathbf{x} \geq \mathbf{0}_N\}$ . It can be seen that this isocost surface and the set lying above it must contain the upper level set  $L(y)$ ; i.e., the following *halfspace*  $M(y, \mathbf{p}^1)$ , contains  $L(y)$ :

$$M(y, \mathbf{p}^1) \equiv \{\mathbf{x} : \mathbf{p}^{1T} \mathbf{x} \geq C(y, \mathbf{p}^1)\}. \quad (3.7)$$

Pick another positive vector of prices,  $\mathbf{p}^2 \gg \mathbf{0}_N$  and it can be seen, repeating the above argument, that the halfspace  $M(y, \mathbf{p}^2) \equiv \{\mathbf{x} : \mathbf{p}^{2T} \mathbf{x} \geq C(y, \mathbf{p}^2)\}$  must also contain the upper level set  $L(y)$ . Thus  $L(y)$  must belong to the intersection of the two halfspaces  $M(y, \mathbf{p}^1)$  and  $M(y, \mathbf{p}^2)$ . Continuing to argue along these lines, it can be seen that  $L(y)$  must be contained in the following set, which is the intersection of all of the supporting halfspaces to  $L(y)$ :

$$M(y) \equiv \bigcap_{\mathbf{p} \gg \mathbf{0}_N} M(y, \mathbf{p}). \quad (3.8)$$

Note that  $M(y)$  is defined using just the given cost function,  $C(y, \mathbf{p})$ . Note also that since each of the sets in the intersection,  $M(y, \mathbf{p})$ , is a convex set, then  $M(y)$  is also a convex set. Since  $L(y)$  is a subset of each  $M(y, \mathbf{p})$ , it must be the case that  $L(y)$  is also a subset of  $M(y)$ ; i.e., we have

$$L(y) \subset M(y). \quad (3.9)$$

Is it the case that  $L(y)$  is equal to  $M(y)$ ? In general, the answer is *no*;  $M(y)$  forms an *outer approximation* to the true production possibilities set  $L(y)$ . To see why this is, see Figure 3.1 above. The boundary of the set  $M(y)$  partly coincides with the boundary of  $L(y)$  but it encloses a bigger set: the backward bending parts of the isoquant  $\{\mathbf{x} : f(\mathbf{x}) = y\}$  are replaced by the dashed lines that are parallel to the  $x_1$  axis and the  $x_2$  axis and the inward bending part of the true isoquant is replaced by the dashed line that is tangent to the two regions where the boundary of  $M(y)$  coincides with the boundary of  $L(y)$ . However, if the producer is a price taker in input markets, then it can be seen that *we will never observe the producer's nonconvex portions or backwards bending parts of the isoquant*. Thus under the assumption of competitive behavior in input markets, there is no loss of generality in assuming that the producer's production function is *nondecreasing* (this will eliminate the backward bending isoquants) or in assuming that the upper level sets of the production function are convex sets (this will eliminate the nonconvex portions of the upper level sets). Recall that a function has convex upper level sets if and only if it is *quasiconcave*.

Putting the above material together, we see that conditions on the production function  $f(\mathbf{x})$  that are necessary for the sets  $M(y)$  and  $L(y)$  to coincide are:

$$\begin{aligned} \cdot f(\mathbf{x}) \text{ is defined for } \mathbf{x} \geq \mathbf{0}_N \text{ and is continuous from above}^{*6} \text{ over this domain of} \\ \text{definition set;} \end{aligned} \quad (3.10)$$

$$\cdot f \text{ is nondecreasing and} \quad (3.11)$$

$$\cdot f \text{ is quasiconcave.} \quad (3.12)$$

**Theorem 2 Shephard Duality Theorem:**<sup>\*7</sup> If  $f$  satisfies (3.10)-(3.12), then the cost function  $C$  defined by (3.1) satisfies the properties listed in Theorem 1 above and the upper level sets  $M(y)$  defined by (3.8) using only the cost function coincide with the upper level sets  $L(y)$  defined using the production function; i.e., under these regularity conditions, the production function and the cost function determine each other.

We now consider how an explicit formula for the production function in terms of the cost function can be obtained. Suppose we have a given cost function,  $C(y, \mathbf{p})$ , and we are given a strictly positive input vector,  $\mathbf{x} \gg \mathbf{0}_N$ , and we ask what is the maximum output that this  $\mathbf{x}$  can produce. It can be seen that

$$\begin{aligned} f(\mathbf{x}) &= \max_y \{y : \mathbf{x} \in M(y)\} \\ &= \max_y \{y : C(y, \mathbf{p}) \leq \mathbf{p}^T \mathbf{x} \text{ for every } \mathbf{p} \gg \mathbf{0}_N\} \quad \text{using definitions (3.7) and (3.8).} \\ &= \max_y \{y : C(y, \mathbf{p}) \leq 1 \text{ for every } \mathbf{p} \gg \mathbf{0}_N \text{ such that } \mathbf{p}^T \mathbf{x} = 1\} \end{aligned} \quad (3.13)$$

<sup>\*6</sup> Since each of the sets  $M(y, \mathbf{p})$  in the intersection set  $M(y)$  defined by (3.8) are closed, it can be shown that  $M(y)$  is also a closed set. Hence if  $M(y)$  is to coincide with  $L(y)$ , we need the upper level sets of  $f$  to be closed sets and this will hold if and only if  $f$  is continuous from above.

<sup>\*7</sup> Shephard (1953)[355] (1967)[357] (1970)[358] was the pioneer in establishing various duality theorems between cost and production functions. See also Samuelson (1953-54)[343], Uzawa (1964)[379], McFadden (1966)[307] (1978)[308], Diewert (1971)[73] (1974a; 116-118)[77] (1982; 537-545)[93] and Blackorby, Primont and Russell (1978)[39] for various duality theorems under alternative regularity conditions. Our exposition follows that of Diewert (1993; 123-132)[108]. These duality theorems are global in nature; i.e., the production and cost functions satisfy their appropriate regularity conditions over their entire domains of definition. However, it is also possible to develop duality theorems that are local rather than global; see Blackorby and Diewert (1979)[38].

where the last equality follows using the fact that  $C(y, \mathbf{p})$  is linearly homogeneous in  $\mathbf{p}$  as is the function  $\mathbf{p}^T \mathbf{x}$  and hence we can normalize the prices so that  $\mathbf{p}^T \mathbf{x} = 1$ .

We now have to make a bit of a digression and consider the continuity properties of  $C(y, \mathbf{p})$  with respect to  $\mathbf{p}$ . We have defined  $C(y, \mathbf{p})$  for all strictly positive price vectors  $\mathbf{p}$  and since this domain of definition set is open, we know that  $C(y, \mathbf{p})$  is also continuous in  $\mathbf{p}$  over this set, using the concavity in prices property of  $C$ . We now would like to extend the domain of definition of  $C(y, \mathbf{p})$  from the strictly positive orthant of prices,  $\Omega \equiv \{\mathbf{p} : \mathbf{p} \gg \mathbf{0}_N\}$ , to the nonnegative orthant,  $\text{Clo } \Omega \equiv \{\mathbf{p} : \mathbf{p} \geq \mathbf{0}_N\}$ , which is the closure of  $\Omega$ . It turns out that it is possible to do this if we make use of some theorems in convex analysis.

**Theorem 3** *Continuity from above of a concave function using the Fenchel closure operation:* Fenchel (1953; 78)[179]: Let  $f(\mathbf{x})$  be a concave function of  $N$  variables defined over the open convex subset  $S$  of  $\mathbb{R}^N$ . Then there exists a unique extension of  $f$  to  $\text{Clo } S$ , the closure of  $S$ , which is concave and continuous from above.

**Proof.** By the second characterization of concavity, the hypograph of  $f$ ,  $H \equiv \{(y, \mathbf{x}) : y \leq f(\mathbf{x}); \mathbf{x} \in S\}$ , is a convex set in  $\mathbb{R}^{N+1}$ . Hence the closure of  $H$ ,  $\text{Clo } H$ , is also a convex set. Hence the following function  $f^*$  defined over  $\text{Clo } S$  is also a concave function:

$$\begin{aligned} f^*(\mathbf{x}) &\equiv \max_y \{y : (y, \mathbf{x}) \in \text{Clo } H\}; & \mathbf{x} \in \text{Clo } S. \\ &= f(\mathbf{x}) & \text{for } \mathbf{x} \in S. \end{aligned} \quad (3.14)$$

Since  $\text{Clo } H$  is a closed set, it turns out that  $f^*$  is continuous from above. ■

To see that the extension function  $f^*$  need not be continuous, consider the following *example*, where the domain of definition set is  $S \equiv \{(x_1, x_2); x_2 \in \mathbb{R}^1, x_1 \geq x_2^2\}$  in  $\mathbb{R}^2$ :

$$f(x_1, x_2) \equiv \begin{cases} -x_2^2/x_1 & \text{if } x_2 \neq 0, x_1 \geq x_2^2; \\ 0 & \text{if } x_1 = 0 \text{ and } x_2 = 0. \end{cases} \quad (3.15)$$

It is possible to show that  $f$  is concave and hence continuous over the interior of  $S$ ; see problem 5 below. However, we show that  $f$  is not continuous at  $(0, 0)$ . Let  $(x_1, x_2)$  approach  $(0, 0)$  along the line  $x_1 = x_2 > 0$ . Then

$$\lim_{x_1 \rightarrow 0} f(x_1, x_2) = \lim_{x_1 \rightarrow 0} [-x_1^2/x_1] = \lim_{x_1 \rightarrow 0} [-x_1] = 0. \quad (3.16)$$

Now let  $(x_1, x_2)$  approach  $(0, 0)$  along the parabolic path  $x_2 > 0$  and  $x_1 = x_2^2$ . Then

$$\lim_{x_2 \rightarrow 0; x_1 = x_2^2} f(x_1, x_2) = \lim_{x_2 \rightarrow 0} -x_2^2/x_2^2 = -1. \quad (3.17)$$

Thus  $f$  is not continuous at  $(0, 0)$ . It can be verified that restricting  $f$  to  $\text{Int } S$  and then extending  $f$  to the closure of  $S$  (which is  $S$ ) leads to the same  $f^*$  as is defined by (3.15). Thus the Fenchel closure operation does not always result in a continuous concave function.

Theorem 4 below states sufficient conditions for the Fenchel closure of a concave function defined over an open domain of definition set to be continuous over the closure of the original domain of definition. Fortunately, the hypotheses of this Theorem are weak enough to cover most economic applications. Before stating the theorem, we need an additional definition.

**Definition** A set  $S$  in  $\mathbb{R}^N$  is a *polyhedral set* iff  $S$  is equal to the intersection of a *finite* number of halfspaces.

**Theorem 4** *Continuity of a concave function using the Fenchel closure operation*; Gale, Klee and Rockafellar (1968)[196], Rockafellar (1970; 85)[335]: Let  $f$  be a concave function of  $N$  variables defined over an open convex polyhedral set  $S$ . Suppose  $f$  is bounded from below over every bounded subset of  $S$ . Then the Fenchel closure extension of  $f$  to the closure of  $S$  results in a continuous concave function defined over  $\text{Clo } S$ .

The proof of this result is a bit too involved for us to reproduce here but we can now apply this result.

Applying Theorem 4, we can extend the domain of definition of  $C(y, \mathbf{p})$  from strictly positive price vectors  $\mathbf{p}$  to nonnegative price vectors using the Fenchel closure operation and hence  $C(y, \mathbf{p})$  will be continuous and concave in  $\mathbf{p}$  over the set  $\{\mathbf{p} : \mathbf{p} \geq \mathbf{0}_N\}$  for each  $y$  in the interval of feasible outputs.\*<sup>8</sup>

Now we can return to the problem where we have a given cost function,  $C(y, \mathbf{p})$ , we are given a strictly positive input vector,  $\mathbf{x} \gg \mathbf{0}_N$ , and we ask what is the maximum output that this  $\mathbf{x}$  can produce. Repeating the analysis in (3.13), we have

$$\begin{aligned}
 f(\mathbf{x}) &= \max_y \{y : \mathbf{x} \in M(y)\} \\
 &= \max_y \{y : C(y, \mathbf{p}) \leq \mathbf{p}^T \mathbf{x} \text{ for every } \mathbf{p} \gg \mathbf{0}_N\} \quad \text{using definitions (3.7) and (3.8).} \\
 &= \max_y \{y : C(y, \mathbf{p}) \leq 1 \text{ for every } \mathbf{p} \gg \mathbf{0}_N \text{ such that } \mathbf{p}^T \mathbf{x} = 1\} \\
 &\quad \text{where we have used the linear homogeneity in prices property of } C \\
 &= \max_y \{y : C(y, \mathbf{p}) \leq 1 \text{ for every } \mathbf{p} \geq \mathbf{0}_N \text{ such that } \mathbf{p}^T \mathbf{x} = 1\} \\
 &\quad \text{where we have extended the domain of definition of } C(y, \mathbf{p}) \text{ to nonnegative prices} \\
 &\quad \text{from positive prices and used the continuity of the extension function over the set} \\
 &\quad \text{of nonnegative prices} \\
 &= \max_y \{y : G(y, \mathbf{x}) \leq 1\} \tag{3.18}
 \end{aligned}$$

where the function  $G(y, \mathbf{x})$  is defined as follows:

$$G(y, \mathbf{x}) \equiv \max_{\mathbf{p}} \{C(y, \mathbf{p}) : \mathbf{p} \geq \mathbf{0}_N \text{ and } \mathbf{p}^T \mathbf{x} = 1\}. \tag{3.19}$$

Note that the maximum in (3.19) will exist since  $C(y, \mathbf{p})$  is continuous in  $\mathbf{p}$  and the feasible region for the maximization problem,  $\{\mathbf{p} : \mathbf{p} \geq \mathbf{0}_N \text{ and } \mathbf{p}^T \mathbf{x} = 1\}$ , is a closed and bounded set.\*<sup>9</sup> Property 7 on the cost function  $C(y, \mathbf{p})$  will imply that the maximum in the last line of (3.18) will exist. Property 6 on the cost function will imply that for fixed  $\mathbf{x}$ ,  $G(y, \mathbf{x})$  is nondecreasing in  $y$ . Typically,  $G(y, \mathbf{x})$  will be continuous in  $y$  for a fixed  $\mathbf{x}$  and so the maximum  $y$  that solves (3.18) will be the  $y^*$  that satisfies the following equation:\*<sup>10</sup>

$$G(y^*, \mathbf{x}) = 1. \tag{3.20}$$

Thus (3.19) and (3.20) implicitly define the production function  $y^* = f(\mathbf{x})$  in terms of the cost function  $C$ .

**Problem 4** Show that the  $f(x_1, x_2)$  defined by (3.15) above is a concave function over the interior of the domain of definition set  $S$ . You do not have to show that  $S$  is a convex set.

\*<sup>8</sup> If  $f(\mathbf{0}_N) = 0$  and  $f(\mathbf{x})$  tends to plus infinity as the components of  $\mathbf{x}$  tend to plus infinity, then the feasible  $y$  set will be  $y \geq 0$  and  $C(y, \mathbf{p})$  will be defined for all  $y \geq 0$  and  $\mathbf{p} \geq \mathbf{0}_N$ .

\*<sup>9</sup> Here is where we use the assumption that  $\mathbf{x} \gg \mathbf{0}_N$  in order to obtain the boundedness of this set.

\*<sup>10</sup> This method for constructing the production function from the cost function may be found in Diewert (1974a; 119)[77].

**Problem 5** In the case where the technology is subject to constant returns to scale, the cost function has the following form:  $C(y, \mathbf{p}) = yc(\mathbf{p})$  where  $c(\mathbf{p})$  is a unit cost function. For  $\mathbf{x} \gg \mathbf{0}_N$ , define the function  $g(\mathbf{x})$  as follows:

$$g(\mathbf{x}) \equiv \max_{\mathbf{p}} \{c(\mathbf{p}) : \mathbf{p}^T \mathbf{x} = 1; \mathbf{p} \geq \mathbf{0}_N\}. \quad (\text{i})$$

Show that in this constant returns to scale case, the function  $G(y, \mathbf{x})$  defined by (3.19) reduces to

$$G(y, \mathbf{x}) = yg(\mathbf{x}). \quad (\text{ii})$$

Show that in this constant returns to scale case, the production function that is dual to the cost function has the following explicit formula for  $\mathbf{x} \gg \mathbf{0}_N$ :

$$f(\mathbf{x}) = 1/g(\mathbf{x}). \quad (\text{iii})$$

**Problem 6** Let  $x \geq 0$  be input (a scalar number) and let  $y = f(x) \geq 0$  be the maximum output that could be produced by input  $x$ , where  $f$  is the production function. Suppose that  $f$  is defined as the following *step function*:

$$f(x) \equiv \begin{cases} 0 & \text{for } 0 \leq x < 1; \\ 1 & \text{for } 1 \leq x < 2; \\ 2 & \text{for } 2 \leq x < 3; \end{cases} \quad (\text{i})$$

and so on. Thus the technology cannot produce fractional units of output and it takes one full unit of input to produce each unit of output. It can be verified that this production function is continuous from above.

(a) Calculate the cost function  $C(y, 1)$  that corresponds to this production function; i.e., set the input price equal to one and try to determine the corresponding total cost function  $C(y, 1)$ . (It will turn out that this cost function is continuous from below in  $y$  but it is not necessary to prove this).

(b) Graph both the production function  $y = f(x)$  and the cost function  $c = C(y, 1)$ .

**Problem 7** Suppose that a producer's cost function is defined as follows for  $y \geq 0, p_1 > 0$  and  $p_2 > 0$ :

$$C(y, p_1, p_2) \equiv [b_{11}p_1 + 2b_{12}(p_1p_2)^{1/2} + b_{22}p_2]y \quad (\text{i})$$

where the  $b_{ij}$  parameters are all positive.

(a) Show that this cost function is concave in the input prices  $p_1, p_2$ . *Note:* this is the two input case of the Generalized Leontief cost function defined by Diewert (1971)[73].

(b) Calculate an explicit functional form for the corresponding production function  $f(x_1, x_2)$  where we assume that  $x_1 > 0$  and  $x_2 > 0$ .

### 3.4 The Derivative Property of the Cost Function

Up to this point, Theorem 2, the Shephard Duality Theorem, is of mainly academic interest: if the production function  $f$  satisfies properties (3.10)-(3.12), then the corresponding cost function  $C$  defined by (3.1) satisfies the properties listed in Theorem 1 above and moreover completely determines the production function. However, it is the next property of the cost function that makes duality theory so useful in applied economics.

**Theorem 5** *Shephard's* (1953; 11)[355] *Lemma:* If the cost function  $C(y, \mathbf{p})$  satisfies the properties listed in Theorem 1 above and in addition is once differentiable with respect to the components of

input prices at the point  $(y^*, \mathbf{p}^*)$  where  $y^*$  is in the range of the production function  $f$  and  $\mathbf{p}^* \gg \mathbf{0}_N$ , then

$$\mathbf{x}^* = \nabla_p C(y^*, \mathbf{p}^*) \quad (3.21)$$

where  $\nabla_p C(y^*, \mathbf{p}^*)$  is the vector of first order partial derivatives of cost with respect to input prices,  $[\partial C(y^*, \mathbf{p}^*)/\partial p_1, \dots, \partial C(y^*, \mathbf{p}^*)/\partial p_N]^T$ , and  $\mathbf{x}^*$  is any solution to the cost minimization problem

$$\min_{\mathbf{x}} \{\mathbf{p}^{*T} \mathbf{x} : f(\mathbf{x}) \geq y^*\} \equiv C(y^*, \mathbf{p}^*). \quad (3.22)$$

Under these differentiability hypotheses, it turns out that the  $\mathbf{x}^*$  solution to (3.22) is unique.

**Proof.** Let  $\mathbf{x}^*$  be any solution to the cost minimization problem (3.22). Since  $\mathbf{x}^*$  is feasible for the cost minimization problem when the input price vector is changed to an arbitrary  $\mathbf{p} \gg \mathbf{0}_N$ , it follows that

$$\mathbf{p}^T \mathbf{x}^* \geq C(y^*, \mathbf{p}) \quad \text{for every } \mathbf{p} \gg \mathbf{0}_N. \quad (3.23)$$

Since  $\mathbf{x}^*$  is a solution to the cost minimization problem (3.22) when  $\mathbf{p} = \mathbf{p}^*$ , we must have

$$\mathbf{p}^{*T} \mathbf{x}^* = C(y^*, \mathbf{p}^*). \quad (3.24)$$

But (3.23) and (3.24) imply that the function of  $N$  variables,  $g(\mathbf{p}) \equiv \mathbf{p}^T \mathbf{x}^* - C(y^*, \mathbf{p})$  is nonnegative for all  $\mathbf{p} \gg \mathbf{0}_N$  with  $g(\mathbf{p}^*) = 0$ . Hence,  $g(\mathbf{p})$  attains a global minimum at  $\mathbf{p} = \mathbf{p}^*$  and since  $g(\mathbf{p})$  is differentiable with respect to the input prices  $\mathbf{p}$  at this point, the following first order necessary conditions for a minimum must hold at this point:

$$\nabla_p g(\mathbf{p}^*) = \mathbf{x}^* - \nabla_p C(y^*, \mathbf{p}^*) = \mathbf{0}_N. \quad (3.25)$$

Now note that (3.25) is equivalent to (3.21). If  $\mathbf{x}^{**}$  is any other solution to the cost minimization problem (3.22), then repeat the above argument to show that

$$\begin{aligned} \mathbf{x}^{**} &= \nabla_p C(y^*, \mathbf{p}^*) \\ &= \mathbf{x}^* \end{aligned} \quad (3.26)$$

where the second equality follows using (3.25). Hence  $\mathbf{x}^{**} = \mathbf{x}^*$  and the solution to (3.22) is unique. ■

The above result has the following implication: postulate a differentiable functional form for the cost function  $C(y, \mathbf{p})$  that satisfies the regularity conditions listed in Theorem 1 above. Then differentiating  $C(y, \mathbf{p})$  with respect to the components of the input price vector  $\mathbf{p}$  generates the firm's system of cost minimizing input demand functions,  $\mathbf{x}(y, \mathbf{p}) \equiv \nabla_p C(y, \mathbf{p})$ .

Shephard (1953)[355] was the first person to establish the above result starting with just a cost function satisfying the appropriate regularity conditions.<sup>\*11</sup> However, Hotelling (1932; 594)[242] stated a version of the result in the context of profit functions and Hicks (1946; 331)[222] and Samuelson (1953-54; 15-16)[343] established the result starting with a differentiable utility or production function.

One application of the above result is its use as an aid in generating systems of cost minimizing input demand functions that are linear in the parameters that characterize the technology. For example, suppose that the cost function had the following *Generalized Leontief functional form*:<sup>\*12</sup>

$$C(y, \mathbf{p}) \equiv \sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{1/2} p_j^{1/2} y; \quad b_{ij} = b_{ji} \text{ for } 1 \leq i < j \leq N \quad (3.27)$$

<sup>\*11</sup> See also Fenchel (1953; 104)[179]. We have used the technique of proof used by McKenzie (1956-57)[310].

<sup>\*12</sup> See Diewert (1971)[73].

where the  $N(N + 1)/2$  independent  $b_{ij}$  parameters are all nonnegative. With these nonnegativity restrictions, it can be verified that the  $C(y, \mathbf{p})$  defined by (3.27) satisfies properties 1 to 7 listed in Theorem 1.\*<sup>13</sup> Applying Shephard's Lemma shows that the system of cost minimizing input demand functions that correspond to this functional form are given by:

$$x_i(y, \mathbf{p}) = \partial C(y, \mathbf{p}) / \partial p_i = \sum_{j=1}^N b_{ij} (p_j / p_i)^{1/2} y; \quad i = 1, 2, \dots, N. \quad (3.28)$$

Errors can be added to the system of equations (3.28) and the parameters  $b_{ij}$  can be estimated using linear regression techniques if we have time series or cross sectional data on output, inputs and input prices.\*<sup>14</sup> If all of the  $b_{ij}$  equal zero for  $i \neq j$ , then the demand functions become:

$$x_i(y, \mathbf{p}) = \partial C(y, \mathbf{p}) / \partial p_i = b_{ii} y; \quad i = 1, 2, \dots, N. \quad (3.29)$$

Note that input prices do not appear in the system of input demand functions defined by (3.29) so that input quantities do not respond to changes in the relative prices of inputs. The corresponding production function is known as the Leontief (1941)[290] production function.\*<sup>15</sup> Hence, it can be seen that the production function that corresponds to (3.28) is a generalization of this production function. The unit output isoquant for the Leontief production function is graphed below in Figure 3.2.

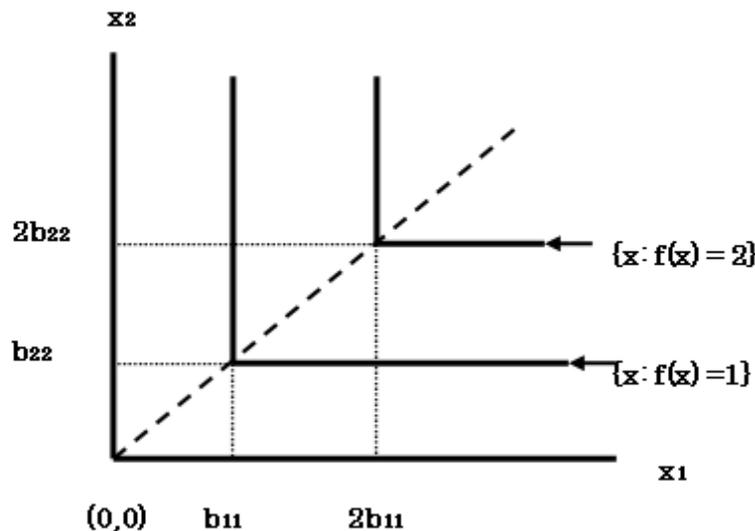


Fig. 3.2 The Two Input Leontief Production Function

\*<sup>13</sup> Using problem 7 above, it can be seen that if the  $b_{ij}$  are nonnegative and  $y$  is positive, then the functions  $b_{ij} p_i^{1/2} p_j^{1/2} y$  are concave in the components of  $\mathbf{p}$ . Hence, since a sum of concave functions is concave, it can be seen that the  $C(y, \mathbf{p})$  defined by (3.27) is concave in the components of  $\mathbf{p}$ .

\*<sup>14</sup> Note that  $b_{12}$  will appear in the first input demand equation and in the second as well using the cross equation symmetry condition,  $b_{21} = b_{12}$ . There are  $N(N - 1)/2$  such cross equation symmetry conditions and we could test for their validity or impose them in order to save degrees of freedom. The nonnegativity restrictions that ensure global concavity of  $C(y, \mathbf{p})$  in  $\mathbf{p}$  can be imposed if we replace each parameter  $b_{ij}$  by a squared parameter,  $(a_{ij})^2$ . However, the resulting system of estimating equations is no longer linear in the unknown parameters.

\*<sup>15</sup> The Leontief production function can be defined as  $f(x_1, \dots, x_N) \equiv \min_i \{x_i / b_{ii} : i = 1, \dots, N\}$ . It is also known as the no substitution production function. Note that this production function is not differentiable even though its cost function is differentiable.

### 3.5 The Comparative Statics Properties of Input Demand Functions

Before we develop the main result in this section, it will be useful to establish some results about the derivatives of a twice continuously differentiable linearly homogeneous function of  $N$  variables. We say that  $f(\mathbf{x})$ , defined for  $\mathbf{x} \gg \mathbf{0}_N$  is *positively homogeneous of degree  $\alpha$*  iff  $f$  has the following property:

$$f(\lambda \mathbf{x}) = \lambda^\alpha f(\mathbf{x}) \quad \text{for all } \mathbf{x} \gg \mathbf{0}_N \text{ and } \lambda > 0. \quad (3.30)$$

A special case of the above definition occurs when the number  $\alpha$  in the above definition equals 1. In this case, we say that  $f$  is (positively) *linearly homogeneous*<sup>\*16</sup> iff

$$f(\lambda \mathbf{x}) = \lambda f(\mathbf{x}) \quad \text{for all } \mathbf{x} \gg \mathbf{0}_N \text{ and } \lambda > 0. \quad (3.31)$$

**Theorem 6** *Euler's Theorems on Differentiable Homogeneous Functions:* Let  $f(\mathbf{x})$  be a (positively) linearly homogeneous function of  $N$  variables, defined for  $\mathbf{x} \gg \mathbf{0}_N$ .

**Part 1:** If the first order partial derivatives of  $f$  exist, then the first order partial derivatives of  $f$  satisfy the following equation:

$$f(\mathbf{x}) = \sum_{n=1}^N x_n \frac{\partial f(x_1, \dots, x_N)}{\partial x_n} = \mathbf{x}^T \nabla f(\mathbf{x}) \quad \text{for all } \mathbf{x} \gg \mathbf{0}_N. \quad (3.32)$$

**Part 2:** If the second order partial derivatives of  $f$  exist, then they satisfy the following equations:

$$\sum_{j=1}^N \frac{\partial^2 f(x_1, \dots, x_N)}{\partial x_n \partial x_j} x_j = 0 \quad \text{for all } \mathbf{x} \gg \mathbf{0}_N \text{ and } n = 1, \dots, N. \quad (3.33)$$

The  $N$  equations in (3.33) can be written using matrix notation in a much more compact form as follows:

$$\nabla^2 f(\mathbf{x}) \mathbf{x} = \mathbf{0}_N \quad \text{for all } \mathbf{x} \gg \mathbf{0}_N. \quad (3.34)$$

**Proof of Part 1:** Let  $\mathbf{x} \gg \mathbf{0}_N$  and  $\lambda > 0$ . Differentiating both sides of (3.31) with respect to  $\lambda$  leads to the following equation using the composite function chain rule:

$$\begin{aligned} f(\mathbf{x}) &= \sum_{n=1}^N \frac{\partial f(\lambda x_1, \dots, \lambda x_N)}{\partial(\lambda x_n)} \frac{\partial(\lambda x_n)}{\partial \lambda} \\ &= \sum_{n=1}^N \frac{\partial f(\lambda x_1, \dots, \lambda x_N)}{\partial(\lambda x_n)} x_n. \end{aligned} \quad (3.35)$$

Now evaluate (3.35) at  $\lambda = 1$  and we obtain (3.32).

**Proof of Part 2:** Let  $\mathbf{x} \gg \mathbf{0}_N$  and  $\lambda > 0$ . For  $n = 1, \dots, N$ , differentiate both sides of (3.31) with respect to  $x_n$  and we obtain the following  $N$  equations:

$$\begin{aligned} f_n(\lambda x_1, \dots, \lambda x_N) \frac{\partial(\lambda x_n)}{\partial x_n} &= \lambda f_n(x_1, \dots, x_N) & \text{for } n = 1, \dots, N \text{ or} \\ f_n(\lambda x_1, \dots, \lambda x_N) \lambda &= \lambda f_n(x_1, \dots, x_N) & \text{for } n = 1, \dots, N \text{ or} \\ f_n(\lambda x_1, \dots, \lambda x_N) &= f_n(x_1, \dots, x_N) & \text{for } n = 1, \dots, N \end{aligned} \quad (3.36)$$

<sup>\*16</sup> Usually in economics, we omit the adjective “positively” but it is understood that the  $\lambda$  which appears in definitions (3.30) and (3.31) is restricted to be positive.

where the  $n$ th first order partial derivative function is defined as  $f_n(x_1, \dots, x_N) \equiv \partial f(x_1, \dots, x_N) / \partial x_n$  for  $n = 1, \dots, N$ .<sup>\*17</sup> Now differentiate both sides of the last set of equations in (3.36) with respect to  $\lambda$  and we obtain the following  $N$  equations:

$$\begin{aligned} 0 &= \sum_{j=1}^N \frac{\partial f_n(\lambda x_1, \dots, \lambda x_N)}{\partial(\lambda x_j)} \frac{\partial(\lambda x_j)}{\partial \lambda} \quad \text{for } n = 1, \dots, N \\ &= \sum_{j=1}^N \frac{\partial f_n(\lambda x_1, \dots, \lambda x_N)}{\partial x_j} x_j. \end{aligned} \quad (3.37)$$

Now evaluate (3.37) at  $\lambda = 1$  and we obtain the  $N$  equations (3.33). ■

The above results can be applied to the cost function,  $C(y, \mathbf{p})$ . From Theorem 1,  $C(y, \mathbf{p})$  is linearly homogeneous in  $\mathbf{p}$ . Hence by part 2 of Euler's Theorem, if the second order partial derivatives of the cost function with respect to the components of the input price vector  $\mathbf{p}$  exist, then these derivatives satisfy the following restrictions:

$$\nabla_{pp}^2 C(y, \mathbf{p}) \mathbf{p} = \mathbf{0}_N. \quad (3.38)$$

**Theorem 7** Diewert (1982; 567)[93]: Suppose the cost function  $C(y, \mathbf{p})$  satisfies the properties listed in Theorem 1 and in addition is twice continuously differentiable with respect to the components of its input price vector at some point,  $(y, \mathbf{p})$ . Then the system of cost minimizing input demand equations,  $\mathbf{x}(y, \mathbf{p}) \equiv [x_1(y, \mathbf{p}), \dots, x_N(y, \mathbf{p})]^T$ , exists at this point and these input demand functions are once continuously differentiable. Form the  $N \times N$  matrix of input demand derivatives with respect to input prices,  $\mathbf{B} \equiv [\partial x_i(y, \mathbf{p}) / \partial p_j]$ , which has  $ij$  element equal to  $\partial x_i(y, \mathbf{p}) / \partial p_j$ . Then the matrix  $\mathbf{B}$  has the following properties:

$$\cdot \mathbf{B} = \mathbf{B}^T \text{ so that } \partial x_i(y, \mathbf{p}) / \partial p_j = \partial x_j(y, \mathbf{p}) / \partial p_i \text{ for all } i \neq j; \text{ }^{*18} \quad (3.39)$$

$$\cdot \mathbf{B} \text{ is negative semidefinite}^{*19} \text{ and} \quad (3.40)$$

$$\cdot \mathbf{B}\mathbf{p} = \mathbf{0}_N. \text{ }^{*20} \quad (3.41)$$

**Proof.** Shephard's Lemma implies that the firm's system of cost minimizing input demand equations,  $\mathbf{x}(y, \mathbf{p}) \equiv [x_1(y, \mathbf{p}), \dots, x_N(y, \mathbf{p})]^T$ , exists and is equal to

$$\mathbf{x}(y, \mathbf{p}) = \nabla_p C(y, \mathbf{p}). \quad (3.42)$$

Differentiating both sides of (3.42) with respect to the components of  $\mathbf{p}$  gives us

$$\mathbf{B} \equiv [\partial x_i(y, \mathbf{p}) / \partial p_j] = \nabla_{pp}^2 C(y, \mathbf{p}). \quad (3.43)$$

Now property (3.39) follows from Young's Theorem in calculus. Property (3.40) follows from (3.43) and the fact that  $C(y, \mathbf{p})$  is concave in  $\mathbf{p}$  and the fourth characterization of concavity. Finally, property (3.41) follows from the fact that the cost function is linearly homogeneous in  $\mathbf{p}$  and hence (3.38) holds. ■

<sup>\*17</sup> Using definition (3.30) for the case where  $\alpha = 0$ , it can be seen that the last set of equations in (3.36) shows that the first order partial derivative functions of a linearly homogenous function are homogeneous of degree 0.

<sup>\*18</sup> These are the Hicks (1946; 311)[222] and Samuelson (1947; 69)[340] symmetry restrictions. Hotelling (1932; 549)[242] obtained analogues to these symmetry conditions in the profit function context.

<sup>\*19</sup> Hicks (1946; 311)[222] and Samuelson (1947; 69)[340] also obtained versions of this result by starting with the production (or utility) function  $F(\mathbf{x})$ , assuming that the first order conditions for solving the cost minimization problem held and that the strong second order sufficient conditions for the primal cost minimization problem also held.

<sup>\*20</sup> Hicks (1946; 331)[222] and Samuelson (1947; 69)[340] also obtained this result using their primal technique.

Note that property (3.40) implies the following properties on the input demand functions:

$$\frac{\partial x_n(y, \mathbf{p})}{\partial p_n} \leq 0 \quad \text{for } n = 1, \dots, N. \quad (3.44)$$

Property (3.44) means that input demand curves cannot be upward sloping.

If the cost function is also differentiable with respect to the output variable  $y$ , then we can deduce an additional property about the first order derivatives of the input demand functions. The linear homogeneity property of  $C(y, \mathbf{p})$  in  $\mathbf{p}$  implies that the following equation holds for all  $\lambda > 0$ :

$$C(y, \lambda \mathbf{p}) = \lambda C(y, \mathbf{p}) \quad \text{for all } \lambda > 0 \text{ and } \mathbf{p} \gg \mathbf{0}_N. \quad (3.45)$$

Partially differentiating both sides of (3.45) with respect to  $y$  leads to the following equation:

$$\frac{\partial C(y, \lambda \mathbf{p})}{\partial y} = \lambda \frac{\partial C(y, \mathbf{p})}{\partial y} \quad \text{for all } \lambda > 0 \text{ and } \mathbf{p} \gg \mathbf{0}_N. \quad (3.46)$$

But (3.46) implies that the function  $\partial C(y, \mathbf{p})/\partial y$  is linearly homogeneous in  $\mathbf{p}$  and hence part 1 of Euler's Theorem applied to this function gives us the following equation:

$$\frac{\partial C(y, \mathbf{p})}{\partial y} = \sum_{n=1}^N p_n \frac{\partial^2 C(y, \mathbf{p})}{\partial y \partial p_n} = \mathbf{p}^T \nabla_{y\mathbf{p}}^2 C(y, \mathbf{p}). \quad (3.47)$$

But using (3.42), it can be seen that (3.47) is equivalent to the following equation:<sup>\*21</sup>

$$\frac{\partial C(y, \mathbf{p})}{\partial y} = \sum_{n=1}^N p_n \frac{\partial x_n(y, \mathbf{p})}{\partial y}. \quad (3.48)$$

**Problem 8** For  $i \neq j$ , the inputs  $i$  and  $j$  are said to be *substitutes* if  $\partial x_i(y, \mathbf{p})/\partial p_j = \partial x_j(y, \mathbf{p})/\partial p_i > 0$ , *unrelated* if  $\partial x_i(y, \mathbf{p})/\partial p_j = \partial x_j(y, \mathbf{p})/\partial p_i = 0$ <sup>\*22</sup>, and *complements* if  $\partial x_i(y, \mathbf{p})/\partial p_j = \partial x_j(y, \mathbf{p})/\partial p_i < 0$ .

(a) If  $N = 2$ , show that the two inputs cannot be complements.

(b) If  $N = 2$  and  $\partial x_1(y, \mathbf{p})/\partial p_1 = 0$ , then show that all of the remaining input demand price derivatives are equal to 0; i.e., show that  $\partial x_1(y, \mathbf{p})/\partial p_2 = \partial x_2(y, \mathbf{p})/\partial p_1 = \partial x_2(y, \mathbf{p})/\partial p_2 = 0$ .

(c) If  $N = 3$ , show that at most one pair of inputs can be complements.<sup>\*23</sup>

**Problem 9** Let  $N \geq 3$  and suppose that  $\partial x_1(y, \mathbf{p})/\partial p_1 = 0$ . Then show that  $\partial x_1(y, \mathbf{p})/\partial p_n = 0$  as well for  $n = 2, 3, \dots, N$ .

*Hint:* You will need to use the definition of negative semidefiniteness in a strategic way. This problem shows that if the own input elasticity of demand for an input is 0, then that input is unrelated to all other inputs.

**Problem 10** Recall the definition (3.27) of the Generalized Leontief cost function where the parameters  $b_{ij}$  were all assumed to be nonnegative. Show that under these nonnegativity restrictions, every input pair is either unrelated or substitutes.

<sup>\*21</sup> This method of deriving these restrictions is due to Diewert (1982; 568)[93] but these restrictions were originally derived by Samuelson (1947; 66)[340] using his primal cost minimization method.

<sup>\*22</sup> Pollak (1969; 67)[330] uses the term "unrelated" in a similar context.

<sup>\*23</sup> This result is due to Hicks (1946; 311-312)[222]: "It follows at once from Rule (5) that, while it is possible for all other goods consumed to be substitutes for  $x_r$ , it is not possible for them all to be complementary with it."

*Hint:* Simply calculate  $\partial^2 C(y, \mathbf{p}) / \partial p_i \partial p_j$  for  $i \neq j$  and look at the resulting formula.

*Comment:* This result shows that if we impose the nonnegativity conditions  $b_{ij} \geq 0$  for  $i \neq j$  on this functional form in order to ensure that it is globally concave in prices, then we have a priori ruled out any form of complementarity between the inputs. This means if the number of inputs  $N$  is greater than 2, this nonnegativity restricted functional form cannot be a *flexible functional form*\*<sup>24</sup> for a cost function; i.e., it cannot attain an arbitrary pattern of demand derivatives that are consistent with microeconomic theory, since the nonnegativity restrictions rule out any form of complementarity.

**Problem 11** Suppose that a producer's three input production function has the following Cobb Douglas (1928)[59] functional form:

$$f(x_1, x_2, x_3) \equiv x_1^{\alpha_1} x_2^{\alpha_2} x_3^{\alpha_3} \quad \text{where } \alpha_1 > 0, \alpha_2 > 0, \alpha_3 > 0 \text{ and } \alpha_1 + \alpha_2 + \alpha_3 = 1. \quad (\text{a})$$

Let the positive input prices  $p_1 > 0, p_2 > 0, p_3 > 0$  and the positive output level  $y > 0$  be given.

(i) Calculate the producer's cost function,  $C(y, p_1, p_2, p_3)$  along with the three input demand functions,  $x_1(y, p_1, p_2, p_3), x_2(y, p_1, p_2, p_3)$  and  $x_3(y, p_1, p_2, p_3)$ .

*Hint:* Use the usual Lagrangian technique for solving constrained minimization problems. You need not check the second order conditions for the problem. The positive constant  $k \equiv \alpha_1^{-\alpha_1} \alpha_2^{-\alpha_2} \alpha_3^{-\alpha_3}$  will appear in the cost function.

(ii) Calculate the input one demand elasticity with respect to output  $[\partial x_1(y, p_1, p_2, p_3) / \partial y][y / x_1(y, p_1, p_2, p_3)]$  and the three input one demand elasticities with respect to input prices  $[\partial x_1(y, p_1, p_2, p_3) / \partial p_n][p_n / x_1(y, p_1, p_2, p_3)]$  for  $n = 1, 2, 3$ .

(iii) Show that  $-1 < [\partial x_1(y, p_1, p_2, p_3) / \partial p_1][p_1 / x_1(y, p_1, p_2, p_3)] < 0$ .

(iv) Show that  $0 < [\partial x_1(y, p_1, p_2, p_3) / \partial p_2][p_2 / x_1(y, p_1, p_2, p_3)] < 1$ .

(v) Show that  $0 < [\partial x_1(y, p_1, p_2, p_3) / \partial p_3][p_3 / x_1(y, p_1, p_2, p_3)] < 1$ .

(vi) Can any pair of inputs be complementary if the technology is a three input Cobb Douglas?

*Comment:* The Cobb Douglas functional form is widely used in macroeconomics and in applied general equilibrium models. However, this problem shows that it is not satisfactory if  $N \geq 3$ . Even in the  $N = 2$  case where analogues to (iii) and (iv) above hold, it can be seen that this functional form is not consistent with technologies where the degree of substitution between inputs is very high or very low.

**Problem 12** Suppose that the second order partial derivatives with respect to input prices of the cost function  $C(y, \mathbf{p})$  exist so that the  $n$ th cost minimizing input demand function  $x_n(y, \mathbf{p}) = \partial C(y, \mathbf{p}) / \partial p_n > 0$  exists for  $n = 1, \dots, N$ . Define the input  $n$  elasticity of demand with respect to input price  $k$  as follows:

$$e_{nk}(y, \mathbf{p}) \equiv [\partial x_n(y, \mathbf{p}) / \partial p_k][p_k / x_n(y, \mathbf{p})] \quad \text{for } n = 1, \dots, N \text{ and } k = 1, \dots, N. \quad (\text{a})$$

\*<sup>24</sup> Diewert (1974; 115 and 133)[75] introduced the term "flexible functional form" to describe a functional form for a cost function (or production function) that could approximate an arbitrary cost function (consistent with microeconomic theory) to the second order around any given point. The Generalized Leontief cost function defined by (3.27) above is flexible for the class of cost functions that are dual to linearly homogeneous production functions if we do not impose any restrictions on the parameters  $b_{ij}$ ; see Diewert (1971)[73] and section 3.9 below for a proof of this fact. However, if we do not impose the nonnegativity restrictions  $b_{ij} \geq 0$  for  $i \neq j$  on this functional form, it will frequently turn out that when these parameters are econometrically estimated, the resulting cost function fails the concavity restrictions,  $\nabla_{pp}^2 C(y^t, \mathbf{p}^t)$  is negative semidefinite, at one or more points  $(y^t, \mathbf{p}^t)$  in the observed data set that was used in the econometric estimation. Thus finding flexible functional forms where the restrictions implied by microeconomic theory can be *imposed* on the functional form without destroying its flexibility is a nontrivial task.

Show that for each  $n$ ,  $\sum_{k=1}^N e_{nk}(y, \mathbf{p}) = 0$ .

**Problem 13** Let the producer's cost function be  $C(y, \mathbf{p})$ , which satisfies the regularity conditions in Theorem 1 and, in addition, is once differentiable with respect to the components of the input price vector  $\mathbf{p}$ . Then the  $n$ th input demand function is  $x_n(y, \mathbf{p}) \equiv \partial C(y, \mathbf{p}) / \partial p_n$  for  $n = 1, \dots, N$ . Input  $n$  is defined to be *normal* at the point  $(y, \mathbf{p})$  if  $\partial x_n(y, \mathbf{p}) / \partial y = \partial^2 C(y, \mathbf{p}) / \partial p_n \partial y > 0$ ; i.e., if the cost minimizing demand for input  $n$  increases as the target output level  $y$  increases. On the other hand, input  $n$  is defined to be *inferior* at the point  $(y, \mathbf{p})$  if  $\partial x_n(y, \mathbf{p}) / \partial y = \partial^2 C(y, \mathbf{p}) / \partial p_n \partial y < 0$ . Prove that not all  $N$  inputs can be inferior at the point  $(y, \mathbf{p})$ .

*Hint:* Make use of (3.48).

**Problem 14** If the production function  $f$  dual to the differentiable cost function  $C(y, \mathbf{p})$  exhibits *constant returns to scale* so that  $f(\lambda \mathbf{x}) = \lambda f(\mathbf{x})$  for all  $\mathbf{x} \geq \mathbf{0}_N$  and all  $\lambda > 0$ , then show that for each  $n$ , the input  $n$  elasticity of demand with respect to the output level  $y$  is 1; i.e., show that for  $n = 1, \dots, N$ ,  $[\partial x_n(y, \mathbf{p}) / \partial y][y / x_n(y, \mathbf{p})] = 1$ .

**Problem 15** Let  $C(y, \mathbf{p})$  be a twice continuously differentiable cost function that satisfies the regularity conditions listed in Theorem 1 in section 3.2 above. By Shephard's Lemma, the input demand functions are given by

$$x_n(y, \mathbf{p}) = \partial C(y, \mathbf{p}) / \partial p_n > 0; \quad n = 1, \dots, N. \quad (\text{i})$$

The Allen (1938; 504)[4] Uzawa (1962)[378] *elasticity of substitution*  $\sigma_{nk}$  between inputs  $n$  and  $k$  is defined as follows:

$$\begin{aligned} \sigma_{nk}(y, \mathbf{p}) &\equiv \{C(y, \mathbf{p}) \partial^2 C(y, \mathbf{p}) / \partial p_n \partial p_k\} / \{[\partial C(y, \mathbf{p}) / \partial p_n][\partial C(y, \mathbf{p}) / \partial p_k]\} \quad 1 \leq n, k \leq N \\ &= \{C(y, \mathbf{p}) \partial^2 C(y, \mathbf{p}) / \partial p_n \partial p_k\} / x_n(y, \mathbf{p}) x_k(y, \mathbf{p}) \quad \text{using (i)}. \end{aligned} \quad (\text{ii})$$

Define  $\Sigma \equiv [\sigma_{nk}(y, \mathbf{p})]$  as the  $N \times N$  matrix of elasticities of substitution.

(a) Show that  $\Sigma$  has the following properties:

$$\cdot \Sigma = \Sigma^T; \quad (\text{iii})$$

$$\cdot \Sigma \text{ is negative semidefinite and} \quad (\text{iv})$$

$$\cdot \Sigma_{\mathbf{s}} = \mathbf{0}_N \quad (\text{v})$$

where  $\mathbf{s} \equiv [s_1, \dots, s_N]^T$  is the vector of cost shares; i.e.,  $s_n \equiv p_n x_n(y, \mathbf{p}) / C(y, \mathbf{p})$  for  $n = 1, \dots, N$ . Now define the  $N \times N$  matrix of cross price elasticities of demand  $\mathbf{E}$  in a manner analogous to definition (ii) above:

$$\begin{aligned} \mathbf{E} &\equiv [e^{nk}] \quad n = 1, \dots, N; \quad k = 1, \dots, N \\ &\equiv [(p_k / x_n) \partial x_n(y, \mathbf{p}) / \partial p_k] \\ &= [(p_k / x_n) \partial^2 C(y, \mathbf{p}) / \partial p_n \partial p_k] \quad \text{using (i)} \\ &= \hat{x}^{-1} \nabla_{pp}^2 C(y, \mathbf{p}) \hat{p}. \end{aligned} \quad (\text{vi})$$

(b) Show that  $\mathbf{E} = \Sigma_{\hat{\mathbf{s}}}$  where  $\hat{\mathbf{s}}$  is an  $N \times N$  diagonal matrix with the elements of the share vector  $\mathbf{s}$  running down the main diagonal.

**Problem 16** Suppose a firm's cost function has the following Constant Elasticity of Substitution (CES) functional form:<sup>\*25</sup>

$$C(y, p_1, \dots, p_N) \equiv ky \left[ \sum_{n=1}^N \alpha_n p_n^r \right]^{1/r}; \quad k > 0; r \leq 1, r \neq 0; \alpha_n > 0 \text{ and } \sum_{n=1}^N \alpha_n = 1. \quad (\text{i})$$

Thus the cost function is equal to a positive constant  $k$  times the output level  $y$  times a mean of order  $r$ . From the chapter on inequalities, we know that  $C(y, \mathbf{p})$  is a concave function of  $\mathbf{p}$  provided that  $r$  is equal to or less than one. Show that

$$\sigma_{nk}(y, \mathbf{p}) = -(r - 1) \quad \text{for all } n, k \text{ such that } n \neq k \quad (\text{ii})$$

where  $\sigma_{nk}(y, \mathbf{p})$  is the elasticity of substitution between inputs  $n$  and  $k$  defined above in problem 15, part (ii).

*Comment:* This problem shows why the CES functional form is unsatisfactory if the number of inputs  $N$  exceeds two, since it is a priori unlikely that all elasticities of substitution between every pair of inputs would equal the same number.

### 3.6 The Application of Cost Functions to Consumer Theory

The cost function and production function framework described in the previous sections can be readily adapted to the consumer context: simply replace output  $y$  by utility  $u$ , reinterpret the production function  $F$  as a utility function, reinterpret the input vector  $\mathbf{x}$  as a vector of commodity demands and reinterpret the vector of input prices  $\mathbf{p}$  as a vector of commodity prices. With these changes, the producer's cost minimization problem (3.1) becomes the following problem of *minimizing the cost or expenditure of attaining a given level of utility*  $u$ :

$$C(u, \mathbf{p}) \equiv \min_{\mathbf{x}} \{ \mathbf{p}^T \mathbf{x} : F(\mathbf{x}) \geq u \}. \quad (3.49)$$

If the cost function is differentiable with respect to the components of the commodity price vector  $\mathbf{p}$ , then Shephard's (1953; 11)[355] Lemma applies and the consumer's system of commodity demand functions as functions of the chosen utility level  $u$  and the commodity price vector  $\mathbf{p}$ ,  $\mathbf{x}(u, \mathbf{p})$ , is equal to the vector of first order partial derivatives of the cost or expenditure function  $C(u, \mathbf{p})$  with respect to the components of  $\mathbf{p}$ :

$$\mathbf{x}(u, \mathbf{p}) = \nabla_{\mathbf{p}} C(u, \mathbf{p}). \quad (3.50)$$

The demand functions  $x_n(u, \mathbf{p})$  defined in (3.50) are known as *Hicksian*<sup>\*26</sup> *demand functions*.

Thus it seems that we can adapt the theory of cost and production functions used in section 3.2 above in a very straightforward way, replacing output  $y$  by utility  $u$  and reinterpreting all of our previous results. However, there is a problem: the output level  $y$  is an *observable* variable but the corresponding utility level  $u$  is *not observable*!

However, this problem can be solved (but as we will see, some of the details are rather complex). We need only equate the cost function  $C(u, \mathbf{p})$  to the consumer's *observable expenditure* in the period under consideration,  $Y$  say, and solve the resulting equation for  $u$  as a function of  $Y$  and  $\mathbf{p}$ , say  $u = g(Y, \mathbf{p})$ . Thus  $u = g(Y, \mathbf{p})$  is the  $u$  solution to the following equation:

$$C(u, \mathbf{p}) = Y \quad (3.51)$$

<sup>\*25</sup> This functional form was introduced into the production literature by Arrow, Chenery, Minhas and Solow (1961)[13].

<sup>\*26</sup> See Hicks (1946; 311-331)[222].

and the resulting solution function  $u = g(Y, \mathbf{p})$  is the *consumer's indirect utility function*. Now replace the  $u$  in the system of Hicksian demand functions (3.50) by  $g(Y, \mathbf{p})$  and we obtain the consumer's system of (observable) *market demand functions*:

$$\mathbf{d}(Y, \mathbf{p}) = \nabla_{\mathbf{p}} C(g(Y, \mathbf{p}), \mathbf{p}). \quad (3.52)$$

We illustrate the above procedure for the generalized Leontief cost function defined by (3.27) above. For this functional form, equation (3.51) becomes:

$$u \sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{1/2} p_j^{1/2} = Y; \quad (b_{ij} = b_{ji} \text{ for all } i \text{ and } j) \quad (3.53)$$

and the  $u$  solution to this equation is:

$$u = g(Y, \mathbf{p}) = Y / \left[ \sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{1/2} p_j^{1/2} \right]. \quad (3.54)$$

The Hicksian demand functions for the  $C(u, \mathbf{p})$  defined by the left hand side of (3.53) are:

$$x_n(u, \mathbf{p}) \equiv \partial C(u, \mathbf{p}) / \partial p_n = \left[ \sum_{j=1}^N b_{nj} (p_j / p_n)^{1/2} \right] u; \quad n = 1, \dots, N. \quad (3.55)$$

Substituting (3.54) into (3.55) leads to the following system of market demand functions:

$$d_n(Y, \mathbf{p}) = \left[ \sum_{j=1}^N b_{nj} (p_j / p_i)^{1/2} \right] Y / \left[ \sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{1/2} p_j^{1/2} \right]; \quad n = 1, \dots, N. \quad (3.56)$$

Equations (3.56) can be used as the basis for the econometric estimation of preferences. Suppose that we have collected data on the quantities  $x_n^t$  purchased over  $T$  time periods for a household as well as the corresponding commodity prices  $p_n^t$ . Then we can define period  $t$  "income"\*<sup>27</sup> or expenditure on the  $n$  commodities as  $Y^t$ :

$$Y^t \equiv \sum_{n=1}^N p_n^t x_n^t; \quad t = 1, \dots, T. \quad (3.57)$$

Evaluating (3.56) at the period  $t$  data and adding a stochastic error term  $e_n^t$  to equation  $n$  in (3.56) for  $n = 1, \dots, N$  leads to the following system of estimating equations:

$$x_n^t = \left[ \sum_{j=1}^N b_{nj} (p_j^t / p_n^t)^{1/2} \right] Y^t / \left[ \sum_{i=1}^N \sum_{j=1}^N b_{ij} (p_i^t)^{1/2} (p_j^t)^{1/2} \right] + e_n^t; \quad t = 1, \dots, T; n = 1, \dots, N. \quad (3.58)$$

Not all  $N$  equations in (3.58) can have independent error terms since if we multiply both sides of equation  $n$  in (3.58) by  $p_n^t$  and sum over  $n$ , we obtain the following equation:

$$\sum_{n=1}^N p_n^t x_n^t = Y^t + \sum_{n=1}^N p_n^t e_n^t. \quad (3.59)$$

Using (3.57), we find that the period  $t$  errors  $e_n^t$  satisfy the following linear restriction exactly:

$$\sum_{n=1}^N p_n^t e_n^t = 0; \quad t = 1, \dots, T. \quad (3.60)$$

There is one other factor that must be taken into account in doing an econometric estimation of preferences using the system of estimation equations (3.58). Note that if we multiply all of the  $b_{ij}$  parameters by the positive number  $\lambda$ , the right hand sides of each equation in (3.58) will remain unchanged; i.e., the demand functions are homogeneous of degree 0 in the  $b_{ij}$  parameters. Thus these parameters *will not be identified* as matters stand. Hence, it will be necessary to impose a

\*<sup>27</sup> Strictly speaking, a household's income will also include savings in addition to expenditures on current goods and services.

normalization on these parameters. One normalization that is frequently used in applied economics is to set unit cost<sup>\*28</sup> equal to 1 for some set of reference prices  $\mathbf{p}^0$  say.<sup>\*29</sup> Thus we impose the following normalization on the  $b_{ij}$ :

$$1 = c(\mathbf{p}^0) = \sum_{i=1}^N \sum_{j=1}^N b_{ij} (p_i^0)^{1/2} (p_j^0)^{1/2}. \quad (3.61)$$

Equation (3.61) can be used to solve for say  $b_{11}$  in terms of the other  $b_{ij}$  and then this equation can be used to eliminate  $b_{11}$  from the  $N - 1$  independent estimating equations in (3.58) and the remaining parameters can be estimated using nonlinear regression techniques.

The technique suggested above for the econometric estimation of preferences is a special case of the following general strategy: (i) Assume that the consumer's preferences can be represented by the cost function  $C(u, \mathbf{p})$  that has the following form:

$$C(u, \mathbf{p}) = uc(\mathbf{p}) \quad (3.62)$$

where  $c(\mathbf{p})$  is a suitable differentiable unit cost function. (ii) Differentiate (3.62) with respect to the components of the commodity price vector  $\mathbf{p}$  to form the following system of Hicksian demand functions:

$$\mathbf{x}(u, \mathbf{p}) = u \nabla_{\mathbf{p}} c(\mathbf{p}). \quad (3.63)$$

(iii) Equate cost  $uc(\mathbf{p})$  to expenditure or income  $Y$  and solve for  $u$  as a function of  $Y$  and  $\mathbf{p}$  to get the consumer's indirect utility function  $u = g(Y, \mathbf{p})$ :

$$u = Y/c(\mathbf{p}). \quad (3.64)$$

(iv) Substitute (3.64) into the right hand side of (3.63) in order to obtain the following system of consumer demand functions:

$$\mathbf{d}(Y, \mathbf{p}) = \nabla_{\mathbf{p}} c(\mathbf{p}) Y/c(\mathbf{p}). \quad (3.65)$$

(v) Finally, impose the normalization (3.61),  $c(\mathbf{p}^0) = 1$ , in order to identify all of the unknown parameters in (3.65).

Unfortunately, there is a problem with the above strategy for estimating a consumer's preferences. The problem is the same one that occurred in problem 14 above; with the  $n$ th consumer demand function,  $d_n(Y, \mathbf{p})$  defined by the  $n$ th equation in (3.52), we find that *all income elasticities of demand are equal to one*; i.e., we have:

$$[Y/d_n(Y, \mathbf{p})] \partial d_n(Y, \mathbf{p}) / \partial Y = 1; \quad n = 1, \dots, N. \quad (3.66)$$

But (3.66) contradicts *Engel's Law*, which says that the income elasticity of demand for food is less than one.

In the following two sections, we show how this problem of unitary income elasticities can be solved.

## 3.7 Flexible Functional Forms and Nonunitary Income Elasticities of Demand

We first define what it means for a unit cost function,  $c(\mathbf{p})$ , to be a *flexible functional form*. Let  $c^*(\mathbf{p})$  be an arbitrary unit cost function that satisfies the appropriate regularity conditions on unit cost functions and in addition, is twice continuously differentiable around a point  $\mathbf{p}^* \gg \mathbf{0}_N$ . Then

<sup>\*28</sup> The Generalized Leontief cost function defined by (3.27) has the form  $C(u, \mathbf{p}) = uc(\mathbf{p})$  where  $c(\mathbf{p}) \equiv C(1, \mathbf{p})$ .

<sup>\*29</sup> Usually,  $\mathbf{p}^0$  is taken to be  $\mathbf{p}^1$ , the vector of prices that prevailed in the base period.

we say that a given unit cost function  $c(\mathbf{p})$  that is also twice continuously differentiable around the point  $\mathbf{p}^*$  is *flexible* if it has enough free parameters so that the following  $1 + N + N^2$  equations can be satisfied:

$$c(\mathbf{p}^*) = c^*(\mathbf{p}^*); \quad (3.67)$$

$$\nabla c(\mathbf{p}^*) = \nabla c^*(\mathbf{p}^*); \quad (3.68)$$

$$\nabla^2 c(\mathbf{p}^*) = \nabla^2 c^*(\mathbf{p}^*). \quad (3.69)$$

Thus  $c(\mathbf{p})$  is a flexible functional form if it has enough free parameters to provide a second order Taylor series approximation to an arbitrary unit cost function.

At first glance, it looks like  $c(\mathbf{p})$  will have to have at least  $1 + N + N^2$  independent parameters in order to be able to satisfy all of the equations (3.67)-(3.69). However, since both  $c$  and  $c^*$  are assumed to be twice continuously differentiable, Young's Theorem in calculus implies that  $\partial^2 c(\mathbf{p}^*)/\partial p_i \partial p_j = \partial^2 c(\mathbf{p}^*)/\partial p_j \partial p_i$  for all  $i \neq j$  (and of course, the same equations hold for the second order partial derivatives of  $c^*(\mathbf{p})$  when evaluated at  $\mathbf{p} = \mathbf{p}^*$ ). Thus the  $N^2$  equations in (3.69) can be replaced with the following  $N(N + 1)/2$  equations:

$$\frac{\partial^2 c(\mathbf{p}^*)}{\partial p_i \partial p_j} = \frac{\partial^2 c^*(\mathbf{p}^*)}{\partial p_j \partial p_i} \quad \text{for } 1 \leq i \leq j \leq N. \quad (3.70)$$

Another property that both unit cost functions must have is homogeneity of degree one in the components of  $\mathbf{p}$ . By part 1 of Euler's Theorem on homogeneous functions,  $c$  and  $c^*$  satisfy the following equations:

$$c(\mathbf{p}^*) = \mathbf{p}^{*T} \nabla c(\mathbf{p}^*) \text{ and } c^*(\mathbf{p}^*) = \mathbf{p}^{*T} \nabla c^*(\mathbf{p}^*). \quad (3.71)$$

Thus if  $c$  and  $c^*$  satisfy equations (3.68), then using (3.71), we see that  $c$  and  $c^*$  automatically satisfy equation (3.67). By part 2 of Euler's Theorem on homogeneous functions,  $c$  and  $c^*$  satisfy the following equations:

$$\nabla^2 c(\mathbf{p}^*) \mathbf{p}^* = \mathbf{0}_N \text{ and } \nabla^2 c^*(\mathbf{p}^*) \mathbf{p}^* = \mathbf{0}_N. \quad (3.72)$$

This means that if we have  $\partial^2 c(\mathbf{p}^*)/\partial p_i \partial p_j = \partial^2 c^*(\mathbf{p}^*)/\partial p_i \partial p_j$  for all  $i \neq j$ , then equations (3.72) will imply that  $\partial^2 c(\mathbf{p}^*)/\partial p_j \partial p_j = \partial^2 c^*(\mathbf{p}^*)/\partial p_j \partial p_j$  as well, for  $j = 1, \dots, N$ .

Summarizing the above material, if  $c(\mathbf{p})$  is linearly homogeneous, then in order for it to be flexible,  $c(\mathbf{p})$  needs to have only enough parameters so that the  $N$  equations in (3.68) can be satisfied and so that the following  $N(N - 1)/2$  equations can be satisfied:

$$\frac{\partial^2 c(\mathbf{p}^*)}{\partial p_j \partial p_j} = \frac{\partial^2 c^*(\mathbf{p}^*)}{\partial p_i \partial p_j} \equiv c_{ij}^* \quad \text{for } 1 \leq i < j \leq N. \quad (3.73)$$

Thus in order to be flexible,  $c(\mathbf{p})$  must have at least  $N + N(N - 1)/2 = N(N + 1)/2$  independent parameters.

Now consider the Generalized Leontief unit cost function defined as follows:<sup>\*30</sup>

$$c(\mathbf{p}) \equiv \sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{1/2} p_j^{1/2}; \quad b_{ij} = b_{ji} \text{ for all } i \text{ and } j. \quad (3.74)$$

Note that there are exactly  $N(N + 1)/2$  independent  $b_{ij}$  parameters in the  $c(\mathbf{p})$  defined by (3.74). For this functional form, the  $N$  equations in (3.68) become:

$$\partial c(\mathbf{p}^*)/\partial p_n = \sum_{j=1}^N b_{nj} (p_j^*/p_n^*)^{1/2} = \partial c^*(\mathbf{p}^*)/\partial p_n \equiv c_n^*; \quad n = 1, \dots, N. \quad (3.75)$$

<sup>\*30</sup> We no longer restrict the  $b_{ij}$  to be nonnegative.

The  $N(N - 1)/2$  equations in (3.73) become:

$$(1/2)b_{ij}/(p_i^*p_j^*)^{1/2} = c_{ij}^*; \quad 1 \leq i < j \leq N. \quad (3.76)$$

However, it is easy to solve equations (3.76) for the  $b_{ij}$ :

$$b_{ij} = 2c_{ij}^*(p_i^*p_j^*)^{1/2}; \quad 1 \leq i < j \leq N. \quad (3.77)$$

Once the  $b_{ij}$  for  $i < j$  have been determined using (3.77), we set  $b_{ji} = b_{ij}$  for  $i < j$  and finally the  $b_{ii}$  are determined using the  $N$  equations in (3.75).

The above material shows how we can find a flexible functional form for a unit cost function<sup>\*31</sup>. We now turn our attention to finding a flexible functional form for a general cost function  $C(u, \mathbf{p})$ . Let  $C^*(u, \mathbf{p})$  be an arbitrary cost function that satisfies the appropriate regularity conditions on cost functions listed in Theorem 1 above and in addition, is twice continuously differentiable around a point  $(u^*, \mathbf{p}^*)$  where  $u^* > 0$  and  $\mathbf{p}^* \gg \mathbf{0}_N$ . Then we say that a given cost function  $C(u, \mathbf{p})$  that is also twice continuously differentiable around the point  $(u^*, \mathbf{p}^*)$  is *flexible* if it has enough free parameters so that the following  $1 + (N + 1) + (N + 1)^2$  equations can be satisfied:

$$C(u^*, \mathbf{p}^*) = C^*(u^*, \mathbf{p}^*); \quad (1 \text{ equation}) \quad (3.78)$$

$$\nabla_{\mathbf{p}} C(u^*, \mathbf{p}^*) = \nabla_{\mathbf{p}} C^*(u^*, \mathbf{p}^*); \quad (N \text{ equations}) \quad (3.79)$$

$$\nabla_{\mathbf{p}\mathbf{p}}^2 C(u^*, \mathbf{p}^*) = \nabla_{\mathbf{p}\mathbf{p}}^2 C^*(u^*, \mathbf{p}^*); \quad (N^2 \text{ equations}) \quad (3.80)$$

$$\nabla_u C(u^*, \mathbf{p}^*) = \nabla_u C^*(u^*, \mathbf{p}^*); \quad (1 \text{ equation}) \quad (3.81)$$

$$\nabla_{\mathbf{p}u}^2 C(u^*, \mathbf{p}^*) = \nabla_{\mathbf{p}u}^2 C^*(u^*, \mathbf{p}^*); \quad (N \text{ equations}) \quad (3.82)$$

$$\nabla_{\mathbf{u}\mathbf{p}}^2 C(u^*, \mathbf{p}^*) = \nabla_{\mathbf{u}\mathbf{p}}^2 C^*(u^*, \mathbf{p}^*); \quad (N \text{ equations}) \quad (3.83)$$

$$\nabla_{uu}^2 C(u^*, \mathbf{p}^*) = \nabla_{uu}^2 C^*(u^*, \mathbf{p}^*) \quad (1 \text{ equation}). \quad (3.84)$$

Equations (3.78)-(3.80) are the counterparts to our earlier unit cost equations (3.67)-(3.69). As was the case with unit cost functions, equation (3.78) is implied by the linear homogeneity in prices of the cost functions and Part 1 of Euler's Theorem on homogeneous functions. Young's Theorem on the symmetry of cross partial derivatives means that the lower triangle of equations in (3.80) is implied by the equalities in the upper triangle of both matrices of partial derivatives. Part 2 of Euler's Theorem on homogeneous functions implies that if all the off diagonal elements in both matrices in (3.80) are equal, then so are the diagonal elements. Hence, in order to satisfy all of the equations in (3.78)-(3.80), we need only satisfy the  $N$  equations in (3.79) and the  $N(N - 1)/2$  in the upper triangle of equations (3.80). Young's Theorem implies that if equations (3.82) are satisfied, then so are equations (3.83). However, Euler's Theorem on homogeneous functions implies that

$$\frac{\partial C(u^*, \mathbf{p}^*)}{\partial u} = \mathbf{p}^{*T} \nabla_{\mathbf{p}u}^2 C(u^*, \mathbf{p}^*) = \mathbf{p}^{*T} \nabla_{\mathbf{p}u}^2 C^*(u^*, \mathbf{p}^*) = \frac{\partial C^*(u^*, \mathbf{p}^*)}{\partial u}. \quad (3.85)$$

Hence, if equations (3.82) are satisfied, then so is the single equation (3.81). Putting this all together, we see that in order for  $C$  to be flexible, we need enough free parameters in  $C$  so that the following equations can be satisfied:

- Equations (3.79);  $N$  equations;
- The upper triangle in equations (3.80);  $N(N - 1)/2$  equations;
- Equations (3.82);  $N$  equations; and
- Equation (3.84); 1 equation.

<sup>\*31</sup> This material can be adapted to the case where we want a flexible functional form for a linearly homogeneous utility or production function  $f(\mathbf{x})$ : just replace  $\mathbf{p}$  by  $\mathbf{x}$  and  $c(\mathbf{p})$  by  $f(\mathbf{x})$ .

Hence, in order for  $C$  to be a flexible functional form, it will require a minimum of  $2N + N(N - 1)/2 + 1 = N(N + 1)/2 + N + 1$  parameters. Thus a fully flexible cost function,  $C(u, \mathbf{p})$ , will require  $N + 1$  additional parameters compared to a flexible unit cost function,  $c(\mathbf{p})$ .

Suppose the unit cost function is the Generalized Leontief unit cost function  $c(\mathbf{p})$  defined by (3.74) above. We now show how terms can be added to it in order to make it a fully flexible cost function. Thus define  $C(u, \mathbf{p})$  as follows:

$$C(u, \mathbf{p}) \equiv uc(\mathbf{p}) + \mathbf{b}^T \mathbf{p} + (1/2)a_0 \boldsymbol{\alpha}^T \mathbf{p} u^2 \quad (3.86)$$

where  $\mathbf{b}^T \equiv [b_1, \dots, b_N]$  is an  $N$  dimensional vector of new parameters,  $a_0$  is a new parameter and  $\boldsymbol{\alpha}^T \equiv [\alpha_1, \dots, \alpha_N] > \mathbf{0}_N$  is a vector of predetermined parameters.\*<sup>32</sup> Using (3.86) as our  $C$ , equations (3.79), (3.80), (3.82) and (3.84) become:

$$u^* \nabla_{\mathbf{p}} c(\mathbf{p}^*) + \mathbf{b} + (1/2)a_0 \boldsymbol{\alpha} u^{*2} = \nabla_{\mathbf{p}} C^*(u^*, \mathbf{p}^*); \quad (3.87)$$

$$u^* \nabla_{\mathbf{p}\mathbf{p}}^2 c(\mathbf{p}^*) = \nabla_{\mathbf{p}\mathbf{p}}^2 C^*(u^*, \mathbf{p}^*); \quad (3.88)$$

$$\nabla_{\mathbf{p}} c(\mathbf{p}^*) + a_0 \boldsymbol{\alpha} u^* = \nabla_{\mathbf{p}u}^2 C^*(u^*, \mathbf{p}^*); \quad (3.89)$$

$$a_0 \boldsymbol{\alpha}^T \mathbf{p}^* = \nabla_{uu}^2 C^*(u^*, \mathbf{p}^*). \quad (3.90)$$

Use equations (3.88) in order to determine the  $b_{ij}$  for  $i \neq j$ . Use (3.90) in order to determine the single parameter  $a_0$ . Use equations (3.89) in order to determine the  $b_{ii}$ . Finally, use equations (3.87) in order to determine the parameters  $b_n$  in the  $\mathbf{b}$  vector. Thus the cost function  $C(u, \mathbf{p})$  defined by (3.86), which uses the Generalized Leontief unit cost function  $c(\mathbf{p})$  defined by (3.74) as a building block, is a *parsimonious flexible functional form* for a general cost function.

In fact, it is not necessary to use the Generalized Leontief unit cost function in (3.86) in order to convert a flexible functional form for a unit cost function into a flexible functional form for a general cost function. Let  $c(\mathbf{p})$  be any flexible functional form for a unit cost function and define  $C(u, \mathbf{p})$  by (3.86). Use equation (3.90) to determine the parameter  $a_0$ . Once  $a_0$  has been determined, equations (3.88) and (3.89) can be used to determine the parameters in the unit cost function  $c(\mathbf{p})$ .<sup>\*33</sup> Finally, equations (3.87) can be used to determine the parameters in the vector  $\mathbf{b}$ .

Obviously, the material in this section can be applied to the problems involved in estimating a flexible cost function in the production context: simply replace utility  $u$  by output  $y$  and reinterpret the commodity price vector  $\mathbf{p}$  as an input price vector. Differentiating (3.86) leads to the following system of estimating equations, where  $\mathbf{x}(y, \mathbf{p})$  equal to  $\nabla_{\mathbf{p}} C(y, \mathbf{p})$  is the producer's system of cost minimizing input demand functions:

$$\mathbf{x}(y, \mathbf{p}) = y \nabla c(\mathbf{p}) + \mathbf{b} + (1/2)a_0 \boldsymbol{\alpha} y^2. \quad (3.91)$$

In order to obtain estimating equations for the general cost function defined by (3.86), there are some normalization issues that need to be discussed. We do this in the following section.

## 3.8 Money Metric Utility Scaling and Other Methods of Cardinalizing Utility

Since utility is unobservable, in order to estimate econometrically a consumer's utility function, it will be necessary to pick a utility scale for that consumer; i.e., it will be necessary to *cardinalize* the

\*<sup>32</sup> The parameter  $a_0$  could be set equal to 1 and the vector of parameters  $\boldsymbol{\alpha}$  could be estimated econometrically. We have defined the cost function  $C$  in this manner so that it has the *minimal* number of parameters required in order to be a flexible functional form. Thus it is a *parsimonious* flexible functional form.

\*<sup>33</sup> It can be seen that equations (3.88) and (3.89) have the same structure as equations (3.68) and (3.69). Hence if  $c(\mathbf{p})$  has enough free parameters to satisfy (3.68) and (3.69), then it has enough free parameters to satisfy (3.88) and (3.89) once  $a_0$  has been determined.

consumer’s utility function.\*34

There are two commonly used methods that have been used to pick a cardinal utility scale for a consumer. The first method is used when we are working with the consumer’s direct utility function,  $F(\mathbf{x})$  say. We simply pick a strictly positive *reference consumption vector*,  $\mathbf{x}^* \gg \mathbf{0}_N$  say, set the utility of this vector equal to some positive number  $F(\mathbf{x}^*)$  and scale the level of utility along the ray through the point  $\mathbf{x}$  as follows:\*35

$$F(\lambda\mathbf{x}^*) = \lambda F(\mathbf{x}^*); \quad \lambda \geq 0. \tag{3.92}$$

Thus all consumption vectors  $\mathbf{x} \geq \mathbf{0}_N$  such that they yield the same utility as  $\mathbf{x}^*$  are assigned the utility level  $F(\mathbf{x}^*)$ ; this is the indifference curve or surface  $\{\mathbf{x} : F(\mathbf{x}) = F(\mathbf{x}^*)\}$ . Then all consumption vectors  $\mathbf{x}$  that are on the same indifference surface as  $2\mathbf{x}^*$  are given the utility level  $2F(\mathbf{x}^*)$ ; this is the indifference surface  $\{\mathbf{x} : F(\mathbf{x}) = F(2\mathbf{x}^*) = 2F(\mathbf{x}^*)\}$ , and so on. Thus the ray through the origin and the reference consumption vector  $\mathbf{x}^*$  is used to scale utility levels. Figure 3.3 illustrates how this cardinalization method works.

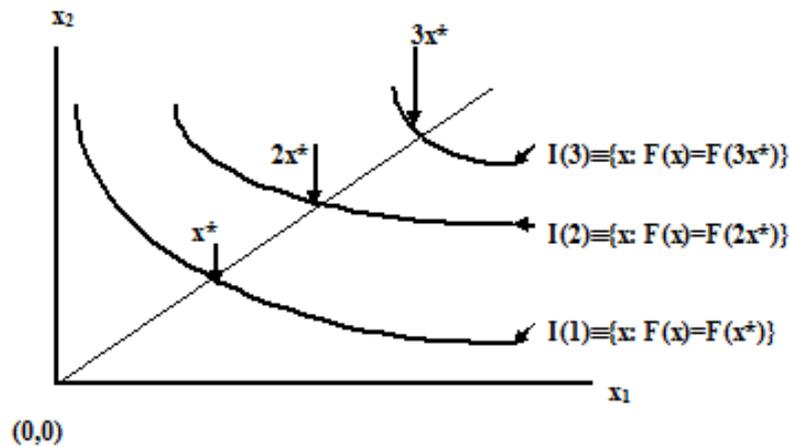


Fig. 3.3 Scaling Utility by a Reference Ray through the Origin

Of course, different choices of the reference consumption vector  $\mathbf{x}^*$  will lead to different cardinalizations of the consumer’s utility function. Usually,  $\mathbf{x}^*$  will be chosen to be the observed consumption vector of a consumer in some base period or situation.

We turn now to a second method of utility scaling that was referred to as *money metric utility scaling* by Samuelson (1974; 1262)[347]. For this method of utility scaling, we choose a reference set of prices, say  $\mathbf{p}^* \gg \mathbf{0}_N$ , and if these reference prices face the consumer, we normalize the consumer’s cost function,  $C(u, \mathbf{p})$ , so that the following restriction holds:

$$C(u, \mathbf{p}^*) = u \quad \text{for all } u > 0. \tag{3.93}$$

Typically, we choose  $\mathbf{p}^*$  to be the prices facing the consumer in some base period situation when the consumer spends the “income”  $Y^*$  on the  $N$  commodities and has utility level  $u^*$  so that equating

\*34 If preferences can be represented by the utility function  $u = F(\mathbf{x})$ , then they can be equally well represented by the utility function  $g\{F(\mathbf{x})\}$  where  $g(u)$  is a monotonically increasing function of one variable.

\*35 This is the type of utility scaling recommended by Blackorby (1975)[36] and other welfare economists because this form of scaling does not depend on prices.

expenditure to income in this base period, we have

$$C(u^*, \mathbf{p}^*) = Y^*. \quad (3.94)$$

Combining (3.93) and (3.94), we see that for this base period situation, we have

$$u^* = Y^*; \quad (3.95)$$

i.e., utility equals expenditure in this base period. Thus the money metric utility scaling convention (3.93) has the effect of making nominal “income”  $Y$  equal to utility  $u$  provided that the consumer is facing the reference prices  $\mathbf{p}^*$ . The geometry of this scaling method is illustrated in Figure 3.4.

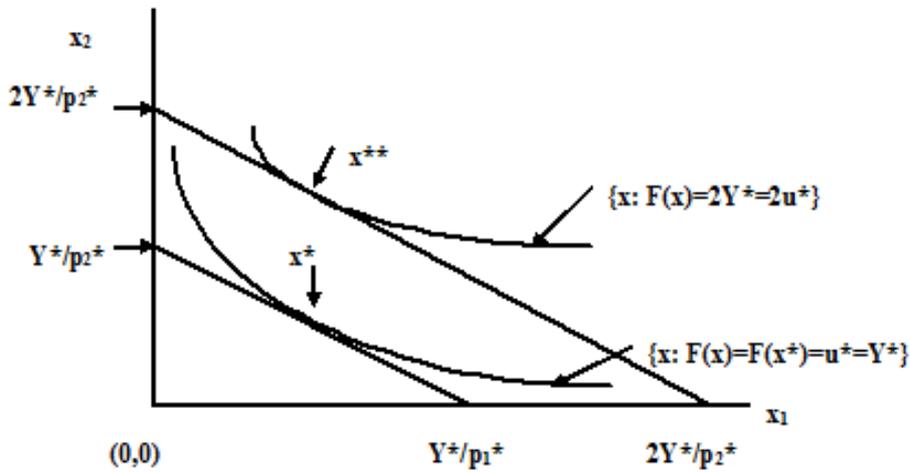


Fig. 3.4 Money Metric Utility Scaling

In Figure 3.4,  $\mathbf{x}^*$  solves the cost minimization problem,  $\min_{\mathbf{x}}\{\mathbf{p}^{*T}\mathbf{x} : F(\mathbf{x}) = u^*\} \equiv C(u^*, \mathbf{p}^*)$ , which in turn is equal to observed expenditure,  $Y^*$ . We scale utility so that  $u^*$  is set equal to  $Y^*$  in this base period situation. Thus all  $\mathbf{x}$  combinations that yield the utility level  $u^* = Y^*$  are assigned this utility level  $Y^*$ . This is the set  $\{\mathbf{x} : F(\mathbf{x}) = F(\mathbf{x}^*)\}$  that is labelled in Figure 3.4. Now double the initial utility level to  $2u^* = 2Y^*$  and solve the cost minimization problem  $\min_{\mathbf{x}}\{\mathbf{p}^{*T}\mathbf{x} : F(\mathbf{x}) = 2u^*\} \equiv C(2u^*, \mathbf{p}^*)$ , which in turn is equal to twice the initial expenditure,  $2Y^*$ . The solution to this cost minimization problem is  $\mathbf{x}^{**}$  in Figure 3.4. All points  $\mathbf{x}$  on the corresponding indifference curve,  $\{\mathbf{x} : F(\mathbf{x}) = F(\mathbf{x}^{**}) = 2u^*\}$ , are assigned the utility level  $2u^*$ , which in turn is equal to  $2Y^*$ .

In general, money metric utility scaling works as follows. For each positive “income” level  $Y > 0$ , define the budget set  $B(Y)$  as follows:

$$B(Y) \equiv \{\mathbf{x} : \mathbf{p}^{*T}\mathbf{x} = Y; \mathbf{x} \geq \mathbf{0}_N\}. \quad (3.96)$$

For each  $Y$  greater than zero, an indifference surface will be tangent to the budget set  $B(Y)$ .<sup>\*36</sup> All points on this indifference surface are assigned the utility level  $Y$ .

Money metric utility scaling suffers from the same disadvantage that ray scaling had; i.e., different choices of the reference vector of consumer prices  $\mathbf{p}^*$  will give rise to different utility scales. However, both money metric and ray scaling are acceptable methods of scaling utility; neither method of scaling can be contradicted by observable data on a consumer.

<sup>\*36</sup> If the cost function is differentiable, the tangent indifference surface is  $\{\mathbf{x} : F(\mathbf{x}) = F[\nabla_p C(Y, \mathbf{p}^*)]\}$ .

The money metric utility scaling assumption (3.93) implies additional restrictions on the derivatives of the cost function. Differentiating both sides of (3.93) with respect to  $u$  gives us the following equation:

$$\frac{\partial C(u, \mathbf{p}^*)}{\partial u} = 1 \quad \text{for all } u > 0. \quad (3.97)$$

Differentiating (3.97) with respect to  $u$  again leads to the following equation:

$$\frac{\partial^2 C(u, \mathbf{p}^*)}{\partial u^2} = 0 \quad \text{for all } u > 0. \quad (3.98)$$

Euler's Theorem on homogeneous functions and (3.97) imply the following additional restriction on the second order partial derivatives of the cost function:

$$\mathbf{p}^{*T} \nabla_{pu}^2 C(u, \mathbf{p}^*) = \frac{\partial C(u, \mathbf{p}^*)}{\partial u} = 1 \quad \text{for all } u > 0. \quad (3.99)$$

The restriction (3.99) is not an independent restriction since it is implied by Euler's Theorem on homogeneous functions and the restriction (3.97). Thus if we impose money metric utility scaling and want a flexible functional form, we will require  $1 + N + N(N + 1)/2$  independent parameters (see the previous section) less 3 parameters, which correspond to the restrictions (3.94), (3.97) and (3.98). We shall use money metric utility scaling quite frequently in this course.

The independent restrictions (3.94), (3.97) and (3.99) imposed by money metric utility scaling have an impact on our earlier discussion of flexible functional forms for the cost function,  $C(u, \mathbf{p})$ . Since empirically, it is harmless to impose money metric utility scaling, we can impose money metric scaling on both  $C(u, \mathbf{p})$  and  $C^*(u, \mathbf{p})$  at the point  $(u^*, \mathbf{p}^*)$  when we are attempting to find a flexible cost function  $C(u, \mathbf{p})$ . This means that equations (3.81) and (3.84) become:

$$\nabla_u C(u^*, \mathbf{p}^*) = \nabla_u C^*(u^*, \mathbf{p}^*) = 1; \quad (1 \text{ equation}) \quad (3.100)$$

$$\nabla_{uu}^2 C(u^*, \mathbf{p}^*) = \nabla_{uu}^2 C^*(u^*, \mathbf{p}^*) = 0; \quad (1 \text{ equation}). \quad (3.101)$$

Thus if money metric utility scaling using the reference prices  $\mathbf{p}^*$  is imposed on both  $C$  and  $C^*$ , equations (3.100) and (3.101) will be satisfied automatically. This reduces the number of free parameters that are required for  $C(u, \mathbf{p})$  to be a flexible functional form by 2 compared to our earlier discussion. The restriction (3.94) reduces the required number of parameters by an additional 1. Hence, in order for  $C$  to be a flexible functional form under the money metric utility scaling assumption,  $C$  will require a minimum of  $N(N + 1)/2 + N - 2$  parameters. Recall the Generalized Leontief cost function  $C(u, \mathbf{p})$  defined earlier by (3.86). Under our new money metric utility scaling assumptions, it is evident that we can set the parameter  $a_0$  equal to 0. Thus define  $C(u, \mathbf{p})$  as follows:

$$C(u, \mathbf{p}) \equiv uc(\mathbf{p}) + \mathbf{b}^T \mathbf{p} \quad (3.102)$$

where  $c(\mathbf{p})$  is defined by (3.74). The equations that we now have to satisfy in order for  $C$  defined by (3.102) to be flexible are as follows:

$$u^* \nabla_{\mathbf{p}} c(\mathbf{p}^*) + \mathbf{b} = \nabla_{\mathbf{p}} C^*(u^*, \mathbf{p}^*); \quad (3.103)$$

$$u^* \nabla_{pp}^2 c(\mathbf{p}^*) = \nabla_{pp}^2 C^*(u^*, \mathbf{p}^*); \quad (3.104)$$

$$\nabla_{\mathbf{p}} c(\mathbf{p}^*) = \nabla_{pu}^2 C^*(u^*, \mathbf{p}^*); \quad (3.105)$$

Use equations (3.104) in order to determine the  $b_{ij}$  for  $i \neq j$ . Use equations (3.105) in order to determine the  $b_{ii}$ . Finally, use equations (3.103) in order to determine the parameters  $b_n$  in the  $\mathbf{b}$  vector. However, using equation (3.100) means that  $c(\mathbf{p}^*)$  must satisfy the following restriction:

$$\nabla_u C(u^*, \mathbf{p}^*) = c(\mathbf{p}^*) = \nabla_u C^*(u^*, \mathbf{p}^*) = 1; \quad (3.106)$$

Thus  $c(\mathbf{p}^*) = 1$ , which means that we need to impose a restriction on the  $b_{ij}$  such as:

$$b_{11} = \left\{ 1 - \left[ \sum_{n=2}^N b_{nn} p_n^* + \sum_{i=1}^N \sum_{j=1, i \neq j}^N b_{ij} (p_i^* p_j^*)^{1/2} \right] \right\} / p_1^*. \quad (3.107)$$

Recall the normalization (3.61) in the previous section, which was similar to (3.107) except that the reference prices  $\mathbf{p}^0$  were used in place of the reference prices  $\mathbf{p}^*$ . Now premultiply both sides of (3.103) by  $\mathbf{p}^{*T}$  in order to obtain the following equation:

$$\begin{aligned} u^* c(\mathbf{p}^*) + \mathbf{p}^{*T} \mathbf{b} &= C^*(u^*, \mathbf{p}^*) && \text{using Euler's Theorem} \\ &= u^* && \text{using (3.93).} \end{aligned} \quad (3.108)$$

Using (3.106), namely that  $c(\mathbf{p}^*) = 1$ , we find that equation (3.108) becomes  $u^* + \mathbf{p}^{*T} \mathbf{b} = u^*$  or

$$\mathbf{p}^{*T} \mathbf{b} = 0. \quad (3.109)$$

Hence we can impose a restriction on the components of  $\mathbf{b}$  such as

$$b_1 = -\sum_{n=2}^N p_n^* b_n / p_1^* \quad (3.110)$$

without destroying the flexibility of the functional form defined by (3.102). Thus there are  $N(N+1)/2 - 1$  independent  $b_{ij}$  parameters in the Generalized Leontief unit cost function  $c(\mathbf{p})$  defined by (3.74) and  $N - 1$  independent  $b_n$  parameters in the  $\mathbf{b}$  vector, which is just the right number for (3.102) to be a parsimonious flexible functional form for a cost function in the context of money metric utility scaling.

In fact, it is not necessary to use the Generalized Leontief unit cost function in (3.102).<sup>\*37</sup> Let  $c(\mathbf{p})$  be any flexible functional form for a unit cost function and define  $C(u, \mathbf{p})$  by (3.102). Use equations (3.104) and (3.105) to determine the parameters in the unit cost function  $c(\mathbf{p})$ . Then use equations (3.103) to determine  $\mathbf{b}$ . Finally, repeat the arguments around equations (3.106), (3.108) and (3.109) to show that  $c(\mathbf{p})$  and  $\mathbf{b}$  satisfy the additional restrictions  $c(\mathbf{p}^*) = 1$  and  $\mathbf{p}^{*T} \mathbf{b} = 0$ .

The reader will note that our suggested flexible functional form for  $C(u, \mathbf{p})$  defined by (3.102) reduces to  $uc(\mathbf{p})$  if the parameters  $b_n$  in the  $\mathbf{b}$  vector all turn out to be zero. If  $c(\mathbf{p})$  is a flexible functional form for a unit cost function, then when  $\mathbf{b} = \mathbf{0}_N$ , our general flexible functional form  $C(u, \mathbf{p})$  can model homothetic (or linearly homogeneous) preferences in a flexible manner. Put another way, we found a flexible functional form for a general cost function,  $C(u, \mathbf{p})$ , by simply adding an extra parameter vector  $\mathbf{b}$  to a cost function that was flexible for homothetic preferences, namely  $uc(\mathbf{p})$ . The indirect utility function that corresponds to the cost function  $C(u, \mathbf{p})$  defined by (3.102) can be obtained by setting the right hand side of (3.102) to  $Y$  and then solving the resulting equation for  $u = g(Y, \mathbf{p})$ , which results in the following formula for  $g$ :

$$u = g(Y, \mathbf{p}) = [Y - \mathbf{b}^T \mathbf{p}] / c(\mathbf{p}). \quad (3.111)$$

In order for utility to be positive (and meaningful in this model), we require that the consumer's income  $Y$  be greater than or equal to *committed expenditures*,  $\mathbf{b}^T \mathbf{p}$ .<sup>\*38</sup> The system of Hicksian demand functions that corresponds to (3.102) is:

$$\mathbf{x}(u, \mathbf{p}) = \mathbf{b} + \nabla_p c(\mathbf{p}) u. \quad (3.112)$$

<sup>\*37</sup> This general argument is due to Diewert (1980; 597)[88].

<sup>\*38</sup> In the case where one or more components of  $\mathbf{b}$  are negative, we will require income to be large enough so that the demands  $d_n$  defined by (3.113) are nonnegative. Thus we require  $Y$  to be large enough so that  $\mathbf{b} + \nabla_p c(\mathbf{p})[Y - \mathbf{b}^T \mathbf{p}] / c(\mathbf{p}) \geq \mathbf{0}_N$ .

Substituting (3.111) into (3.112) leads to the following system of market demand functions:

$$d(Y, \mathbf{p}) \equiv \mathbf{x}[g(Y, \mathbf{p}), \mathbf{p}] = \mathbf{b} + \nabla_p c(\mathbf{p})[Y - \mathbf{b}^T \mathbf{p}] / c(\mathbf{p}). \tag{3.113}$$

To see how the geometry of this method of adding the vector of committed expenditures  $\mathbf{b}$  to a homothetic preferences cost function works, consider the case  $N = 2$  and let the vector of reference prices  $[p_1^*, p_2^*]$  be  $[1, 1]$ . In this case, the constraint (3.109) implies that

$$b_2 = -b_1. \tag{3.114}$$

In Figure 3.5 below, we assumed that  $b_1$  is positive so that  $b_2 = -b_1$  is negative. We drew a quadrant in Figure 3.5 with an origin at the point  $\mathbf{b} = [b_1, b_2] = [b_1, -b_1]$ . Now fill in this quadrant with the family of indifference curves that are dual to the unit cost function  $c(\mathbf{p})$ . Three of these indifference curves are graphed in Figure 3.5 that pass through the  $\mathbf{x}$  points,  $\mathbf{x}^0, \mathbf{x}^*$  (our point of approximation that has utility level  $u^*$ ) and  $\mathbf{x}^1$ .

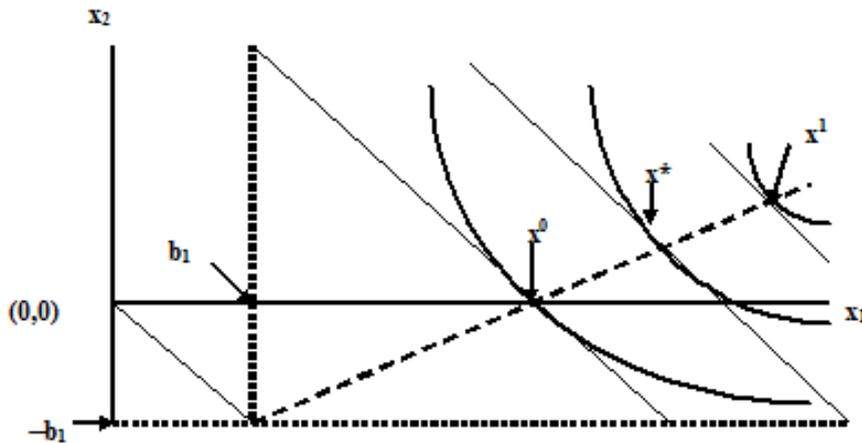


Fig. 3.5 The Addition of  $\mathbf{b}$  to a Flexible Unit Cost Function

There are also 4 parallel budget lines that correspond to income  $Y = 0$  (this is the line that passes through  $(0, 0)$  and  $(b_1, b_2)$ ),  $Y = Y^0$  (the corresponding consumer demand vector  $\mathbf{x}^0 = [x_1^0, x_2^0]$  has its  $x_2$  component equal to 0, i.e.,  $x_2^0 = 0$ ),<sup>\*39</sup>  $Y = Y^* \equiv \mathbf{p}^{*T} \mathbf{x}^*$  where  $\mathbf{x}^* = \nabla_p C(u^*, \mathbf{p}^*)$  is the consumer demand vector at the point of approximation for the flexible functional form, and  $Y = Y^1 = \mathbf{p}^{*T} \mathbf{x}^1$  where  $Y^1 > Y^*$ . Note that the points  $\mathbf{x}^0, \mathbf{x}^*$  and  $\mathbf{x}^1$  lie on the dashed line that starts at the point  $\mathbf{b}$ . Draw a straight line starting at the origin  $(0, 0)$  and passing through the point  $\mathbf{x}^*$ . It can be seen that this straight line is below the dashed line to the right of  $\mathbf{x}^*$ ; this means that the income elasticity of demand for commodity 1 is less than one while the income elasticity of demand for commodity 2 is greater than one. Thus it can be seen how this model can be consistent with arbitrary income elasticities of demand around a point of approximation,  $\mathbf{x}^*$ .

**Problem 17** Instead of estimating preferences using dual methods, it is possible to estimate primal utility functions directly. Suppose that  $\mathbf{x}^t \gg \mathbf{0}_N$  solves the consumer's period  $t$  utility maximization problem:

$$\max_{\mathbf{x}} \{F(\mathbf{x}) : \mathbf{p}^{tT} \mathbf{x} = Y^t; \mathbf{x} \geq \mathbf{0}_N\} \tag{i}$$

where  $\mathbf{p}^t \gg \mathbf{0}_N, Y^t > 0$  and  $F(\mathbf{x})$  is the consumer's differentiable utility function.

<sup>\*39</sup> In order to obtain nonnegative demands, we require that income  $Y$  be equal to or greater than  $Y^0$ .

(a) Show that the consumer's period  $t$  normalized price vector,  $\mathbf{p}^t/Y^t$ , satisfies the following system of equations (Hotelling (1935; 71)[243], Wold (1944; 69-71)[403] (1953; 145))[404]:

$$\mathbf{p}^t/Y^t = \nabla F(\mathbf{x}^t)/\mathbf{x}^{tT}\nabla F(\mathbf{x}^t). \quad (\text{ii})$$

*Hint:* Set up the Lagrangian for the constrained maximization problem and look at the resulting first order necessary conditions for an interior solution. Eliminate the Lagrange multiplier from these  $N + 1$  equations. The remaining  $N$  equations can be rewritten in the form (ii).

(b) Now assume that the utility function is  $g\{F(\mathbf{x})\}$ , where  $g(u)$  is a monotonic once differentiable function of one variable with  $g'(u) > 0$  for all  $u > 0$ . Find the counterparts to equations (ii) above.

(c) Find a flexible functional form for  $F(\mathbf{x})$  in the class of functions with no restrictions on  $F$ . You do not have to formally prove its flexibility; just exhibit what you think might be a flexible functional form. What system of equations does  $F$  have to satisfy in order to be flexible at a point  $\mathbf{x}^* \gg \mathbf{0}_N$ ? Thus what is the minimal number of parameters that  $F$  must have in order to be flexible?

(d) Substitute this candidate flexible functional form into the system of econometric estimating equations defined above by (ii). Can all of the parameters of  $F$  be identified?

(e) The material in this section treats commodity demand vectors  $\mathbf{x}^t$  as the dependent variables in a system of econometric estimating equations while income  $Y^t$  and the commodity price vector  $\mathbf{p}^t$  are regarded as independent variables. The parameters of a cost function are estimated using this framework. However, the material in this problem treats the commodity price vectors deflated by income or expenditure,  $\mathbf{p}^t/Y^t$ , as the dependent variables and the commodity demand vectors  $\mathbf{x}^t$  as the independent variables. The parameters of a utility function are estimated in this framework. Which framework is preferable in applied work?

**Problem 18** Assume that the twice continuously differentiable utility function  $F(\mathbf{x})$  satisfies the ray scaling assumption, (3.92).

(a) Show that the utility function and its derivatives must satisfy the following 2 restrictions:

$$\mathbf{x}^{*T}\nabla F(\mathbf{x}^*) = F(\mathbf{x}^*); \quad (\text{i})$$

$$\mathbf{x}^{*T}\nabla^2 F(\mathbf{x}^*)\mathbf{x}^* = 0. \quad (\text{ii})$$

*Hint:* define the functions  $g(\lambda) \equiv F(\lambda\mathbf{x}^*)$  and  $h(\lambda) \equiv \lambda F(\mathbf{x}^*)$ . Since (3.92) holds,  $g(\lambda) = h(\lambda)$  for  $\lambda \geq 0$ . Now differentiate  $g$  and  $h$  with respect to  $\lambda$  once and then again and set  $\lambda = 1$ .

(b) In view of part (a) of this problem, how many independent parameters must  $F(\mathbf{x})$  have in order to be a parsimonious flexible functional form at the point  $\mathbf{x}^*$  where  $F$  and  $F^*$  both satisfy the ray scaling assumption (3.92)?

*Hint:* In addition to the restrictions (i) and (ii) in part (a) of this problem, we need to cardinalize the utility function by adding the following restriction:

$$F(\mathbf{x}^*) = \alpha > 0 \quad \text{where } \alpha \text{ is an arbitrary positive number.} \quad (\text{iii})$$

(c) Use the flexible functional form for  $F$  that you suggested in problem 17 above and impose ray scaling on it. What additional restrictions on the parameters does ray scaling imply on your suggested functional form? Set up a system of estimating equations using Wold's Identity using your suggested flexible functional form with ray scaling imposed. Are all of the parameters in your estimating model identified?

### 3.9 Variable Profit Functions

Up to now, we have only considered technologies that produce one output. In reality, firms (and industries) usually produce many outputs. Hence, in this section, we consider technologies that produce many outputs while using many inputs.

Let  $S$  denote the technology set of a firm. We decompose the inputs and outputs of the firm into two sets of commodities: variable and fixed. Let  $\mathbf{y} \equiv [y_1, \dots, y_M]$  denote a vector of *variable net outputs* (if  $y_m > 0$ , then commodity  $m$  is an output while if  $y_m < 0$ , then commodity  $m$  is an input) and let  $\mathbf{x} \equiv [x_1, \dots, x_N]$  denote a nonnegative vector of “*fixed*” inputs<sup>\*40</sup>. Thus the technology set  $S$  is a set of feasible variable net output and fixed input vectors,  $(\mathbf{x}, \mathbf{y})$ .

Let  $\mathbf{p} \gg \mathbf{0}_M$  be a strictly positive vector of variable net output prices that the firm faces during a production period. Then conditional on a given vector of fixed inputs  $\mathbf{x}$ , we assume that the firm attempts to solve the following *variable profit maximization problem*:

$$\max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \mathbf{x}) \in S\} \equiv \pi(\mathbf{p}, \mathbf{x}). \quad (3.115)$$

Some regularity conditions on the technology set  $S$  are required in order to ensure that the maximum in (3.115) exists. A simple set of sufficient conditions are:<sup>\*41</sup>

- $S$  is a closed set in  $\mathbb{R}^{M+N}$ ; (3.116)
- For each  $\mathbf{x} \geq \mathbf{0}_N$ , the set of  $\mathbf{y}$  such that  $(\mathbf{y}, \mathbf{x}) \in S$  is not empty and is bounded from above; i.e., for each  $\mathbf{x} \geq \mathbf{0}_N$  and  $\mathbf{y}$  such that  $(\mathbf{y}, \mathbf{x}) \in S$ , there exists a number  $b(\mathbf{x})$  such that  $\mathbf{y} \leq b(\mathbf{x})\mathbf{1}_M$ . (3.117)

Condition (3.117) means that for each vector of fixed inputs,  $\mathbf{x} \geq \mathbf{0}_N$ , the amount of each variable net output that can be produced by the technology is bounded from above, which is not a restrictive condition.

Note that (3.115) serves to define the firm’s *variable profit function*,<sup>\*42</sup>  $\pi(\mathbf{p}, \mathbf{x})$ ; i.e.,  $\pi(\mathbf{p}, \mathbf{x})$  is equal to the optimized objective function in (3.115) and is regarded as a function of the net output prices for variable commodities that the firm faces,  $\mathbf{p}$ , as well as a function of the vector of fixed inputs,  $\mathbf{x}$ , that the firm has at its disposal. Just as in section 3.2 above where we showed that the cost function  $C(\mathbf{y}, \mathbf{p})$  satisfied a number of regularity conditions without assuming much about the production function, we can now show that the profit function  $\pi(\mathbf{p}, \mathbf{x})$  satisfies some regularity conditions without assuming much about the technology set  $S$ .

**Theorem 8** McFadden (1966)[307] (1978)[308], Gorman (1968)[201], Diewert (1973)[74]: Suppose the technology set  $S$  satisfies (3.116) and (3.117). Then the variable profit function  $\pi(\mathbf{p}, \mathbf{x})$  defined by (3.115) has the following properties with respect to  $\mathbf{p}$  for each  $\mathbf{x} \geq \mathbf{0}_N$ :

*Property 1:*  $\pi(\mathbf{p}, \mathbf{x})$  is *positively linearly homogeneous* in  $\mathbf{p}$  for each fixed  $\mathbf{x} \geq \mathbf{0}_N$ ; i.e.,

$$\pi(\lambda \mathbf{p}, \mathbf{x}) = \lambda \pi(\mathbf{p}, \mathbf{x}) \quad \text{for all } \lambda > 0, \mathbf{p} \gg \mathbf{0}_M \text{ and } \mathbf{x} \geq \mathbf{0}_N. \quad (3.118)$$

<sup>\*40</sup> These “fixed” inputs may only be fixed in the short run.

<sup>\*41</sup> Let  $\mathbf{x} \geq \mathbf{0}_N$ . Then by (3.117), there exists  $\mathbf{y}_x$  such that  $(\mathbf{y}_x, \mathbf{x}) \in S$ . Define the closed and bounded set  $B(\mathbf{x}, \mathbf{p}) \equiv \{\mathbf{y} : \mathbf{y} \leq b(\mathbf{x})\mathbf{1}_M; \mathbf{p}^T \mathbf{y} \geq \mathbf{p}^T \mathbf{y}_x\}$ . It can be seen that the constraint  $(\mathbf{y}, \mathbf{x}) \in S$  in (3.115) can be replaced by the constraint  $(\mathbf{y}, \mathbf{x}) \in S \cap B(\mathbf{x}, \mathbf{p})$ . Using (3.114),  $S \cap B(\mathbf{x}, \mathbf{p})$  is a closed and bounded set so that the maximum in (3.115) will exist.

<sup>\*42</sup> This concept is due to Hicks (1946; 319)[222] and Samuelson (1953-54)[343], who determined many of its properties using primal optimization techniques. For more general approaches to this function using duality theory, see Gorman (1968)[201], McFadden (1966)[307] (1978)[308] and Diewert (1973)[74]. McFadden used the term “conditional profit function” while Diewert used the term “variable profit function”.

*Property 2:*  $\pi(\mathbf{p}, \mathbf{x})$  is a convex function of  $\mathbf{p}$  for each  $\mathbf{x} \geq \mathbf{0}_N$ ; i.e.,

$$\begin{aligned} \mathbf{x} \geq \mathbf{0}_N, \mathbf{p}^1 \gg \mathbf{0}_M; \mathbf{p}^2 \gg \mathbf{0}_M; 0 < \lambda < 1 \text{ implies} \\ \pi(\lambda \mathbf{p}^1 + (1 - \lambda) \mathbf{p}^2, \mathbf{x}) \leq \lambda \pi(\mathbf{p}^1, \mathbf{x}) + (1 - \lambda) \pi(\mathbf{p}^2, \mathbf{x}). \end{aligned} \quad (3.119)$$

**Problem 19** Prove Theorem 8.

*Hint:* Properties 1 and 2 above for  $\pi(\mathbf{p}, \mathbf{x})$  are analogues to Properties 2 and 4 for the cost function  $C(\mathbf{y}, \mathbf{p})$  in Theorem 1 above and can be proven in the same manner.

We now ask whether a knowledge of the profit function  $\pi(\mathbf{p}, \mathbf{x})$  is sufficient to determine the underlying technology set  $S$ . As was the case in section 3.3 above, the answer to this question is *yes*, but with some qualifications.

To see how to use a given profit function satisfying the 2 regularity conditions listed in Theorem 8 to determine the technology set that generated it, pick an arbitrary vector of fixed inputs  $\mathbf{x} \geq \mathbf{0}_N$  and an arbitrary vector of positive prices,  $\mathbf{p}^1 \gg \mathbf{0}_M$ . Now use the given profit function  $\pi$  to define the following isoprofit surface:  $\{\mathbf{y} : \mathbf{p}^{1T} \mathbf{y} = \pi(\mathbf{p}^1, \mathbf{x})\}$ . This isoprofit surface must be tangent to the set of net output combinations  $\mathbf{y}$  that are feasible, given that the vector of fixed inputs  $\mathbf{x}$  is available to the firm, which is the conditional on  $\mathbf{x}$  production possibilities set,  $S(\mathbf{x}) \equiv \{\mathbf{x} : (\mathbf{y}, \mathbf{x}) \in S\}$ . It can be seen that this isoprofit surface and the set lying below it must contain the set  $S(\mathbf{x})$ ; i.e., the following *halfspace*  $M(\mathbf{x}, \mathbf{p}^1)$ , contains  $S(\mathbf{x})$ :

$$M(\mathbf{x}, \mathbf{p}^1) \equiv \{\mathbf{y} : \mathbf{p}^{1T} \mathbf{y} \leq \pi(\mathbf{p}^1, \mathbf{x})\}. \quad (3.120)$$

Pick another positive vector of prices,  $\mathbf{p}^2 \gg \mathbf{0}_M$  and it can be seen, repeating the above argument, that the halfspace  $M(\mathbf{x}, \mathbf{p}^2) \equiv \{\mathbf{y} : \mathbf{p}^{2T} \mathbf{y} \leq \pi(\mathbf{p}^2, \mathbf{x})\}$  must also contain the conditional on  $\mathbf{x}$  production possibilities set  $S(\mathbf{x})$ . Thus  $S(\mathbf{x})$  must belong to the intersection of the two halfspaces  $M(\mathbf{x}, \mathbf{p}^1)$  and  $M(\mathbf{x}, \mathbf{p}^2)$ . Continuing to argue along these lines, it can be seen that  $S(\mathbf{x})$  must be contained in the following set, which is the intersection over all  $\mathbf{p} \gg \mathbf{0}_M$  of all of the supporting halfspaces to  $S(\mathbf{x})$ :

$$M(\mathbf{x}) \equiv \bigcap_{\mathbf{p} \gg \mathbf{0}_M} M(\mathbf{x}, \mathbf{p}). \quad (3.121)$$

Note that  $M(\mathbf{x})$  is defined using just the given profit function,  $\pi(\mathbf{p}, \mathbf{x})$ . Note also that since each of the sets in the intersection,  $M(\mathbf{x}, \mathbf{p})$ , is a convex set, then  $M(\mathbf{x})$  is also a convex set. Since  $S(\mathbf{x})$  is a subset of each  $M(\mathbf{x}, \mathbf{p})$ , it must be the case that  $S(\mathbf{x})$  is also a subset of  $M(\mathbf{x})$ ; i.e., we have

$$S(\mathbf{x}) \subset M(\mathbf{x}). \quad (3.122)$$

Is it the case that  $S(\mathbf{x})$  is equal to  $M(\mathbf{x})$ ? In general, the answer is *no*;  $M(\mathbf{x})$  forms an *outer approximation* to the true conditional production possibilities set  $S(\mathbf{x})$ . To see why this is, see Figure 3.6 below. The boundary of the set  $M(\mathbf{x})$  partly coincides with the boundary of  $S(\mathbf{x})$  but it encloses a bigger set: the backward bending parts of the true production frontier are replaced by the dashed lines that are parallel to the  $y_1$  axis and the  $y_2$  axis and the inward bending part of the true production frontier is replaced by the dashed line that is tangent to the two regions where the boundary of  $M(\mathbf{x})$  coincides with the boundary of  $S(\mathbf{x})$ . However, if the producer is a price taker in the two output markets, then it can be seen that *we will never observe the producer's nonconvex or backward bending parts of the production frontier*.

Figure 3.6 illustrated the case where the two variable commodities were both outputs. Figure 3.7 illustrates the one variable output, one variable input geometry that corresponds to (3.122). In Figure 3.7,  $y_1$  is the variable input and  $y_2$  is the variable output. Again, the boundary of the set  $M(\mathbf{x})$  partly coincides with the boundary of  $S(\mathbf{x})$  but it encloses a bigger set: the downward bending

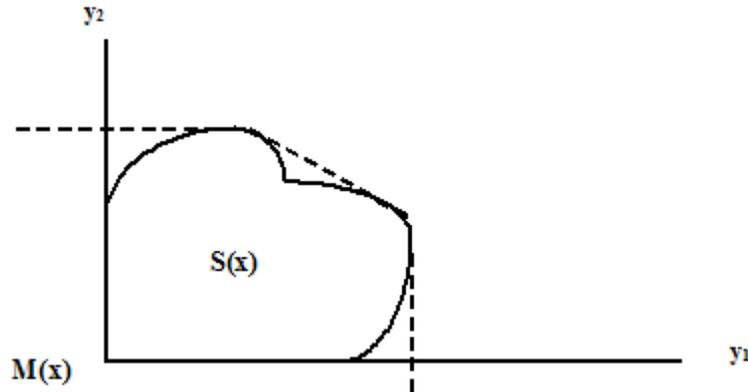


Fig. 3.6 The Geometry of the Two Output Maximization Problem

part of the true production frontier is replaced by the dashed line that is parallel to the  $y_1$  axis and the nonconvex part of the true production frontier is replaced by the dashed line that is tangent to the two regions where the boundary of  $M(\mathbf{x})$  coincides with the boundary of  $S(\mathbf{x})$ . Again, if the producer is a price taker in the two variable markets, then it can be seen that *we will never observe the producer's nonconvex or downward bending parts of the production frontier*.

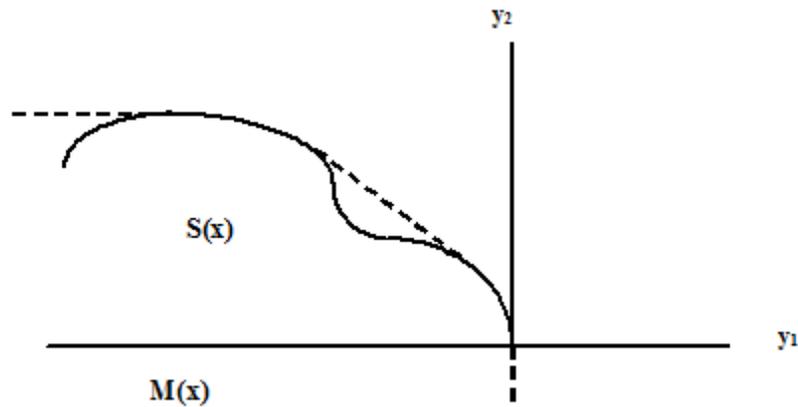


Fig. 3.7 The Geometry of the One Output One Input Maximization Problem

What are conditions on the technology set  $S$  (and hence on the conditional technology sets  $S(\mathbf{x})$ ) that will ensure that the outer approximation sets  $M(\mathbf{x})$ , constructed using the variable profit function  $\pi(\mathbf{p}, \mathbf{x})$ , will equal the true technology sets  $S(\mathbf{x})$ ? It can be seen that the following two conditions on  $S$  (in addition to conditions (3.116) and (3.117)) are the required conditions:

- For every  $\mathbf{x} \geq \mathbf{0}_N$ , the set  $S(\mathbf{x}) \equiv \{\mathbf{y} : (\mathbf{y}, \mathbf{x}) \in S\}$  has the following *free disposal property*:  $\mathbf{y}^1 \in S(\mathbf{x}), \mathbf{y}^2 \leq \mathbf{y}^1$  implies  $\mathbf{y}^2 \in S(\mathbf{x})$ ; (3.123)
- For every  $\mathbf{x} \geq \mathbf{0}_N$ , the set  $S(\mathbf{x}) \equiv \{\mathbf{y} : (\mathbf{y}, \mathbf{x}) \in S\}$  is convex.<sup>\*43</sup> (3.124)

<sup>\*43</sup> If  $N = 1$  so that there is only one fixed input, then given a producible net output vector  $\mathbf{y} \in \mathbb{R}^M$ , we can define the

Conditions (3.123) and (3.124) are the conditions on the technology set  $S$  that are counterparts to the two regularity conditions of nondecreasingness and quasiconcavity<sup>\*44</sup> that were made on the production function,  $F(\mathbf{x})$ , in section 3.3 above in order to obtain a duality between cost and production functions. If the firm is behaving as a price taker in variable commodity markets, it can be seen that it is not restrictive from an empirical point of view to assume that  $S$  satisfies conditions (3.123) and (3.124), just as it was not restrictive to assume that the production function was nondecreasing and quasiconcave in the context of the producer's (competitive) cost minimization problem studied earlier.

The next result provides a counterpart to Shephard's Lemma, Theorem 5 in section 3.4 above.

**Theorem 9** *Hotelling's (1932; 594)[242] Lemma:*<sup>\*45</sup> If the profit function  $\pi(\mathbf{p}, \mathbf{x})$  satisfies the properties listed in Theorem 8 above and in addition is once differentiable with respect to the components of the variable commodity prices at the point  $(\mathbf{p}^*, \mathbf{x}^*)$  where  $\mathbf{x}^* \geq \mathbf{0}_N$  and  $\mathbf{p}^* \gg \mathbf{0}_M$ , then

$$\mathbf{y}^* = \nabla_{\mathbf{p}}\pi(\mathbf{p}^*, \mathbf{x}^*) \quad (3.125)$$

where  $\nabla_{\mathbf{p}}\pi(\mathbf{p}^*, \mathbf{x}^*)$  is the vector of first order partial derivatives of variable profit with respect to variable commodity prices and  $\mathbf{y}^*$  is any solution to the profit maximization problem

$$\max_{\mathbf{y}}\{\mathbf{p}^{*T}\mathbf{y} : (\mathbf{y}, \mathbf{x}^*) \in S\} \equiv \pi(\mathbf{p}^*, \mathbf{x}^*). \quad (3.126)$$

Under these differentiability hypotheses, it turns out that the  $\mathbf{y}^*$  solution to (3.126) is unique.

**Proof.** Let  $\mathbf{y}^*$  be any solution to the profit maximization problem (3.126). Since  $\mathbf{y}^*$  is feasible for the profit maximization problem when the variable commodity price vector is changed to an arbitrary  $\mathbf{p} \gg \mathbf{0}_M$ , it follows that

$$\mathbf{p}^T\mathbf{y}^* \leq \pi(\mathbf{p}, \mathbf{x}^*) \quad \text{for every } \mathbf{p} \gg \mathbf{0}_M. \quad (3.127)$$

Since  $\mathbf{y}^*$  is a solution to the profit maximization problem (3.126) when  $\mathbf{p} = \mathbf{p}^*$ , we must have

$$\mathbf{p}^{*T}\mathbf{y}^* = \pi(\mathbf{p}^*, \mathbf{x}^*). \quad (3.128)$$

But (3.127) and (3.128) imply that the function of  $M$  variables,  $g(\mathbf{p}) \equiv \mathbf{p}^T\mathbf{y}^* - \pi(\mathbf{p}, \mathbf{x}^*)$  is nonpositive for all  $\mathbf{p} \gg \mathbf{0}_M$  with  $g(\mathbf{p}^*) = 0$ . Hence,  $g(\mathbf{p})$  attains a global maximum at  $\mathbf{p} = \mathbf{p}^*$  and since  $g(\mathbf{p})$  is differentiable with respect to the variable commodity prices  $\mathbf{p}$  at this point, the following first order necessary conditions for a maximum must hold at this point:

$$\nabla_{\mathbf{p}}g(\mathbf{p}^*) = \mathbf{y}^* - \nabla_{\mathbf{p}}\pi(\mathbf{p}^*, \mathbf{x}^*) = \mathbf{0}_M. \quad (3.129)$$

Now note that (3.129) is equivalent to (3.125). If  $\mathbf{y}^{**}$  is any other solution to the profit maximization problem (3.126), then repeat the above argument to show that

$$\begin{aligned} \mathbf{y}^{**} &= \nabla_{\mathbf{p}}\pi(\mathbf{p}^*, \mathbf{x}^*) \\ &= \mathbf{y}^* \end{aligned} \quad (3.130)$$

where the second equality follows using (3.129). Hence  $\mathbf{y}^{**} = \mathbf{y}^*$  and the solution to (3.126) is unique. ■

---

(fixed) *input requirements function* that corresponds to the technology set  $S$  as  $g(\mathbf{y}) \equiv \min_{\mathbf{x}}\{\mathbf{x} : (\mathbf{y}, \mathbf{x}) \in S\}$ . In this case, condition (3.123) becomes the following condition: the input requirements function  $g(\mathbf{y})$  is *quasiconvex* in  $\mathbf{y}$ . For additional material on this one fixed input model, see Diewert (1974c)[79].

<sup>\*44</sup> Recall conditions (3.11) and (3.12) in section 3.3.

<sup>\*45</sup> See also Gorman (1968)[201] and Diewert (1974a, 137)[77].

Hotelling's Lemma may be used in order to derive systems of variable commodity output supply and input demand functions just as we used Shephard's Lemma to generate systems of cost minimizing input demand functions; for examples of this use of Hotelling's Lemma, see Diewert (1974a; 137-139)[77].

If we are willing to make additional assumptions about the underlying firm production possibilities set  $S$ , then we can deduce that  $\pi(\mathbf{p}, \mathbf{x})$  satisfies some additional properties. One such additional property is the following one:  $S$  is subject to the *free disposal of fixed inputs* if it has the following property:

$$\mathbf{x}^2 > \mathbf{x}^1 \geq \mathbf{0}_N \text{ and } (\mathbf{y}, \mathbf{x}^1) \in S \text{ implies } (\mathbf{y}, \mathbf{x}^2) \in S. \quad (3.131)$$

The above property means if the vector of fixed inputs  $\mathbf{x}^1$  is sufficient to produce the vector of variable inputs and outputs  $\mathbf{y}$  and if we have at our disposal a bigger vector of fixed inputs  $\mathbf{x}^2$ , then  $\mathbf{y}$  is still producible by the technology that is represented by the set  $S$ .

**Theorem 10** <sup>\*46</sup> Suppose the technology set  $S$  satisfies assumptions (3.116) and (3.117) above.

(a) If in addition,  $S$  has the following property:<sup>\*47</sup>

$$\text{For every } \mathbf{x} \geq \mathbf{0}_N, (\mathbf{0}_M, \mathbf{x}) \in S; \quad (3.132)$$

then for every  $\mathbf{p} \gg \mathbf{0}_M$  and  $\mathbf{x} \geq \mathbf{0}_N$ ,  $\pi(\mathbf{p}, \mathbf{x}) \geq 0$ ; i.e., the variable profit function is *nonnegative* if (a) holds.

(b) If  $S$  is a convex set, then for each  $\mathbf{p} \gg \mathbf{0}_M$ , then  $\pi(\mathbf{p}, \mathbf{x})$  is a *concave function* of  $\mathbf{x}$  over the set  $\Omega \equiv \{\mathbf{x} : \mathbf{x} \geq \mathbf{0}_N\}$ .

(c) If  $S$  is a cone so that the technology is subject to constant returns to scale, then  $\pi(\mathbf{p}, \mathbf{x})$  is (positively) *homogeneous of degree one* in the components of  $\mathbf{x}$ .

(d) If  $S$  is subject to the free disposal of fixed inputs, then

$$\mathbf{p} \gg \mathbf{0}_M, \mathbf{x}^2 > \mathbf{x}^1 \geq \mathbf{0}_N \text{ implies } \pi(\mathbf{p}, \mathbf{x}^2) \geq \pi(\mathbf{p}, \mathbf{x}^1); \quad (3.133)$$

i.e.,  $\pi(\mathbf{p}, \mathbf{x})$  is *nondecreasing* in the components of  $\mathbf{x}$ .

**Proof of (a):** Let  $\mathbf{p} \gg \mathbf{0}_M$  and  $\mathbf{x} \geq \mathbf{0}_N$ . Then

$$\begin{aligned} \pi(\mathbf{p}, \mathbf{x}) &\equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \mathbf{x}) \in S\} \\ &\geq \mathbf{p}^T \mathbf{0}_M \quad \text{since by (3.132), } (\mathbf{0}_M, \mathbf{x}) \in S \text{ and hence is feasible for the problem} \\ &= 0. \end{aligned} \quad (3.134)$$

**Proof of (b):** Let  $\mathbf{p} \gg \mathbf{0}_M, \mathbf{x}^1 \geq \mathbf{0}_N, \mathbf{x}^2 \geq \mathbf{0}_N$  and  $0 < \lambda < 1$ . Then

$$\begin{aligned} \pi(\mathbf{p}, \mathbf{x}^1) &\equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \mathbf{x}^1) \in S\} \\ &= \mathbf{p}^T \mathbf{y}^1 \quad \text{where } (\mathbf{y}^1, \mathbf{x}^1) \in S; \end{aligned} \quad (3.135)$$

$$\begin{aligned} \pi(\mathbf{p}, \mathbf{x}^2) &\equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \mathbf{x}^2) \in S\} \\ &= \mathbf{p}^T \mathbf{y}^2 \quad \text{where } (\mathbf{y}^2, \mathbf{x}^2) \in S. \end{aligned} \quad (3.136)$$

Since  $S$  is assumed to be a convex set, we have

$$\lambda(\mathbf{y}^1, \mathbf{x}^1) + (1 - \lambda)(\mathbf{y}^2, \mathbf{x}^2) = [\lambda\mathbf{y}^1 + (1 - \lambda)\mathbf{y}^2, \lambda\mathbf{x}^1 + (1 - \lambda)\mathbf{x}^2] \in S. \quad (3.137)$$

<sup>\*46</sup> The results in this Theorem are essentially due to Samuelson (1953-54; 20)[343], Gorman (1968)[201] and Diewert (1973)[74] (1974a; 136)[77] but they are packaged in a somewhat different form in this chapter.

<sup>\*47</sup> This property says that the technology can always produce no variable outputs and utilize no variable inputs given any vector of fixed inputs  $\mathbf{x}$ .

Using the definition of  $\pi$ , we have:

$$\begin{aligned}
\pi(\mathbf{p}, \lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) &\equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \lambda \mathbf{x}^1 + (1 - \lambda) \mathbf{x}^2) \in S\} \\
&\geq \mathbf{p}^T [\lambda \mathbf{y}^1 + (1 - \lambda) \mathbf{y}^2] \quad \text{since by (3.137), } \lambda \mathbf{y}^1 + (1 - \lambda) \mathbf{y}^2 \text{ is feasible for the problem} \\
&= \lambda \mathbf{p}^T \mathbf{y}^1 + (1 - \lambda) \mathbf{p}^T \mathbf{y}^2 \\
&= \lambda \pi(\mathbf{p}, \mathbf{x}^1) + (1 - \lambda) \pi(\mathbf{p}, \mathbf{x}^2) \quad \text{using (3.135) and (3.136)}.
\end{aligned} \tag{3.138}$$

**Proof of (c):** Let  $\mathbf{p} \gg \mathbf{0}_M$ ,  $\mathbf{x}^* \geq \mathbf{0}_N$  and  $\lambda > 0$ . Then

$$\begin{aligned}
\pi(\mathbf{p}, \mathbf{x}^*) &\equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \mathbf{x}^*) \in S\} \\
&= \mathbf{p}^T \mathbf{y}^* \quad \text{where } (\mathbf{y}^*, \mathbf{x}^*) \in S.
\end{aligned} \tag{3.139}$$

Since  $S$  is a cone and since  $(\mathbf{y}^*, \mathbf{x}^*) \in S$ , then we have  $(\lambda \mathbf{y}^*, \lambda \mathbf{x}^*) \in S$  as well. Hence, using a feasibility argument:

$$\begin{aligned}
\pi(\mathbf{p}, \lambda \mathbf{x}^*) &\equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \lambda \mathbf{x}^*) \in S\} \\
&\geq \mathbf{p}^T \lambda \mathbf{y}^* \quad \text{since } (\lambda \mathbf{y}^*, \lambda \mathbf{x}^*) \in S \text{ and hence is feasible for the problem} \\
&= \lambda \mathbf{p}^T \mathbf{y}^*.
\end{aligned} \tag{3.140}$$

Now *suppose* that the strict inequality in (3.140) holds so that

$$\begin{aligned}
\pi(\mathbf{p}, \lambda \mathbf{x}^*) &\equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \lambda \mathbf{x}^*) \in S\} \\
&= \mathbf{p}^T \mathbf{y}^{**} \quad \text{where } (\mathbf{y}^{**}, \lambda \mathbf{x}^*) \in S \\
&> \lambda \mathbf{p}^T \mathbf{y}^*.
\end{aligned} \tag{3.141}$$

Since  $S$  is a cone and since  $(\mathbf{y}^{**}, \lambda \mathbf{x}^*) \in S$ , then we have  $(\lambda^{-1} \mathbf{y}^{**}, \mathbf{x}^*) \in S$  as well. Thus  $\lambda^{-1} \mathbf{y}^{**}$  is feasible for the maximization problem (3.139) that defined  $\pi(\mathbf{p}, \mathbf{x}^*)$  and so

$$\begin{aligned}
\mathbf{p}^T \mathbf{y}^* &= \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \mathbf{x}^*) \in S\} \quad \text{using (3.139)} \\
&\geq \mathbf{p}^T \lambda^{-1} \mathbf{y}^{**} \quad \text{since } \lambda^{-1} \mathbf{y}^{**} \text{ is feasible for the problem} \\
&= \lambda^{-1} \mathbf{p}^T \mathbf{y}^{**}
\end{aligned} \tag{3.142}$$

or since  $\lambda > 0$ , (3.142) is equivalent to

$$\lambda \mathbf{p}^T \mathbf{y}^* \geq \mathbf{p}^T \mathbf{y}^{**} > \lambda \mathbf{p}^T \mathbf{y}^* \quad \text{using (3.141)}. \tag{3.143}$$

But (3.143) implies that  $\lambda \mathbf{p}^T \mathbf{y}^* > \lambda \mathbf{p}^T \mathbf{y}^*$ , which is impossible and hence our *supposition* is false and the desired result follows.

**Proof of (d):** Let  $\mathbf{p} \gg \mathbf{0}_M$ ,  $\mathbf{x}^2 > \mathbf{x}^1 \geq \mathbf{0}_N$ . Using the definition of  $\pi(\mathbf{p}, \mathbf{x}^1)$ , we have

$$\begin{aligned}
\pi(\mathbf{p}, \mathbf{x}^1) &\equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \mathbf{x}^1) \in S\} \\
&= \mathbf{p}^T \mathbf{y}^1 \quad \text{where } (\mathbf{y}^1, \mathbf{x}^1) \in S.
\end{aligned} \tag{3.144}$$

Using the free disposal property (3.131) for  $S$ , since  $(\mathbf{y}^1, \mathbf{x}^1) \in S$  and  $\mathbf{x}^2 > \mathbf{x}^1$ , we have

$$(\mathbf{y}^1, \mathbf{x}^2) \in S. \tag{3.145}$$

Using the definition of  $\pi(\mathbf{p}, \mathbf{x}^2)$ , we have

$$\begin{aligned}\pi(\mathbf{p}, \mathbf{x}^2) &\equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \mathbf{x}^2) \in S\} \\ &\geq \mathbf{p}^T \mathbf{y}^1 \quad \text{since by (3.145), } (\mathbf{y}^1, \mathbf{x}^2) \text{ is feasible} \\ &= \pi(\mathbf{p}, \mathbf{x}^1) \quad \text{using (3.144).}\end{aligned}\tag{3.146}$$

■

Note that if the technology set  $S$  satisfies the minimal regularity conditions (3.116) and (3.117) plus all of the additional conditions that are listed in Theorem 10 above (we shall call such a technology set a *regular technology set*), then the associated variable profit function  $\pi(\mathbf{p}, \mathbf{x})$  will have *all* of the regularity conditions with respect to its fixed input vector  $\mathbf{x}$  that a nonnegative, nondecreasing, concave and linearly homogeneous production function  $f(\mathbf{x})$  possesses with respect to its input vector  $\mathbf{x}$ .

Hotelling's Lemma enabled us to interpret the vector of first order partial derivatives of the variable profit function with respect to the components of the variable commodity price vector  $\mathbf{p}$ ,  $\nabla_{\mathbf{p}}\pi(\mathbf{p}, \mathbf{x})$ , as the producer's vector of variable profit maximizing output supply (and the negative of variable input demand) functions,  $\mathbf{y}(\mathbf{p}, \mathbf{x})$ , provided that the derivatives existed. If the first order partial derivatives of the variable profit function  $\pi(\mathbf{p}, \mathbf{x})$  with respect to the components of the fixed input vector  $\mathbf{x}$  exist, then this vector of derivatives,  $\nabla_{\mathbf{x}}\pi(\mathbf{p}, \mathbf{x})$ , can also be given an *economic interpretation as a vector of shadow prices or imputed contributions to profit of adding marginal units of fixed inputs*. The following result also shows that these derivatives can be interpreted as competitive input prices for the "fixed" factors if they are allowed to become variable.

**Theorem 11** <sup>\*48</sup> Suppose the technology set  $S$  satisfies assumptions (3.116) and (3.117) above and in addition is a convex set. Suppose in addition that  $\mathbf{p}^* \gg \mathbf{0}_M$ ,  $\mathbf{x}^* \geq \mathbf{0}_N$  and that the vector of derivatives,  $\nabla_{\mathbf{x}}\pi(\mathbf{p}^*, \mathbf{x}^*) \equiv \mathbf{w}^*$ , exists. Then  $\mathbf{x}^*$  is a solution to the following *long run profit maximization problem* that allows the "fixed" inputs  $\mathbf{x}$  to be variable:

$$\max_{\mathbf{x}} \{\pi(\mathbf{p}^*, \mathbf{x}) - \mathbf{w}^{*T} \mathbf{x} : \mathbf{x} \geq \mathbf{0}_N\}.\tag{3.147}$$

**Proof.** Part (b) of Theorem 10 above implies that  $\pi(\mathbf{p}^*, \mathbf{x})$  is a concave function of  $\mathbf{x}$  over the set  $\Omega \equiv \{\mathbf{x} : \mathbf{x} \geq \mathbf{0}_N\}$ . The function  $-\mathbf{w}^{*T} \mathbf{x}$  is linear in  $\mathbf{x}$  and hence is also a concave function of  $\mathbf{x}$  over  $\Omega$ . Hence  $f(\mathbf{x})$  defined for  $\mathbf{x} \geq \mathbf{0}_N$  as

$$f(\mathbf{x}) \equiv \pi(\mathbf{p}^*, \mathbf{x}) - \mathbf{w}^{*T} \mathbf{x}\tag{3.148}$$

is also a concave function in  $\mathbf{x}$  over the set  $\Omega$ . Since  $\mathbf{x}^* \geq \mathbf{0}_N$ ,  $\mathbf{x}^* \in \Omega$ . Hence using the third characterization of concavity and the differentiability of  $f(\mathbf{x})$  with respect to  $\mathbf{x}$  at  $\mathbf{x}^*$ , we have:

$$\begin{aligned}f(\mathbf{x}) &\leq f(\mathbf{x}^*) + \nabla_{\mathbf{x}}f(\mathbf{x}^*)^T(\mathbf{x} - \mathbf{x}^*) \quad \text{for all } \mathbf{x} \geq \mathbf{0}_N \\ &= \pi(\mathbf{p}^*, \mathbf{x}^*) - \mathbf{w}^{*T} \mathbf{x}^* + \mathbf{0}_N^T(\mathbf{x} - \mathbf{x}^*) \quad \text{since } \nabla_{\mathbf{x}}f(\mathbf{x}^*) = \nabla_{\mathbf{x}}\pi(\mathbf{p}^*, \mathbf{x}^*) - \mathbf{w}^* = \mathbf{0}_N \\ &= \pi(\mathbf{p}^*, \mathbf{x}^*) - \mathbf{w}^{*T} \mathbf{x}^*.\end{aligned}\tag{3.149}$$

But (3.148) and (3.149) show that  $\mathbf{x}^*$  solves the profit maximization problem (3.147). ■

**Corollary** If in addition to the above assumptions,  $\pi(\mathbf{p}, \mathbf{x})$  is differentiable with respect to the components of  $\mathbf{p}$  at the point  $(\mathbf{p}^*, \mathbf{x}^*)$ , so that  $\mathbf{y}^* \equiv \nabla_{\mathbf{p}}\pi(\mathbf{p}^*, \mathbf{x}^*)$  exists, then  $(\mathbf{y}^*, \mathbf{x}^*)$  solves the following long run profit maximization problem:

$$\Pi(\mathbf{p}^*, \mathbf{w}^*) \equiv \max_{\mathbf{y}, \mathbf{x}} \{\mathbf{p}^{*T} \mathbf{y} - \mathbf{w}^{*T} \mathbf{x} : (\mathbf{y}, \mathbf{x}) \in S\}.\tag{3.150}$$

<sup>\*48</sup> Related results can be found in Samuelson (1953-54; 10)[343] and Diewert (1974a; 140)[77].

**Proof.** Using Hotelling's Lemma, we know that  $\mathbf{y}^*$  solves the following variable profit maximization problem:

$$\pi(\mathbf{p}^*, \mathbf{x}^*) \equiv \max_{\mathbf{y}} \{\mathbf{p}^{*T} \mathbf{y} : (\mathbf{y}, \mathbf{x}^*) \in S\} = \mathbf{p}^{*T} \mathbf{y}^*. \quad (3.151)$$

Now look at the long run profit maximization problem defined by (3.150):

$$\begin{aligned} \Pi(\mathbf{p}^*, \mathbf{w}^*) &\equiv \max_{\mathbf{y}, \mathbf{x}} \{\mathbf{p}^{*T} \mathbf{y} - \mathbf{w}^{*T} \mathbf{x} : (\mathbf{y}, \mathbf{x}) \in S\} \\ &= \max_{\mathbf{x}} [\max_{\mathbf{y}} \{\mathbf{p}^{*T} \mathbf{y} : (\mathbf{y}, \mathbf{x}) \in S\} - \mathbf{w}^{*T} \mathbf{x}] \quad \text{where we have rewritten the} \\ &\quad \text{maximization problem as a two stage maximization problem} \\ &= \max_{\mathbf{x}} [\pi(\mathbf{p}^*, \mathbf{x}) - \mathbf{w}^{*T} \mathbf{x}] \quad \text{using the definition of } \pi(\mathbf{p}^*, \mathbf{x}) \\ &= \pi(\mathbf{p}^*, \mathbf{x}^*) - \mathbf{w}^{*T} \mathbf{x}^* \quad \text{using Theorem 11.} \end{aligned} \quad (3.152)$$

Hence with  $\mathbf{x} = \mathbf{x}^*$  being an  $\mathbf{x}$  solution to (3.152), we must have

$$\begin{aligned} \Pi(\mathbf{p}^*, \mathbf{w}^*) &\equiv \max_{\mathbf{y}, \mathbf{x}} \{\mathbf{p}^{*T} \mathbf{y} - \mathbf{w}^{*T} \mathbf{x} : (\mathbf{y}, \mathbf{x}) \in S\} \\ &= [\max_{\mathbf{y}} \{\mathbf{p}^{*T} \mathbf{y} : (\mathbf{y}, \mathbf{x}^*) \in S\} - \mathbf{w}^{*T} \mathbf{x}^*] \quad \text{letting } \mathbf{x} = \mathbf{x}^* \\ &= \mathbf{p}^{*T} \mathbf{y}^* - \mathbf{w}^{*T} \mathbf{x}^* \quad \text{using (3.151).} \end{aligned} \quad (3.153)$$

■

Hotelling's Lemma and Theorem 11 can be used as a convenient method for obtaining econometric estimating equations for determining the parameters that characterize a producer's technology set  $S$ . Assuming that  $S$  satisfies (3.116) and (3.117), we need only postulate a differentiable functional form for the producer's variable profit function,  $\pi(\mathbf{p}, \mathbf{x})$ , that is linearly homogeneous and convex in  $\mathbf{p}$ . Suppose that we have collected data on the fixed input vectors used by the firm in period  $t$ ,  $\mathbf{x}^t$ , and the net supply vectors for variable commodities produced in period  $t$ ,  $\mathbf{y}^t$ , for  $t = 1, \dots, T$  time periods as well as the corresponding variable commodity price vectors  $\mathbf{p}^t$ . Then the following  $M$  equations can be used in order to estimate the unknown parameters in  $\pi(\mathbf{p}, \mathbf{x})$ :

$$\mathbf{y}^t = \nabla_{\mathbf{p}} \pi(\mathbf{p}^t, \mathbf{x}^t) + \mathbf{u}^t; \quad t = 1, \dots, T \quad (3.154)$$

where  $\mathbf{u}^t$  is a vector of errors. If in addition, it can be assumed that the firm is optimizing with respect to its vector of fixed inputs in each period, where it faces the fixed input price vector  $\mathbf{w}^t$  in period  $t$ , then the following  $N$  equations can be added to (3.154) as additional estimating equations:

$$\mathbf{w}^t = \nabla_{\mathbf{x}} \pi(\mathbf{p}^t, \mathbf{x}^t) + \mathbf{v}^t; \quad t = 1, \dots, T \quad (3.155)$$

where  $\mathbf{v}^t$  is a vector of errors.\*<sup>49</sup>

### 3.10 The Comparative Statics Properties of Net Supply and Fixed Input Demand Functions

From Theorem 8 above, we know that the firm's variable profit function  $\pi(\mathbf{p}, \mathbf{x})$  is convex and linearly homogeneous in the components of the vector of variable commodity prices  $\mathbf{p}$  for each fixed input

\*<sup>49</sup> If the technology set  $S$  is subject to constant returns to scale and the data reflect this fact by "adding up" (so that  $\mathbf{p}^{tT} \mathbf{y}^t = \mathbf{w}^{tT} \mathbf{x}^t$  for  $t = 1, \dots, T$ ), then the error vectors  $\mathbf{u}^t$  and  $\mathbf{v}^t$  in (3.154) and (3.155) cannot be statistically independent since they will satisfy the constraint  $\mathbf{p}^{tT} \mathbf{u}^t = \mathbf{w}^{tT} \mathbf{v}^t$  for  $t = 1, \dots, T$ . Hence, under these circumstances, one of the  $M + N$  equations in (3.154) and (3.155) must be dropped in the system of estimating equations.

vector  $\mathbf{x}$ . Thus if  $\pi(\mathbf{p}, \mathbf{x})$  is twice continuously differentiable with respect to the components of  $\mathbf{p}$  at some point  $(\mathbf{p}, \mathbf{x})$ , then using Hotelling's Lemma, we can prove the following counterpart to Theorem 7 for the cost function.

**Theorem 12** Hotelling (1932; 597)[242], Hicks (1946; 321)[222], Diewert (1974a; 142-146)[77]: Suppose the variable profit function  $\pi(\mathbf{p}, \mathbf{x})$  is linearly homogeneous and convex in  $\mathbf{p}$  and in addition is twice continuously differentiable with respect to the components of  $\mathbf{p}$  at some point,  $(\mathbf{p}, \mathbf{x})$ . Then the system of variable profit maximizing net supply functions,  $\mathbf{y}(\mathbf{p}, \mathbf{x}) \equiv [y_1(\mathbf{p}, \mathbf{x}), \dots, y_M(\mathbf{p}, \mathbf{x})]^T$ , exists at this point and these net supply functions are once continuously differentiable. Form the  $M \times M$  matrix of net supply derivatives with respect to variable commodity prices,  $\mathbf{B} \equiv [\partial y_i(\mathbf{p}, \mathbf{x})/\partial p_j]$ , which has  $ij$  element equal to  $\partial y_i(\mathbf{p}, \mathbf{x})/\partial p_j$ . Then the matrix  $\mathbf{B}$  has the following properties:

$$\cdot \mathbf{B} = \mathbf{B}^T \quad \text{so that } \partial y_i(\mathbf{p}, \mathbf{x})/\partial p_j = \partial y_j(\mathbf{p}, \mathbf{x})/\partial p_i \text{ for all } i \neq j; \text{ }^{*50} \quad (3.156)$$

$$\cdot \mathbf{B} \text{ is positive semidefinite and} \quad (3.157)$$

$$\cdot \mathbf{B}\mathbf{p} = \mathbf{0}_M. \quad (3.158)$$

**Proof.** Hotelling's Lemma implies that the firm's system of variable profit maximizing net supply functions,  $\mathbf{y}(\mathbf{p}, \mathbf{x}) \equiv [y_1(\mathbf{p}, \mathbf{x}), \dots, y_M(\mathbf{p}, \mathbf{x})]^T$ , exists and is equal to

$$\mathbf{y}(\mathbf{p}, \mathbf{x}) = \nabla_{\mathbf{p}}\pi(\mathbf{p}, \mathbf{x}). \quad (3.159)$$

Differentiating both sides of (3.159) with respect to the components of  $\mathbf{p}$  gives us

$$\mathbf{B} \equiv [\partial y_i(\mathbf{p}, \mathbf{x})/\partial p_j] = \nabla_{\mathbf{p}\mathbf{p}}^2\pi(\mathbf{p}, \mathbf{x}). \quad (3.160)$$

Now property (3.156) follows from Young's Theorem in calculus. Property (3.157) follows from (3.160) and the fact that  $\pi(\mathbf{p}, \mathbf{x})$  is convex in  $\mathbf{p}$  and the fourth characterization of convexity. Finally, property (3.158) follows from the fact that the profit function is linearly homogeneous in  $\mathbf{p}$  and hence, using Part 2 of Euler's Theorem on homogeneous functions, (3.158) holds. ■

Note that property (3.157) implies the following properties on the net supply functions:

$$\frac{\partial y_m(\mathbf{p}, \mathbf{x})}{\partial p_m} \geq 0 \quad \text{for } m = 1, \dots, M. \quad (3.161)$$

Property (3.161) means that output supply curves cannot be downward sloping. However, if variable commodity  $m$  is an input, then  $y_m(\mathbf{p}, \mathbf{x})$  is negative. If we define the positive input demand function as

$$d_m(\mathbf{p}, \mathbf{x}) \equiv -y_m(\mathbf{p}, \mathbf{x}) \geq 0, \quad (3.162)$$

then the restriction (3.161) translates into  $\partial d_m(\mathbf{p}, \mathbf{x})/\partial p_m \leq 0$ , which means that variable input demand curves cannot be upward sloping.

Obviously, if the technology set is a convex cone, then the firm's competitive fixed input price functions,  $\mathbf{w}(\mathbf{p}, \mathbf{x}) \equiv \nabla_{\mathbf{x}}\pi(\mathbf{p}, \mathbf{x})$ , will satisfy properties analogous to the properties of cost minimizing input demand functions in Theorem 7.

**Theorem 13** Samuelson (1953-54; 10)[343], Diewert (1974a; 144-146)[77]: Suppose that the firm's technology set  $S$  is regular. Define the firm's variable profit function  $\pi(\mathbf{p}, \mathbf{x})$  by (3.115). Suppose that  $\pi(\mathbf{p}, \mathbf{x})$  is twice continuously differentiable with respect to the components of  $\mathbf{x}$  at some point

---

<sup>\*50</sup> These are the Hotelling (1932; 549)[242] and Hicks (1946; 321)[222] symmetry restrictions on supply functions.

$(\mathbf{p}, \mathbf{x})$  where  $\mathbf{p} \gg \mathbf{0}_M$  and  $\mathbf{x} \geq \mathbf{0}_N$ . Then the *system of fixed input price functions*<sup>\*51</sup>,  $\mathbf{w}(\mathbf{p}, \mathbf{x}) \equiv [w_1(\mathbf{p}, \mathbf{x}), \dots, w_N(\mathbf{p}, \mathbf{x})]^T$ , exists at this point<sup>\*52</sup> and these input price functions are once continuously differentiable. Form the  $N \times N$  matrix of fixed input price derivatives with respect to the fixed inputs,  $\mathbf{C} \equiv [\partial w_i(\mathbf{p}, \mathbf{x})/\partial x_j]$ , which has  $ij$  element equal to  $\partial w_i(\mathbf{p}, \mathbf{x})/\partial x_j$ . Then the matrix  $\mathbf{C}$  has the following properties:

$$\cdot \mathbf{C} = \mathbf{C}^T \quad \text{so that } \partial w_i(\mathbf{p}, \mathbf{x})/\partial x_j = \partial w_j(\mathbf{p}, \mathbf{x})/\partial x_i \text{ for all } i \neq j; \quad (3.163)$$

$$\cdot \mathbf{C} \text{ is negative semidefinite and} \quad (3.164)$$

$$\cdot \mathbf{C}\mathbf{x} = \mathbf{0}_N. \quad (3.165)$$

**Proof.** Using the results of Theorem 11, the firm's system of fixed input price functions,  $\mathbf{w}(\mathbf{p}, \mathbf{x}) \equiv [w_1(\mathbf{p}, \mathbf{x}), \dots, w_N(\mathbf{p}, \mathbf{x})]^T$ , exists and is equal to

$$\mathbf{w}(\mathbf{p}, \mathbf{x}) \equiv \nabla_{\mathbf{x}}\pi(\mathbf{p}, \mathbf{x}) \quad (\text{Samuelson's Lemma}). \quad (3.166)$$

Differentiating both sides of (3.166) with respect to the components of  $\mathbf{x}$  gives us

$$\mathbf{C} \equiv [\partial w_i(\mathbf{p}, \mathbf{x})/\partial x_j] = \nabla_{\mathbf{x}\mathbf{x}}^2\pi(\mathbf{p}, \mathbf{x}). \quad (3.167)$$

Now property (3.163) follows from Young's Theorem in calculus. Property (3.164) follows from (3.167) and the fact that  $\pi(\mathbf{p}, \mathbf{x})$  is concave in  $\mathbf{x}$ <sup>\*53</sup> and the fourth characterization of concavity. Finally, property (3.165) follows from the fact that the profit function is linearly homogeneous in  $\mathbf{x}$ <sup>\*54</sup> and hence, using Part 2 of Euler's Theorem on homogeneous functions, (3.165) holds. ■

Note that property (3.164) implies the following properties on the fixed input price functions:

$$\frac{\partial w_n(\mathbf{p}, \mathbf{x})}{\partial x_n} \leq 0 \quad \text{for } n = 1, \dots, N. \quad (3.168)$$

Property (3.168) means that the inverse input demand curves cannot be upward sloping.

If the firm's production possibilities set  $S$  is regular and if the corresponding variable profit function  $\pi(\mathbf{p}, \mathbf{x})$  is twice continuously differentiable with respect to all of its variables, then there will be additional restrictions on the derivatives of the variable net output supply functions  $\mathbf{y}(\mathbf{p}, \mathbf{x}) = \nabla_{\mathbf{p}}\pi(\mathbf{p}, \mathbf{x})$  and on the derivatives of the fixed input price functions  $\mathbf{w}(\mathbf{p}, \mathbf{x}) \equiv \nabla_{\mathbf{x}}\pi(\mathbf{p}, \mathbf{x})$ . Define the  $M \times N$  matrix of derivatives of the net output supply functions  $\mathbf{y}(\mathbf{p}, \mathbf{x})$  with respect to the components of the vector of fixed inputs  $\mathbf{x}$  as follows:

$$\mathbf{D} \equiv [\partial y_i(\mathbf{p}, \mathbf{x})/\partial x_j] = \nabla_{\mathbf{p}\mathbf{x}}^2\pi(\mathbf{p}, \mathbf{x}); \quad i = 1, \dots, M; \quad j = 1, \dots, N, \quad (3.169)$$

where the equalities in (3.169) follow by differentiating both sides of the Hotelling's Lemma relations,  $\mathbf{y}(\mathbf{p}, \mathbf{x}) = \nabla_{\mathbf{p}}\pi(\mathbf{p}, \mathbf{x})$ , with respect to the components of  $\mathbf{x}$ . Similarly, define the  $N \times M$  matrix of derivatives of the fixed input price functions  $\mathbf{w}(\mathbf{p}, \mathbf{x})$  with respect to the components of the vector of variable commodity prices  $\mathbf{p}$  as follows:

$$\mathbf{E} \equiv [\partial w_i(\mathbf{p}, \mathbf{x})/\partial p_j] = \nabla_{\mathbf{x}\mathbf{p}}^2\pi(\mathbf{p}, \mathbf{x}); \quad i = 1, \dots, N; \quad j = 1, \dots, M, \quad (3.170)$$

<sup>\*51</sup> The functions  $\mathbf{w}(\mathbf{p}, \mathbf{x})$  can also be interpreted as the producer's system of *inverse demand functions for fixed inputs*.

<sup>\*52</sup> The assumption that  $S$  is regular implies that  $S$  has the free disposal property in fixed inputs property (3.131), which implies by part (d) of Theorem 10 that  $\pi(\mathbf{p}, \mathbf{x})$  is nondecreasing in  $\mathbf{x}$  and this in turn implies that  $\mathbf{w}(\mathbf{p}, \mathbf{x}) \equiv \nabla_{\mathbf{x}}\pi(\mathbf{p}, \mathbf{x})$  is nonnegative.

<sup>\*53</sup> The assumption that  $S$  is regular implies that  $S$  is a convex set and this in turn implies that  $\pi(\mathbf{p}, \mathbf{x})$  is concave in  $\mathbf{x}$ .

<sup>\*54</sup> The assumption that  $S$  is regular implies that  $S$  is a cone and this in turn implies that  $\pi(\mathbf{p}, \mathbf{x})$  is linearly homogeneous in  $\mathbf{x}$ .

where the equalities in (3.170) follows by differentiating both sides of the Samuelson's Lemma relations,  $\mathbf{w}(\mathbf{p}, \mathbf{x}) \equiv \nabla_x \pi(\mathbf{p}, \mathbf{x})$ , with respect to the components of  $\mathbf{p}$ .

**Theorem 14** Samuelson (1953-54; 10)[343], Diewert (1974a; 144-146)[77]: Suppose that the firm's technology set  $S$  is regular. Define the firm's variable profit function  $\pi(\mathbf{p}, \mathbf{x})$  by (3.115). Suppose that  $\pi(\mathbf{p}, \mathbf{x})$  is twice continuously differentiable with respect to the components of  $\mathbf{x}$  at some point  $(\mathbf{p}, \mathbf{x})$  where  $\mathbf{p} \gg \mathbf{0}_M$  and  $\mathbf{x} \geq \mathbf{0}_N$  and define the matrices of derivatives  $\mathbf{D}$  and  $\mathbf{E}$  by (3.169) and (3.170) respectively. Then these matrices have the following properties:

$$\cdot \mathbf{D} = \mathbf{E}^T \quad \text{so that } \partial y_m(\mathbf{p}, \mathbf{x}) / \partial x_n = \partial w_n(\mathbf{p}, \mathbf{x}) / \partial x_m \text{ for } m = 1, \dots, M \text{ and } n = 1, \dots, N; \quad (3.171)$$

$$\cdot \mathbf{w}(\mathbf{p}, \mathbf{x}) = \mathbf{E}\mathbf{p} \geq \mathbf{0}_N; \quad (3.172)$$

$$\cdot \mathbf{y}(\mathbf{p}, \mathbf{x}) = \mathbf{D}\mathbf{x}. \quad (3.173)$$

**Proof.** The symmetry restrictions (3.171) follow from definitions (3.169) and (3.170) and Young's Theorem in calculus.

Since  $\pi(\mathbf{p}, \mathbf{x})$  is linearly homogeneous in the components of  $\mathbf{p}$ , we have

$$\pi(\lambda\mathbf{p}, \mathbf{x}) = \lambda\pi(\mathbf{p}, \mathbf{x}) \quad \text{for all } \lambda > 0. \quad (3.174)$$

Partially differentiate both sides of (3.174) with respect to  $x_n$  and we obtain:

$$\partial\pi(\lambda\mathbf{p}, \mathbf{x}) / \partial x_n = \lambda\partial\pi(\mathbf{p}, \mathbf{x}) / \partial x_n \quad \text{for all } \lambda > 0 \text{ and } n = 1, \dots, N. \quad (3.175)$$

But (3.175) implies that the functions  $w_n(\mathbf{p}, \mathbf{x}) \equiv \partial\pi(\mathbf{p}, \mathbf{x}) / \partial x_n$  are homogeneous of degree one in  $\mathbf{p}$ . Hence, we can apply Part 1 of Euler's Theorem on homogeneous functions to these functions  $w_n(\mathbf{p}, \mathbf{x})$  and conclude that

$$w_n(\mathbf{p}, \mathbf{x}) = \sum_{m=1}^M [\partial w_n(\mathbf{p}, \mathbf{x}) / \partial p_m] p_m; \quad n = 1, \dots, N. \quad (3.176)$$

But equations (3.176) are equivalent to the equations in (3.172). The inequalities in (3.172) follow from  $\mathbf{w}(\mathbf{p}, \mathbf{x}) \equiv \nabla_x \pi(\mathbf{p}, \mathbf{x}) \geq \mathbf{0}_N$ , which in turn follows from the fact that regularity of  $S$  implies that  $\pi(\mathbf{p}, \mathbf{x})$  is nondecreasing in the components of  $\mathbf{x}$ .

Since  $S$  is regular, part (c) of Theorem 10 implies that  $\pi(\mathbf{p}, \mathbf{x})$  is linearly homogeneous in  $\mathbf{x}$ , so that

$$\pi(\mathbf{p}, \lambda\mathbf{x}) = \lambda\pi(\mathbf{p}, \mathbf{x}) \quad \text{for all } \lambda > 0. \quad (3.177)$$

Partially differentiate both sides of (3.177) with respect to  $p_m$  and we obtain:

$$\partial\pi(\mathbf{p}, \lambda\mathbf{x}) / \partial p_m = \lambda\partial\pi(\mathbf{p}, \mathbf{x}) / \partial p_m \quad \text{for all } \lambda > 0 \text{ and } m = 1, \dots, M. \quad (3.178)$$

But (3.178) implies that the functions  $y_m(\mathbf{p}, \mathbf{x}) \equiv \partial\pi(\mathbf{p}, \mathbf{x}) / \partial p_m$  are homogeneous of degree one in  $\mathbf{x}$ . Hence, we can apply Part 1 of Euler's Theorem on homogeneous functions to these functions  $y_m(\mathbf{p}, \mathbf{x})$  and conclude that

$$y_m(\mathbf{p}, \mathbf{x}) = \sum_{n=1}^N [\partial y_m(\mathbf{p}, \mathbf{x}) / \partial x_n] x_n; \quad m = 1, \dots, M. \quad (3.179)$$

But equations (3.179) are equivalent to equations (3.173). ■

**Problem 20** Under the hypotheses of Theorem 14, show that  $\mathbf{y}(\mathbf{p}, \mathbf{x})$  and  $\mathbf{w}(\mathbf{p}, \mathbf{x})$  satisfy the following equation:

$$\mathbf{p}^T \mathbf{y}(\mathbf{p}, \mathbf{x}) = \mathbf{x}^T \mathbf{w}(\mathbf{p}, \mathbf{x}). \quad (i)$$

**Problem 21** Let  $S$  be a technology set that satisfies assumptions (3.116) and (3.117) and let  $\pi(\mathbf{p}, \mathbf{x})$  be the corresponding differentiable variable profit function defined by (3.115). Variable commodities  $m$  and  $k$  (where  $m \neq k$ ) are said to be *substitutes* if (i) below holds, *unrelated* if (ii) below holds and *complements* if (iii) below holds:

$$\partial y_m(\mathbf{p}, \mathbf{x}) / \partial p_k < 0; \quad (\text{i})$$

$$\partial y_m(\mathbf{p}, \mathbf{x}) / \partial p_k = 0; \quad (\text{ii})$$

$$\partial y_m(\mathbf{p}, \mathbf{x}) / \partial p_k > 0. \quad (\text{iii})$$

(a) If the number of variable commodities  $M = 2$ , then show that the two variable commodities cannot be complements.

(b) If  $M = 2$  and the two variable commodities are unrelated, then show that:

$$\partial y_1(\mathbf{p}, \mathbf{x}) / \partial p_1 = \partial y_2(\mathbf{p}, \mathbf{x}) / \partial p_2 = 0. \quad (\text{iv})$$

(c) If  $M = 3$ , then show that at most one pair of variable commodities can be complements.\*<sup>55</sup>

**Problem 22** Let  $S$  be a regular technology and let  $\pi(\mathbf{p}, \mathbf{x})$  be the corresponding differentiable variable profit function. Define the producer's system of inverse fixed input demand functions as  $\mathbf{w}(\mathbf{p}, \mathbf{x}) \equiv \nabla_{\mathbf{x}} \pi(\mathbf{p}, \mathbf{x})$ . Fixed inputs  $n$  and  $k$  (where  $n \neq k$ ) are said to be *substitutes* if (i) below holds, *unrelated* if (ii) below holds and *complements* if (iii) below holds:

$$\partial w_n(\mathbf{p}, \mathbf{x}) / \partial x_k > 0; \quad (\text{i})$$

$$\partial w_n(\mathbf{p}, \mathbf{x}) / \partial x_k = 0; \quad (\text{ii})$$

$$\partial w_n(\mathbf{p}, \mathbf{x}) / \partial x_k < 0. \quad (\text{iii})$$

(a) If the number of fixed inputs  $N = 2$ , then, assuming that  $x_1 > 0$  and  $x_2 > 0$ , show that the two fixed inputs cannot be complements.

(b) If  $N = 2$  and the two fixed inputs are unrelated, then show that (assume  $x_1 > 0$  and  $x_2 > 0$ ):

$$\partial w_1(\mathbf{p}, \mathbf{x}) / \partial x_1 = \partial w_2(\mathbf{p}, \mathbf{x}) / \partial x_2 = 0. \quad (\text{iv})$$

(c) If  $N = 3$ , then show that at most one pair of fixed inputs can be complements.

**Problem 23** *Application to International Trade Theory.* Suppose that the technology set of a small open economy can be represented by a *regular production possibilities set*,  $S \equiv \{(\mathbf{y}, \mathbf{x})\}$  where  $\mathbf{y}$  is a vector of internationally traded goods (the components of  $C + I + G + X - M$  where imported commodities have negative signs) and  $\mathbf{x} \geq \mathbf{0}_N$  is a nonnegative vector of input factors that are available for use by the aggregate production sector. Let  $\mathbf{p} \gg \mathbf{0}_M$  be a vector of international prices for traded goods that the economy faces. Thus in this case,

$$\pi(\mathbf{p}, \mathbf{x}) \equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : (\mathbf{y}, \mathbf{x}) \in S\} \quad (\text{i})$$

is the *economy's GDP function*,\*<sup>56</sup> regarded as a function of the vector of world prices  $\mathbf{p}$  that the economy faces and of the factor endowment vector or vector of primary resources  $\mathbf{x}$  that the economy

\*<sup>55</sup> This type of argument (that substitutability tends to be more predominant than complementarity) is again due to Hicks (1946; 322-323)[222] but we have not followed his terminology exactly.

\*<sup>56</sup> For applications of duality theory to the theory of international trade, see Samuelson (1953-54)[343], Chipman (1972)[51], Diewert (1974a; 142-146)[77], Diewert and Woodland (1977)[158], Kohli (1978)[272] (1991)[275] and Woodland (1982)[405].

has available to produce goods and services. Assume that  $\pi(\mathbf{p}, \mathbf{x})$  is twice continuously differentiable with respect to its variables at an initial equilibrium for the economy.

(a) Show that if the amount of the first primary input,  $x_1$ , increases a small amount, then GDP does not decrease; i.e., show that

$$\partial\pi(\mathbf{p}, \mathbf{x})/\partial x_1 \geq 0. \quad (\text{ii})$$

(b) Show that as the amount of the first primary input increases a small amount, then the corresponding factor price does not increase and an input quantity weighted sum of the other factor prices does not decrease; i.e., show that

$$\partial w_1(\mathbf{p}, \mathbf{x})/\partial x_1 \leq 0 \text{ and} \quad (\text{iii})$$

$$\sum_{n=2}^N x_n \partial w_n(\mathbf{p}, \mathbf{x})/\partial x_1 \geq 0. \quad (\text{iv})$$

If the inequalities (iii) and (iv) hold strictly, then they show that input 1 experiences a decrease in its price as the amount of input 1 increases but at least one other input must gain as a result of this increase in input 1.

(c) Show that if the price of the first internationally traded good,  $p_1$ , increases a small amount and the first traded good is not imported,<sup>\*57</sup> then GDP increases; i.e., show that

$$\partial\pi(\mathbf{p}, \mathbf{x})/\partial p_1 > 0. \quad (\text{v})$$

(d) Continuation of (c). Show that as the first traded commodity price increases a small amount, then the production of commodity 1 does not decrease and a traded commodity price weighted sum of the other components of GDP does not increase; i.e., show that

$$\partial y_1(\mathbf{p}, \mathbf{x})/\partial p_1 \geq 0 \text{ and} \quad (\text{vi})$$

$$\sum_{m=2}^M p_m \partial y_m(\mathbf{p}, \mathbf{x})/\partial p_1 \leq 0. \quad (\text{vii})$$

If the inequalities (vi) and (vii) hold strictly, then they show that output 1 experiences an increase in production as the price of output 1 increases but at least one other output must decrease (or at least one other imported commodity must increase in magnitude) as a result of this increase in the price of output 1.

(e) Show that if the price of the first internationally traded good,  $p_1$ , increases a small amount and the first traded good is imported, then GDP decreases; i.e., show that

$$\partial\pi(\mathbf{p}, \mathbf{x})/\partial p_1 < 0. \quad (\text{viii})$$

(f) Continuation of (e). Show that as the first traded commodity price increases a small amount, then the importation of commodity 1 does not increase and a traded commodity price weighted sum of the other components of GDP does not increase; i.e., show that<sup>\*58</sup>

$$-\partial y_1(\mathbf{p}, \mathbf{x})/\partial p_1 \leq 0 \text{ and} \quad (\text{ix})$$

$$\sum_{m=2}^M p_m \partial y_m(\mathbf{p}, \mathbf{x})/\partial p_1 \leq 0. \quad (\text{x})$$

If the inequalities (ix) and (x) hold strictly, then they show that imports of traded commodity 1 decline as the price of output 1 increases and in addition, at least one output must decrease (or at least one other imported commodity must increase in magnitude) as a result of this increase in the price of output 1.

<sup>\*57</sup> In fact, we assume that in the initial equilibrium, a positive amount of this first traded commodity is produced by the aggregate production sector.

<sup>\*58</sup> Since  $y_1$  is imported in the initial equilibrium,  $y_1(\mathbf{p}, \mathbf{x}) < 0$ . Thus  $-y_1(\mathbf{p}, \mathbf{x}) > 0$  is the magnitude of imports in the initial equilibrium.

**Problem 24** Let  $S$  be a regular technology set and let  $\pi(\mathbf{p}, \mathbf{x})$  be the corresponding twice continuously differentiable variable profit function defined by (3.115). Variable commodities  $m$  and fixed input  $n$  are said to be *normal* if (i) below holds, *unrelated* if (ii) below holds and *inferior* if (iii) below holds (we assume  $\mathbf{p} \gg \mathbf{0}_M$  and  $\mathbf{x} \gg \mathbf{0}_N$ ):

$$\frac{\partial y_m(\mathbf{p}, \mathbf{x})}{\partial x_n} = \frac{\partial w_n(\mathbf{p}, \mathbf{x})}{\partial p_m} > 0; \tag{i}$$

$$\frac{\partial y_m(\mathbf{p}, \mathbf{x})}{\partial x_n} = \frac{\partial w_n(\mathbf{p}, \mathbf{x})}{\partial p_m} = 0; \tag{ii}$$

$$\frac{\partial y_m(\mathbf{p}, \mathbf{x})}{\partial x_n} = \frac{\partial w_n(\mathbf{p}, \mathbf{x})}{\partial p_m} < 0. \tag{iii}$$

- (a) If  $w_n(\mathbf{p}, \mathbf{x}) > 0$ , then there exists at least one variable commodity  $m$  such that commodity  $m$  and fixed input  $n$  are normal.
- (b) If  $w_n(\mathbf{p}, \mathbf{x}) \geq 0$ , then there exists at least one variable commodity  $m$  such that commodity  $m$  and fixed input  $n$  are either normal or unrelated.
- (c) If  $y_m(\mathbf{p}, \mathbf{x}) > 0$ , then there exists at least one fixed input  $n$  such that commodity  $m$  and fixed input  $n$  are normal.
- (d) If  $y_m(\mathbf{p}, \mathbf{x}) < 0$ , then there exists at least one fixed input  $n$  such that commodity  $m$  and fixed input  $n$  are inferior.

We illustrate the concepts of normality and inferiority for the case of two variable outputs,  $y_1$  and  $y_2$ , and a varying amount of the first fixed input  $x_1$ .<sup>\*59</sup> In Figure 3.8 below, the case where variable commodities 1 and 2,  $y_1$  and  $y_2$ , are both normal with respect to the first fixed input  $x_1$  is illustrated. In this figure, it can be seen that as  $x_1$  increases from its initial level of  $x_1^0$  to the greater level  $x_1^1$ , the production possibilities set  $\{(y_1, y_2) : (y_1, y_2, x_1^0) \in S\}$  shifts outwards to the production possibilities set  $\{(y_1, y_2) : (y_1, y_2, x_1^1) \in S\}$ . The initial variable profit maximizing  $(y_1, y_2)$  point is at point A. After  $x_1$  increases from  $x_1^0$  to  $x_1^1$ , the new variable profit maximizing  $(y_1, y_2)$  point is at point B.

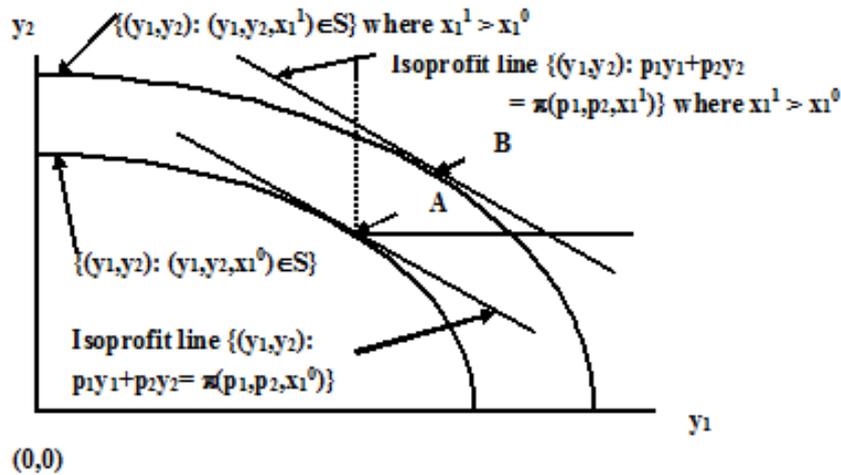


Fig. 3.8 Variable Commodities 1 and 2 Normal with Fixed Input 1

It can be seen as long as the point B is to the northeast of the point A, both  $y_1$  and  $y_2$  will be normal with the fixed input  $x_1$ .<sup>\*60</sup>

<sup>\*59</sup> Since the amounts of the other fixed inputs,  $x_2, \dots, x_N$  remain fixed in the figures to follow, they will be suppressed from the notation.

<sup>\*60</sup> The similarity of this normality concept in production theory with the corresponding normality concept in consumer theory can be seen by looking at Figures 3.8 and 3.9; the fixed input  $x_1$  now plays the role of utility

In Figure 3.9, as  $x_1$  increases from the initial level  $x_1^0$  to the higher level  $x_1^1$ , the revenue maximizing  $(y_1, y_2)$  point moves from A to B. It can be seen that  $y_1$  decreases as  $x_1$  increases and thus the first variable output and the first fixed input are an *inferior* pair of commodities. On the other hand,  $y_2$  increases as  $x_1$  increases and thus the second variable output and the first fixed input are a *normal* pair of commodities.

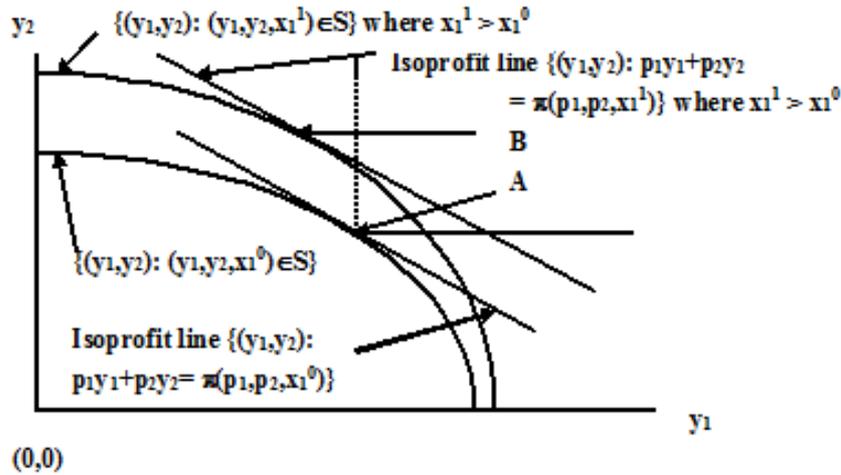


Fig. 3.9 Variable Commodity 1 Inferior with Respect to Fixed Input 1

**Problem 25** Draw counterparts to Figures 3.8 and 3.9 to illustrate the concepts of normality and inferiority for the case where  $y_1 < 0$  is a variable input and  $y_2 > 0$  is a variable output.

### 3.11 Flexible Functional Forms for a Variable Profit Function

We consider functional forms for a variable profit function,  $\pi(\mathbf{p}, \mathbf{x})$ , that are twice continuously differentiable in some region around the point  $\mathbf{p}^* \gg \mathbf{0}_M$  and  $\mathbf{x}^* \gg \mathbf{0}_N$ . We say that  $\pi$  is a flexible functional form if it has a sufficient number of free parameters so that the following equations can be satisfied:

$$\pi(\mathbf{p}^*, \mathbf{x}^*) = \pi^*(\mathbf{p}^*, \mathbf{x}^*); \quad (1 \text{ equation}) \quad (3.180)$$

$$\nabla_{\mathbf{p}} \pi(\mathbf{p}^*, \mathbf{x}^*) = \nabla_{\mathbf{p}} \pi^*(\mathbf{p}^*, \mathbf{x}^*); \quad (M \text{ equations}) \quad (3.181)$$

$$\nabla_{\mathbf{x}} \pi(\mathbf{p}^*, \mathbf{x}^*) = \nabla_{\mathbf{x}} \pi^*(\mathbf{p}^*, \mathbf{x}^*); \quad (N \text{ equations}) \quad (3.182)$$

$$\nabla_{\mathbf{p}\mathbf{p}}^2 \pi(\mathbf{p}^*, \mathbf{x}^*) = \nabla_{\mathbf{p}\mathbf{p}}^2 \pi^*(\mathbf{p}^*, \mathbf{x}^*); \quad (M(M + 1)/2 \text{ equations}) \quad (3.183)$$

$$\nabla_{\mathbf{x}\mathbf{x}}^2 \pi(\mathbf{p}^*, \mathbf{x}^*) = \nabla_{\mathbf{x}\mathbf{x}}^2 \pi^*(\mathbf{p}^*, \mathbf{x}^*); \quad (N(N + 1)/2 \text{ equations}) \quad (3.184)$$

$$\nabla_{\mathbf{p}\mathbf{x}}^2 \pi(\mathbf{p}^*, \mathbf{x}^*) = \nabla_{\mathbf{p}\mathbf{x}}^2 \pi^*(\mathbf{p}^*, \mathbf{x}^*); \quad (MN \text{ equations}) \quad (3.185)$$

where  $\pi^*(\mathbf{p}, \mathbf{x})$  is an arbitrary twice continuously differentiable (at the point  $\mathbf{p}^*, \mathbf{x}^*$ ) variable profit function. There are actually  $M^2$  separate equations in the matrix equation (3.183) and  $N^2$  separate equations in the matrix equation (3.184) but since we are assuming that both  $\pi$  and  $\pi^*$  are twice

---

$u$  and the frontiers of the production possibility sets  $\{(y_1, y_2) : (y_1, y_2, x_1) \in S\}$  indexed by  $x_1$  now replace the indifference curves  $\{(x_1, x_2) : F(x_1, x_2) = u\}$  indexed by  $u$ .

continuously differentiable at the point  $(p^*, \mathbf{x}^*)$ , Young's Theorem from calculus implies that the matrices in (3.183) and (3.184) are symmetric and hence there are only  $M(M+1)/2$  independent equations in (3.183) that need to be satisfied and only  $N(N+1)/2$  equations in (3.184) that need to be satisfied.\*61

Both  $\pi$  and  $\pi^*$  must be linearly homogeneous in the components of the price vector  $\mathbf{p}$  and thus by applying Part 1 of Euler's Theorem on Homogeneous Functions,  $\pi$  and  $\pi^*$  must satisfy the following equations:

$$\pi(p^*, \mathbf{x}^*) = \mathbf{p}^{*T} \nabla_p \pi(p^*, \mathbf{x}^*); \quad \pi^*(p^*, \mathbf{x}^*) = \mathbf{p}^{*T} \nabla_p \pi^*(p^*, \mathbf{x}^*). \quad (3.186)$$

Thus if equations (3.181) are satisfied, then equation (3.180) must be satisfied and the number of free parameters that are required for a flexible functional form can be reduced by 1. Similarly, we can apply Part 2 of Euler's Theorem on Homogeneous Functions and deduce that  $\pi$  and  $\pi^*$  must satisfy the following equations:

$$\nabla_{pp}^2 \pi(p^*, \mathbf{x}^*) p^* = \mathbf{0}_M; \quad \nabla_{pp}^2 \pi^*(p^*, \mathbf{x}^*) p^* = \mathbf{0}_M. \quad (3.187)$$

Thus if the  $M(M-1)/2$  equations in the upper triangle of equations (3.183) are satisfied, then by Young's Theorem, the  $M(M-1)/2$  equations in the lower triangle of equations (3.183) will also be satisfied and then equations (3.187) will imply that the  $M$  equations on the main diagonal of equations (3.183) will also be satisfied. Thus we need only satisfy the  $M(M-1)/2$  equations in the upper triangle of equations (3.183) in order to satisfy all  $M^2$  equations in the matrix equation (3.183).

Since  $\pi(\mathbf{p}, \mathbf{x})$  is linearly homogeneous in the components of  $\mathbf{p}$ , if we partially differentiate the equation  $\pi(\lambda \mathbf{p}, \mathbf{x}) = \lambda \pi(\mathbf{p}, \mathbf{x})$  with respect to each  $x_n$ , we obtain the following equations:

$$\frac{\partial \pi(\lambda \mathbf{p}, \mathbf{x})}{\partial x_n} = \frac{\lambda \partial \pi(\mathbf{p}, \mathbf{x})}{\partial x_n} \quad \text{for all } \lambda > 0 \text{ for } n = 1, \dots, N. \quad (3.188)$$

Equations (3.188) tell us that the  $N$  partial derivative functions,  $\partial \pi(\mathbf{p}, \mathbf{x}) / \partial x_n$  for  $n = 1, \dots, N$ , are linearly homogeneous in their  $\mathbf{p}$  components. Hence, we can apply Part 1 of Euler's Theorem on homogeneous functions to these functions and deduce that the following equations must hold for both  $\partial \pi(\mathbf{p}, \mathbf{x}) / \partial x_n$  (and  $\partial \pi^*(\mathbf{p}, \mathbf{x}) / \partial x_n$  as well since these functions are also linearly homogeneous in the components of  $\mathbf{p}$ ):

$$\frac{\partial \pi(\mathbf{p}, \mathbf{x})}{\partial x_n} = \sum_{m=1}^M p_m \frac{\partial^2 \pi(\mathbf{p}, \mathbf{x})}{\partial x_n \partial p_m}; \quad n = 1, \dots, N; \quad (3.189)$$

$$\frac{\partial \pi^*(\mathbf{p}, \mathbf{x})}{\partial x_n} = \sum_{m=1}^M p_m \frac{\partial^2 \pi^*(\mathbf{p}, \mathbf{x})}{\partial x_n \partial p_m}; \quad n = 1, \dots, N. \quad (3.190)$$

Evaluating equations (3.189) and (3.190) at  $(p^*, \mathbf{x}^*)$  leads to the following matrix equations:

$$\nabla_x \pi(p^*, \mathbf{x}^*) = \nabla_{xp}^2 \pi(p^*, \mathbf{x}^*) p^*; \quad \nabla_x \pi^*(p^*, \mathbf{x}^*) = \nabla_{xp}^2 \pi^*(p^*, \mathbf{x}^*) p^*. \quad (3.191)$$

Thus if equations (3.185) hold, then equations (3.191) imply that equations (3.182) will automatically hold and thus we do not need extra free parameters for  $\pi(\mathbf{p}, \mathbf{x})$  in order to satisfy equations (3.182). Summarizing the above material, in order for  $\pi$  to be a flexible functional form, it will be necessary for  $\pi$  to satisfy the  $M$  equations (3.181), the  $M(M-1)/2$  equations in the upper (or lower) triangle in equations (3.183), the  $N(N+1)/2$  equations in the upper triangle and main diagonal of equations (3.184) and the  $MN$  equations in (3.185). Thus a flexible functional form for a profit function will

\*61 Another application of Young's Theorem implies that if equations (3.185) are satisfied, then the following equations (where we have changed the order of differentiation) will also be satisfied:  $\nabla_{px}^2 \pi(p^*, \mathbf{x}^*) = \nabla_{xp}^2 \pi(p^*, \mathbf{x}^*)$ .

require at least  $M + M(M - 1)/2 + N(N + 1)/2 + MN = M(M + 1)/2 + N(N + 1)/2 + MN$  free parameters in order to satisfy equations (3.180)-(3.185).

Consider the following *Generalized Leontief functional form* for a variable profit function  $\pi$ :

$$\begin{aligned} \pi(\mathbf{p}, \mathbf{x}) \equiv & \sum_{m=1}^M \sum_{k=1}^M a_{mk} p_m^{1/2} p_k^{1/2} [\sum_{n=1}^N \alpha_n x_n] + \sum_{m=1}^M \sum_{n=1}^N b_{mn} p_m x_n \\ & + \sum_{n=1}^N \sum_{i=1}^N c_{ni} x_n^{1/2} x_i^{1/2} [\sum_{m=1}^M \beta_m p_m] + \sum_{m=1}^M d_m p_m + (1/2) \sum_{n=1}^N e_n x_n^2 [\sum_{m=1}^M \beta_m p_m] \end{aligned} \quad (3.192)$$

where the  $a_{mk}$ ,  $b_{mn}$ ,  $c_{ni}$ ,  $d_m$ ,  $e_n$ ,  $\alpha_n$  and  $\beta_m$  are parameters. We impose the following restrictions on the  $a_{mk}$  and  $c_{ni}$  parameters:<sup>\*62</sup>

$$a_{mm} = 0 \text{ for } m = 1, \dots, M; \quad a_{mk} = a_{km} \text{ for } 1 \leq m < k \leq M; \quad (3.193)$$

$$c_{nn} = 0 \text{ for } n = 1, \dots, N; \quad c_{ni} = c_{in} \text{ for } 1 \leq n < i \leq N. \quad (3.194)$$

Thus the  $M \times M$  matrix of parameters  $\mathbf{A} \equiv [a_{mk}]$  is symmetric and has 0 entries down its main diagonal and the  $N \times N$  matrix of parameters  $\mathbf{C} \equiv [c_{ni}]$  is also symmetric and has 0 entries down its main diagonal. In what follows, we will establish the flexibility of the above functional form defined by (3.192)-(3.194) using just the  $a_{mk}$ ,  $b_{mn}$ ,  $c_{ni}$ ,  $d_m$  and  $e_n$  as the free parameters; i.e., we will assume that the  $\alpha_n$  and  $\beta_m$  are predetermined positive numbers.<sup>\*63</sup> Thus there are  $M(M - 1)/2$  free  $a_{mk}$  parameters in the  $\mathbf{A}$  matrix,  $MN$  free parameters  $b_{mn}$  in the  $\mathbf{B} \equiv [b_{mn}]$  matrix,  $N(N - 1)/2$  free  $c_{ni}$  parameters in the  $\mathbf{C}$  matrix,  $M$  free  $d_m$  parameters and  $N$  free  $e_n$  parameters in the above functional form, which is just the minimal number required for a flexible functional form.

Upon partially differentiating the  $\pi$  defined by (3.192)-(3.194) with respect to  $p_m$  for  $m = 1, \dots, M$  and evaluating the resulting partial derivatives at  $(p^*, \mathbf{x}^*)$ , we find that the  $M$  equations in (3.181) become the following equations:<sup>\*64</sup>

$$\begin{aligned} & \sum_{k=1}^M a_{mk} [p_k^*/p_m^*]^{1/2} \boldsymbol{\alpha}^T \mathbf{x}^* + \sum_{n=1}^N b_{mn} x_n^* + \sum_{n=1}^N \sum_{i=1}^N c_{ni} [x_n^* x_i^*]^{1/2} \beta_m \\ & + d_m + (1/2) \sum_{n=1}^N e_n x_n^{*2} \beta_m = \partial \pi^*(p^*, \mathbf{x}^*) / \partial p_m; \quad m = 1, \dots, M. \end{aligned} \quad (3.195)$$

Upon partially differentiating the  $\pi$  defined by (3.192)-(3.194) with respect to  $x_n$  for  $n = 1, \dots, N$  and evaluating the resulting partial derivatives at  $(p^*, \mathbf{x}^*)$ , we find that the  $N$  equations in (3.182) become the following equations:

$$\begin{aligned} & \sum_{m=1}^M \sum_{k=1}^M a_{mk} [p_m^* p_k^*]^{1/2} \alpha_n + \sum_{m=1}^M b_{mn} p_m^* + \sum_{i=1}^N c_{ni} [x_i^*/x_n^*]^{1/2} \boldsymbol{\beta}^T \mathbf{p}^* \\ & + e_n x_n^* \boldsymbol{\beta}^T \mathbf{p}^* = \partial \pi^*(p^*, \mathbf{x}^*) / \partial x_n; \quad n = 1, \dots, N. \end{aligned} \quad (3.196)$$

Now partially differentiate the  $\pi$  defined by (3.192)-(3.194) with respect to  $p_m$  and then  $p_k$  for  $m < k$ . Evaluating the resulting partial derivatives at  $(p^*, \mathbf{x}^*)$ , we find that the  $M(M - 1)/2$  equations in the upper triangle of the matrix equation (3.183) become the following equations:

$$(1/2) a_{mk} (p_m^* p_k^*)^{-1/2} \boldsymbol{\alpha}^T \mathbf{x}^* = \partial^2 \pi^*(p^*, \mathbf{x}^*) / \partial p_m \partial p_k; \quad 1 \leq m < k \leq M. \quad (3.197)$$

It can be seen that the  $M(M - 1)/2$  equations in (3.197) determine the  $a_{mk}$  for  $1 \leq m < k \leq M$ .

<sup>\*62</sup> These restrictions on the parameters are imposed in order to be able to identify all of the unknown parameters in the functional form in econometric applications; i.e., there is a linear dependence in the independent variables associated with the  $a_{mm}$ ,  $c_{nn}$  and  $b_{mn}$  parameters.

<sup>\*63</sup> In econometric applications, we generally set each  $\alpha_n$  and  $\beta_m$  equal to 1. Another popular choice is to set  $\alpha_n$  equal to the sample average of the observed prices for variable output  $m$   $p_m^t$  and set  $\beta_m$  equal to the sample average of the observed "fixed"  $n$ th inputs  $x_n^t$ .

<sup>\*64</sup> For notational convenience, the  $a_{mm}$  and  $c_{nn}$  parameters appear in equations (3.195) but remember, these parameters are set equal to 0. We also write  $\sum_{n=1}^N \alpha_n x_n^*$  as  $\boldsymbol{\alpha}^T \mathbf{x}^*$  and  $\sum_{m=1}^M \beta_m p_m^*$  as  $\boldsymbol{\beta}^T \mathbf{p}^*$ .

Now partially differentiate the  $\pi$  defined by (3.192)-(3.194) with respect to  $x_n$  and then  $x_i$  for  $n < i$ . Evaluating the resulting partial derivatives at  $(p^*, \mathbf{x}^*)$ , we find that the  $N(N-1)/2$  equations in the upper triangle of the matrix equation (3.184) become the following equations:

$$(1/2)c_{ni}(x_n^*x_i^*)^{-1/2}\boldsymbol{\beta}^T\mathbf{p}^* = \partial^2\pi^*(p^*, \mathbf{x}^*)/\partial x_n\partial x_i; \quad 1 \leq n < i \leq N. \quad (3.198)$$

It can be seen that the  $N(N-1)/2$  equations in (3.198) determine the  $c_{ni}$  for  $1 \leq n < i \leq N$ .

Now partially differentiate the  $\pi$  defined by (3.192)-(3.194) with respect to  $x_n$  twice. Evaluating the resulting partial derivatives at  $(p^*, \mathbf{x}^*)$ , we find that the  $N$  equations along the main diagonal of the matrix equation (3.184) become the following equations:

$$-(1/2)\sum_{n=1}^N \sum_{i \neq n} c_{ni}(x_i^*)^{-1/2}(x_n^*)^{-3/2}\boldsymbol{\beta}^T\mathbf{p}^* + e_n\boldsymbol{\beta}^T\mathbf{p}^* = \partial^2\pi^*(p^*, \mathbf{x}^*)/\partial x_n^2; \quad n = 1, \dots, N. \quad (3.199)$$

Since the off diagonal  $c_{ni}$  have already been determined by the symmetry conditions in the restrictions (3.194) and by solving equations (3.198), the  $e_n$  parameters can be determined by solving equations (3.199).

Now partially differentiate the  $\pi$  defined by (3.192)-(3.194) with respect to  $p_m$  and then  $x_n$ . Evaluating the resulting partial derivatives at  $(p^*, \mathbf{x}^*)$ , we find that the  $MN$  equations in the matrix equation (3.185) become the following equations:

$$\begin{aligned} \sum_{k=1}^M a_{mk}[p_k^*/p_m^*]^{1/2}\alpha_n + b_{mn} + \sum_{i=1}^N c_{ni}[x_i^*/x_n^*]^{1/2}\beta_m + e_n x_n^*\beta_m \\ = \partial^2\pi^*(p^*, \mathbf{x}^*)/\partial p_m\partial x_n; \quad m = 1, \dots, M; n = 1, \dots, N. \end{aligned} \quad (3.200)$$

Since the  $a_{mk}$ ,  $c_{ni}$  and  $e_n$  have already been determined, the  $m$ th equation in (3.200) can be used to solve for  $b_{mn}$  for  $m = 1, \dots, M; n = 1, \dots, N$ . Thus all of the unknown parameters in the Generalized Leontief profit function have been determined except for the  $M$   $d_m$  parameters. But now use the  $M$  equations in (3.195) to determine the  $d_m$ . Using the various restrictions that we developed above, it can be seen that the resulting Generalized Leontief profit function will satisfy all of the equations in (3.180)-(3.185) and thus is a flexible functional form.

Now suppose that we have collected data on output and input prices and quantities for  $T$  periods and the data for period  $t$  is denoted by the vectors  $\mathbf{y}^t$ ,  $\mathbf{x}^t$ ,  $\mathbf{p}^t$  and  $\mathbf{w}^t$ . Application of Hotelling's Lemma leads to the following estimating equations for the net outputs  $y_m^t$ :

$$\begin{aligned} y_m^t = \sum_{k=1}^M \sum_{k \neq m} a_{mk}[p_k^t/p_m^t]^{1/2}\boldsymbol{\alpha}^T\mathbf{x}^t + \sum_{n=1}^N b_{mn} + \sum_{n=1}^N \sum_{i=1}^N \sum_{i \neq n} c_{ni}[x_n^t x_i^t]^{1/2}\beta_m \\ + d_m + (1/2)\sum_{n=1}^N e_n(x_n^t)^2\beta_m + \varepsilon_m^t; \quad m = 1, \dots, M; t = 1, \dots, T \end{aligned} \quad (3.201)$$

where the  $\varepsilon_m^t$  are error terms. Note that the unknown parameters appear in a linear fashion on the right hand side of equations (3.201) and thus linear regression techniques can be used in order to facilitate econometric estimation of the unknown parameters.\*65

If it can be assumed that the "fixed" inputs were chosen in a cost minimizing manner, then we can apply Samuelson's Lemma and obtain the following estimating equations for the input prices  $w_n^t$ :

$$\begin{aligned} w_n^t = \sum_{m=1}^M \sum_{k=1}^M \sum_{m \neq k} a_{mk}[p_m^t p_k^t]^{1/2}\alpha_n + \sum_{m=1}^M b_{mn} p_m^t \\ + \sum_{i=1}^N \sum_{i \neq n} c_{ni}[x_i^t/x_n^t]^{1/2}\boldsymbol{\beta}^T\mathbf{p}^t + e_n x_n^t \boldsymbol{\beta}^T\mathbf{p}^t + \eta_n^t; \quad n = 1, \dots, N; t = 1, \dots, T \end{aligned} \quad (3.202)$$

where the  $\eta_n^t$  are error terms. Note that the unknown parameters appear in a linear fashion on the right hand side of equations (3.202) and thus linear regression techniques can be used in order to facilitate econometric estimation of the unknown parameters.\*66

\*65 There are cross equation symmetry conditions on the  $a_{mk}$  parameters which need to be imposed or tested.

\*66 There are cross equation symmetry conditions on the  $c_{ni}$  parameters which need to be imposed or tested and there are additional symmetry conditions on the  $b_{mn}$  between equations which need to be imposed or tested.

Once the unknown parameters have been estimated (denote these estimated parameters by  $a_{mk}^*, b_{mn}^*, c_{ni}^*, d_m^*$  and  $e_n^*$ ), it is necessary to check whether  $\pi(\mathbf{p}, \mathbf{x})$  is locally convex with respect to the components of  $\mathbf{p}$  at the observed data points,  $(\mathbf{p}^t, \mathbf{x}^t)$ , for  $t = 1, \dots, T$ . Thus we need to calculate the matrix of second order partial derivatives,  $\nabla_{pp}^2 \pi(\mathbf{p}^t, \mathbf{x}^t)$ , for  $t = 1, \dots, T$  and check whether the resulting matrix is positive semidefinite (so that the estimated  $\pi(\mathbf{p}, \mathbf{x})$  is at least locally convex in  $\mathbf{p}$  at the observed data points). These second derivatives have the following form:

$$\partial^2 \pi(\mathbf{p}^t, \mathbf{x}^t) / \partial p_m \partial p_k = (1/2) a_{mk}^* (p_m^t p_k^t)^{-1/2} \boldsymbol{\alpha}^T \mathbf{x}^t; \quad m \neq k; \quad (3.203)$$

$$\partial^2 \pi(\mathbf{p}^t, \mathbf{x}^t) / \partial p_m^2 = -(1/2) \sum_{k=1, k \neq m}^M a_{mk}^* [p_k^t]^{1/2} [p_m^t]^{-3/2} \boldsymbol{\alpha}^T \mathbf{x}^t; \quad m = 1, \dots, M. \quad (3.204)$$

Sufficient conditions for global convexity of  $\pi(\mathbf{p}, \mathbf{x})$  in  $\mathbf{p}$  for all  $\mathbf{p} \gg \mathbf{0}_M$  are that the estimated  $a_{mk}^*$  be nonpositive for  $m \neq k$ ; i.e.,

$$a_{mk}^* \leq 0 \text{ for all } m \neq k. \quad (3.205)$$

These sufficient conditions for global convexity can be imposed by replacing the parameters  $a_{mk}$  in (3.192) by  $-[f_{mk}]^2$ ; i.e., by a squaring technique.<sup>\*67</sup> Using equations (3.203), imposing the restrictions (3.205) and using Problem 21 above, it can be seen that all outputs must be substitutes. This is not restrictive if  $M = 2$  but if  $M > 2$ , it is restrictive to impose conditions (3.205). Thus the use of the squaring technique to impose global convexity of  $\pi(\mathbf{p}, \mathbf{x})$  in the components of  $\mathbf{p}$  cannot be recommended if  $M > 2$  since we do not want to restrict a priori elasticities of substitution between outputs. However, if we do not impose restrictions on the  $a_{mk}$ , empirical experience has shown that the convexity of  $\pi(\mathbf{p}, \mathbf{x})$  in the components of  $\mathbf{p}$  will generally fail if  $M$  is equal to or greater than 4 or 5. Thus the Generalized Leontief profit function defined by (3.192)-(3.194) above is not a completely satisfactory flexible functional form for empirical applications.<sup>\*68</sup>

We conclude this section by modifying the above analysis to cover the case where the underlying technology is *regular*; i.e., the technology exhibits constant returns to scale. In this case,  $\pi(\mathbf{p}, \mathbf{x})$  and  $\pi^*(\mathbf{p}, \mathbf{x})$  are linearly homogeneous in the components of  $\mathbf{x}$  (as well as in the components of  $\mathbf{p}$ ). This means that the first and second order partial derivatives of these functions will satisfy additional restrictions at the point  $(p^*, \mathbf{x}^*)$ . We will now derive these extra conditions.

Both  $\pi$  and  $\pi^*$  are linearly homogeneous in the components of the input vector  $\mathbf{x}$  if the technology is regular and thus by applying Part 1 of Euler's Theorem on Homogeneous Functions,  $\pi$  and  $\pi^*$  must satisfy the following equations:

$$\pi(p^*, \mathbf{x}^*) = \mathbf{x}^{*T} \nabla_x \pi(p^*, \mathbf{x}^*); \quad \pi^*(p^*, \mathbf{x}^*) = \mathbf{x}^{*T} \nabla_x \pi^*(p^*, \mathbf{x}^*). \quad (3.206)$$

Similarly, we can apply Part 2 of Euler's Theorem on Homogeneous Functions and deduce that  $\pi$  and  $\pi^*$  must satisfy the following equations:

$$\nabla_{xx}^2 \pi(p^*, \mathbf{x}^*) \mathbf{x}^* = \mathbf{0}_N; \quad \nabla_{xx}^2 \pi^*(p^*, \mathbf{x}^*) \mathbf{x}^* = \mathbf{0}_N. \quad (3.207)$$

Thus if the  $N(N-1)/2$  equations in the upper triangle of equations (3.184) are satisfied, then by Young's Theorem, the  $N(N-1)/2$  equations in the lower triangle of equations (3.184) will also be satisfied and then equations (3.207) will imply that the  $N$  equations on the main diagonal of equations (3.184) will also be satisfied. Thus we need only satisfy the  $N(N-1)/2$  equations in the upper triangle of equations (3.184) in order to satisfy all  $N^2$  equations in the matrix equation (3.184). Thus equations (3.207) reduce the number of free parameters that a flexible functional form

<sup>\*67</sup> The estimating equations (3.201) and (3.202) then become nonlinear in the unknown parameters but the degree of nonlinearity is not too severe and hence nonlinear regression techniques can be successfully employed.

<sup>\*68</sup> In Chapter 9, we will show how a more satisfactory flexible functional form can be defined: one which will always satisfy the appropriate convexity and concavity conditions.

for a regular technology profit function must have by  $N$  as compared to the previous case of a general technology.

Since  $\pi(\mathbf{p}, \mathbf{x})$  is now linearly homogeneous in the components of  $\mathbf{x}$ , if we partially differentiate the equation  $\pi(\mathbf{p}, \lambda \mathbf{x}) = \lambda \pi(\mathbf{p}, \mathbf{x})$  with respect to each  $p_m$ , we obtain the following equations:

$$\partial \pi(\mathbf{p}, \lambda \mathbf{x}) / \partial p_m = \lambda \partial \pi(\mathbf{p}, \mathbf{x}) / \partial p_m \quad \text{for all } \lambda > 0 \text{ for } m = 1, \dots, M. \quad (3.208)$$

Equations (3.208) tell us that the  $M$  partial derivative functions,  $\partial \pi(\mathbf{p}, \mathbf{x}) / \partial p_m$  for  $m = 1, \dots, M$ , are linearly homogeneous in their  $\mathbf{x}$  components. Hence, we can apply Part 1 of Euler's Theorem on homogeneous functions to these functions and deduce that the following equations must hold for both  $\partial \pi(\mathbf{p}, \mathbf{x}) / \partial p_m$  (and  $\partial \pi^*(\mathbf{p}, \mathbf{x}) / \partial p_m$  as well since these functions are also linearly homogeneous in the components of  $\mathbf{x}$ ):

$$\partial \pi(\mathbf{p}, \mathbf{x}) / \partial p_m = \sum_{n=1}^N x_n \partial^2 \pi(\mathbf{p}, \mathbf{x}) / \partial p_m \partial x_n; \quad m = 1, \dots, M; \quad (3.209)$$

$$\partial \pi^*(\mathbf{p}, \mathbf{x}) / \partial p_m = \sum_{n=1}^N x_n \partial^2 \pi^*(\mathbf{p}, \mathbf{x}) / \partial p_m \partial x_n; \quad m = 1, \dots, M. \quad (3.210)$$

Evaluating equations (3.209) and (3.210) at  $(p^*, \mathbf{x}^*)$  leads to the following matrix equations:

$$\nabla_p \pi(p^*, \mathbf{x}^*) = \nabla_{px}^2 \pi(p^*, \mathbf{x}^*) \mathbf{x}^*; \quad \nabla_p \pi^*(p^*, \mathbf{x}^*) = \nabla_{px}^2 \pi^*(p^*, \mathbf{x}^*) \mathbf{x}^*. \quad (3.211)$$

Thus if equations (3.185) hold, then equations (3.211) imply that equations (3.181) will automatically hold and thus we do not need extra free parameters for  $\pi(\mathbf{p}, \mathbf{x})$  in order to satisfy equations (3.181). Thus equations (3.211) reduce the number of free parameters that a flexible functional form for a regular technology profit function must have by  $M$  as compared to the previous case of a general technology.

Summarizing the above material, in order for  $\pi$  to be a flexible functional form for a regular technology, it will be necessary for  $\pi$  to satisfy the  $M(M-1)/2$  equations in the upper (or lower) triangle in equations (3.183), the  $N(N-1)/2$  equations in the upper triangle of equations (3.184) and the  $MN$  equations in (3.185). Thus a flexible functional form for a profit function for a regular technology will require at least  $M(M-1)/2 + N(N-1)/2 + MN$  free parameters in order to satisfy equations (3.180)-(3.185).

Consider the following *Generalized Leontief functional form* for a variable profit function  $\pi$  for a regular technology:

$$\begin{aligned} \pi(\mathbf{p}, \mathbf{x}) \equiv & \sum_{m=1}^M \sum_{k=1}^M a_{mk} p_m^{1/2} p_k^{1/2} [\sum_{n=1}^N \alpha_n x_n] + \sum_{m=1}^M \sum_{n=1}^N b_{mn} p_m x_n \\ & + \sum_{n=1}^N \sum_{i=1}^N c_{ni} x_n^{1/2} x_i^{1/2} [\sum_{m=1}^M \beta_m p_m] \end{aligned} \quad (3.212)$$

where the  $a_{mk}$  and  $c_{ni}$  satisfy the restrictions (3.193) and (3.194). Thus the  $\pi$  defined by (3.212) is a special case of the more general  $\pi$  defined by (3.192) where the previous parameters  $d_m$  and  $e_n$  are now set equal to 0. Note that the new functional form has just the minimal number of parameters ( $M(M-1)/2$   $a_{mk}$ ,  $N(N-1)/2$   $c_{ni}$  and  $MN$   $b_{mn}$ ) required for the functional form to be flexible for a regular technology.

**Problem 26** Prove that the profit function  $\pi(\mathbf{p}, \mathbf{x})$  defined by (3.212) where the  $a_{mk}$  and  $c_{ni}$  satisfy the restrictions (3.193) and (3.194) is a flexible functional form for a regular technology.

*Hint:* You need only satisfy the upper triangle equations in the matrix equations (3.183) and (3.184) and satisfy the matrix equation (3.185).

**Problem 27** (a) Calculate the matrix of second order partial derivatives,  $\nabla_{xx}^2 \pi(\mathbf{p}^t, \mathbf{x}^t)$  for the  $\pi$  defined by (3.212), (3.193) and (3.194); i.e., calculate the analogues to equations (3.203) and (3.204) except now differentiate with respect to  $x_n$  and  $x_i$  instead of  $p_m$  and  $p_k$ .

(b) Should  $\nabla_{xx}^2 \pi(\mathbf{p}^t, \mathbf{x}^t)$  have a definiteness property? If so, what is it?

*Hint:* Remember that we are assuming that the technology is regular.

(c) How can one impose the property that  $\pi(\mathbf{p}, \mathbf{x})$  defined by (3.212) and (3.193) and (3.194) is a concave function in  $\mathbf{x}$  for fixed  $\mathbf{p}$  over the set of  $\mathbf{x}$  such that  $\mathbf{x} \gg \mathbf{0}_N$ ? Is your method of imposing global concavity in  $\mathbf{x}$  for the  $\pi(\mathbf{p}, \mathbf{x})$  defined by (3.212) at all restrictive?

## 3.12 References

Allen, R.G.C. (1938), *Mathematical Analysis for Economists*, London: Macmillan.

Arrow, K.J., H.B. Chenery, B.S. Minhas and R.M. Solow (1961), "Capital-Labor Substitution and Economic Efficiency", *Review of Economic Statistics* 63, 225-250.

Blackorby, C. (1975), "Degrees of Cardinality and Aggregate Partial Orderings", *Econometrica* 43, 845-852.

Blackorby, C. and W.E. Diewert (1979), "Expenditure Functions, Local Duality and Second Order Approximations", *Econometrica* 47, 579-601.

Blackorby, C., D. Primont and R.R. Russell (1978), *Duality, Separability and Functional Structure: Theory and Economic Applications*, New York: North-Holland.

Chipman, J.S. (1966), "A Survey of the Theory of International Trade: Part 3: The Modern Theory", *Econometrica* 34, 18-76.

Chipman, J.S. (1972), "The Theory of Exploitative Trade and Investment Policies: A Reformulation and Synthesis", in *International Economics and Development: Essays in Honor of Raul Prebisch*, L.D. de Marco (ed.), New York: Academic Press.

Cobb, C. and P.H. Douglas (1928), "A Theory of Production", *American Economic Review*, Supplement, 18, 139-165.

Diewert, W.E. (1971), "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function", *Journal of Political Economy* 79, 481-507.

Diewert, W.E. (1973), "Functional Forms for Profit and Transformation Functions", *Journal of Economic Theory* 6, 284-316.

Diewert, W.E. (1974), "Intertemporal Consumer Theory and the Demand for Durables", *Econometrica* 42, 497-516.

Diewert, W.E. (1974a), "Applications of Duality Theory", pp. 106-171 in *Frontiers of Quantitative Economics*, Volume 2, M.D. Intriligator and D.A. Kendrick (eds.), Amsterdam: North-Holland.

Diewert, W.E. (1974c), "Functional Forms for Revenue and Factor Requirements Functions", *International Economic Review* 15, 119-130.

Diewert, W.E. (1978), "Hicks' Aggregation Theorem and the Existence of a Real Value Added Function", pp. 17-51, Vol. 2, in *Production Economics: A Dual Approach to Theory and Applications*, M. Fuss and D. McFadden, editors, North-Holland, Amsterdam.

Diewert, W.E. (1980), "Symmetry Conditions for Market Demand Functions", *Review of Economic Studies* 47, 595-601.

Diewert, W.E. (1982), "Duality Approaches to Microeconomic Theory", pp. 535-599 in *Handbook of Mathematical Economics*, Volume 2, K.J. Arrow and M.D. Intriligator (eds.), Amsterdam: North-Holland.

- Diewert, W.E. (1993), "Duality Approaches to Microeconomic Theory", pp. 105-175 in *Essays in Index Number Theory*, Volume 1, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North-Holland. This paper is a rewrite of Diewert (1982) but it also includes proofs.
- Diewert, W.E. and T.J. Wales (1987), "Flexible Functional Forms and Global Curvature Conditions", *Econometrica* 55, 43-68.
- Diewert, W.E. and T.J. Wales (1992), "Quadratic Spline Models for Producer's Supply and Demand Functions", *International Economic Review* 33, 705-722.
- Diewert, W.E. and A.D. Woodland (1977), "Frank Knight's Theorem in Linear Programming Revisited", *Econometrica* 45, 375-398.
- Fenchel, W. (1953), "Convex Cones, Sets and Functions", Lecture Notes at Princeton University, Department of Mathematics, Princeton, N.J.
- Gale, D, V.L. Klee and R.T. Rockafellar (1968), "Convex Functions on Convex Polytopes", *Proceedings of the American Mathematical Society* 19, 867-873.
- Gorman, W.M. (1968), "Measuring the Quantities of Fixed Factors", pp. 141-172 in *Value, Capital and Growth: Papers in Honour of Sir John Hicks*, J.N. Wolfe (ed.), Chicago: Aldine.
- Hicks, J.R. (1946), *Value and Capital*, Second Edition, Oxford: Clarendon Press.
- Hotelling, H. (1932), "Edgeworth's Taxation Paradox and the Nature of Demand and Supply Functions", *Journal of Political Economy* 40, 577-616.
- Hotelling, H. (1935), "Demand Functions with Limited Budgets", *Econometrica* 3, 66-78.
- Kohli, U.R.J. (1978), "A Gross National Product Function and the Derived Demand for Imports and Supply of Exports", *Canadian Journal of Economics* 11, 167-182.
- Kohli, U. (1991), *Technology, Duality and Foreign Trade: The GNP Function Approach to Modelling Imports and Exports*, Ann Arbor, MI: University of Michigan Press.
- Leontief, W.W. (1941), *The Structure of the American Economy 1919-1929*, Cambridge, MA: Harvard University Press.
- McFadden, D. (1966), "Cost, Revenue and Profit Functions: A Cursory Review", IBER Working Paper No. 86, University of California, Berkeley.
- McFadden, D. (1978), "Cost, Revenue and Profit Functions", pp. 3-109 in *Production Economics: A Dual Approach*, Volume 1, M. Fuss and D. McFadden (eds.), Amsterdam: North-Holland.
- McKenzie, L.W. (1956-7), "Demand Theory without a Utility Index", *Review of Economic Studies* 24, 184-189.
- Pollak, R.A. (1969), "Conditional Demand Functions and Consumption Theory", *Quarterly Journal of Economics* 83, 60-78.
- Rockafellar, R.T. (1970), *Convex Analysis*, Princeton, N.J.: Princeton University Press.
- Samuelson, P.A. (1947), *Foundations of Economic Analysis*, Cambridge, MA: Harvard University Press.
- Samuelson, P.A. (1953-54), "Prices of Factors and Goods in General Equilibrium", *Review of Economic Studies* 21, 1-20.
- Samuelson, P.A. (1974), "Complementarity—An Essay on the 40th Anniversary of the Hicks-Allen Revolution in Demand Theory", *The Journal of Economic Literature* 12, 1255-1289.
- Shephard, R.W. (1953), *Cost and Production Functions*, Princeton N.J.: Princeton University Press.
- Shephard, R.W. (1967), "The Notion of a Production Function", *Unternehmensforschung* 11, 209-232.
- Shephard, R.W. (1970), *Theory of Cost and Production Functions*, Princeton N.J.: Princeton University Press.
- Uzawa, H. (1962), "Production Functions with Constant Elasticities of Substitution", *Review of Economic Studies* 29, 291-299.

- Uzawa, H. (1964), "Duality Principles in the Theory of Cost and Production", *International Economic Review* 5, 291-299.
- Walters, A.A. (1961), "Production and Cost Functions: An Econometric Survey", *Econometrica* 31, 1-66.
- Wold, H. (1944), "A Synthesis of Pure Demand Analysis; Part 3", *Skandinaviske Aktuarietidskrift* 27, 69-120.
- Wold, H. (1953), *Demand Analysis*, New York: John Wiley.
- Woodland, A.D. (1982), *International Trade and Resource Allocation*. Amsterdam: North Holland.



## Chapter 4

# Notes on the Construction of a Data Set for an O.E.C.D. Country

### 4.1 Overview

Our goal is to collect enough data on an OECD country (plus a few other countries that have good sources of data) so that we can construct a small applied general equilibrium model or macro model based on micro theory.

On the producer side, the model will have 5 commodities:

- Domestic output; an aggregate of consumption, government expenditures and investment;
- Exports;
- Imports;
- Labour input;
- Capital input.

Ideally, we would like to have land and natural resource inputs in the model as well but the old system of national accounts did not recognize the contributions of these inputs explicitly and thus historical data on these inputs are not available.

On the consumer side, we will not model the savings decision and so the consumer side model will have only 2 commodities:

- Consumption;
- Leisure demand (or the negative of labour supply).

We will also collect data on the various tax wedges that consumers and producers face.

We will collect data for the years 1960 to the latest year available for most OECD countries. Students can choose to collect data for a non OECD country but typically, it will prove to be difficult to find all of the necessary data (the student will have to use the data generated by the relevant national statistical agency). However, the Asian Productivity Organization has recently developed some very usable data sets for 6 non OECD Asian countries so these data can also be used. The data for recent joined OECD countries has typically not been pushed back to 1960. The list of OECD countries for which the data are available back to 1960 except where noted are as follows (I have excluded Canada since I will use Canada as an example in the files that I send you):

1. Australia;
2. Austria;
3. Belgium;
4. Denmark;

5. Finland;
6. France;
7. Germany (some complications here after the two Germanys unified in 1991);
8. Greece;
9. Iceland;
10. Ireland;
11. Italy;
12. Japan;
13. Korea (1970);
14. Mexico (1970);
15. Netherlands;
16. New Zealand;
17. Norway;
18. Portugal;
19. Spain;
20. Sweden;
21. Switzerland;
22. Turkey;
23. United Kingdom;
24. USA.

The OECD also has data for the following recently admitted countries to the OECD:

25. Chile (data go back to 1996);
26. Czech Republic (1990);
27. Hungary (1995);
28. Poland (1991);
29. Slovak Republic (1993);
30. Estonia (1995);
31. Israel (1995 but I believe that we can go back much further than this);
32. Slovenia (1995).

The OECD is also publishing data on some important non OECD countries (see OECD.STAT):

33. Brazil (data go back to 1990);
34. China (data back to 1970);
35. India (data back to 1997);
36. Indonesia (data back to 1990);
37. Russia (data back to 1995);
38. South Africa (data back to 1970).

The Asian Productivity Organization has usable data back to 1970 for China, Indonesia and India and the following three developing countries:

39. Republic of China (Taiwan);
40. Philippines;
41. Thailand.

The first 24 countries are lower risk in the sense that students in this class have successfully found enough data to the various country estimations that we will be doing later in the course.

The Asian Productivity Organization has just published the *APO Productivity Databook 2010*, Tokyo: The Asian Productivity Organization, which lists basic national accounts output data back to 1970 for the following countries: Bangladesh, Cambodia, ROC (Taiwan), Fiji, Hong Kong, India, Indonesia, Iran, Japan, Korea, Lao PDR, Malaysia, Mongolia, Nepal, Pakistan, Philippines, Singapore, Sri Lanka, Thailand, Vietnam, China and the US. They also list employment data. The main

data that are missing have to do with commodity taxes and labour input but this information may be available from the UN national accounts or national statistical agency information.\*<sup>1</sup>

## 4.2 Basic National Accounts Data

**Source 1:** *National Accounts; Main Aggregates; 1960-1997*; Volume 1, 1999 Edition, Organisation for Economic Co-Operation and Development; Paris, 1999. Call Number: HC 79 I5 O751; *Location: Koerner Level 1 Stacks*. This is a thin volume. The most recent volume will be available at the Reference section of the Koerner Library, Level 2. The data for the years 1970 to 2009 are also available from the OECD online; see OECD.Stat. This is actually the most convenient source but the data does not always go back to 1960; i.e., it often starts at 1970 (or later for newer OECD countries).

Copy the data for your country that have the title: **Main Aggregates** or for the more recent publications, **Gross Domestic Product: Expenditure Approach**. We will be using the following series:

At current prices:

- Government final consumption expenditure;
- Private final consumption expenditure;\*<sup>2</sup>
- Increase in stocks;
- Gross fixed capital formation;
- Exports of goods and services;
- Imports of goods and services.

At constant price levels:

- Government final consumption expenditure;
- Private final consumption expenditure;
- Increase in stocks;
- Gross fixed capital formation;
- Exports of goods and services;
- Imports of goods and services.

In addition, we will use the following current dollar series that are listed under **Cost Components of the GDP** or in the more recent publications, listed as **Gross Domestic Product: Income Approach**:

- Indirect taxes;
- Subsidies;
- Consumption of fixed capital (depreciation);
- Compensation of employees;
- Operating surplus.

For each country, the above series are available for the years 1960 and 1969-1997 from the 1999 publication *National Accounts; Main Aggregates; 1960-1997*; Volume 1, 1999 Edition. Here is our first problem: how do we get data for the missing years? For the expenditure components of the GDP,  $(C + G + I + X - M)$ , we can fill in for the missing years using some additional data series that are tabled in the same publication on pages 150-157. Tables 26-30 on these pages give us volume

\*<sup>1</sup> Students have successfully chosen Hong Kong, China and Taiwan as their country in the past.

\*<sup>2</sup> The System of National Accounts was revised in 1993 and so for the more recent data, Private final consumption expenditure has been replaced by two categories: (1) Household final consumption expenditure and (2) Final consumption expenditure of NPISH's (Non Profit Institutions Serving Households), such as trade unions, professional societies, political parties, religious organizations, sports, cultural or recreational clubs and charities.

indexes for the 5 output components of GDP for each of the “old-time” OECD countries; these are essentially the constant dollar series divided by a constant. Tables 32-36 give us price indexes for these components of GDP. With the help of these series, we can fill in the data for the missing years, 1961-1968. Thus, please copy pages 150-157 in addition to the earlier pages that list the data for your specific country.

**Source 2:** *National Accounts; Main Aggregates; 1960-1989*, Volume 1, Organisation for Economic Co-Operation and Development; Paris, 1991. Call Number: HC 79 I5 O751; *Location: Koerner Level 1 Stacks.*

This volume will enable you to fill in the Cost components of the GDP (indirect taxes, subsidies, depreciation, compensation of employees and operating surplus) for the missing years, 1961-1968.

Here is where another problem can arise: namely, the data in this publication may not agree with the data in the Source 1 publication for the years 1960 and 1969. If the differences in your data are small for these two years in the two sources, just ignore the differences. However, if the differences are “large”, try and adjust the data from the second source to blend in with the data from the first source. We assume that the data in a later publication is more reliable than the data from an earlier publication!

If you cannot find this volume, you can use other volumes of the same publication (which can be found at the Level 1 Stacks of the Koerner Library) to obtain information on the cost components of GDP for the missing years 1961-1968;

**Source 3:** *National Accounts of OECD Countries; 1991-2002*; Volume 1, *Main Aggregates* or Volume IIa, *Detailed Tables*, Organisation for Economic Co-Operation and Development; Paris, 2004. Call Number: HC 79 I5 O751; *Location: Koerner Level 2 Reference.*

These two publications are the most recent publications for the OECD National Accounts. Information for the years 1991-2007 can be obtained from this publication. In order to obtain information for the most recent years, try to find the appropriate National Statistical Agency online and check whether the information for the last few years is available. Alternatively, almost everything from 1970 on is available online at OECD.Stat.

### 4.3 Labour and Population Statistics

The system of national accounts provides current and constant dollar estimates for the output components of a country (alternatively, current dollar estimates are provided along with price indexes or deflators for the current dollar estimates). However, until recently, constant dollar components for the inputs into production have not been routinely provided in the system of national accounts. In this section, we will focus on the problems involved in finding price indexes for labour input or finding estimates of constant dollar labour input.

From the tables on the Cost Components of GDP (see sources 1 and 2 above), we can obtain estimates of the annual compensation of employees; i.e., on the value of labour input for employees for each year. However, these current dollar values do *not* include the value of labour inputs from:

- unpaid family workers and
- self employed workers (or employers and persons working on their own account).

In recent years, for most OECD countries, the omission of unpaid family workers is not a big problem but for some countries, in the 1960’s and 1970’s, there was a considerable amount of unpaid family work (particularly in agriculture) and so we do have to make an adjustment to our data to allow for the input of these workers. With respect to the contribution of the self employed, during the past decade, the proportion of self employed workers has been increasing in many OECD countries so we cannot simply ignore the labour input of these workers. In the current system of national

accounts, the value of the labour input of the self employed is part of *operating surplus*, which is the value of outputs less the value of employee labour input. Thus operating surplus includes capital depreciation, the return to capital (land rent, interest paid, dividends paid and an imputed return to equity capital employed in production) as well as an imputed payment for the labour services of the self employed and the unpaid family workers. We will make some rather arbitrary assumptions in order to obtain an imputation for the value of the labour input of the self employed and unpaid family workers such as assuming that these workers earn 70 % of the amounts that regular employees earn.

**Source 4:** *Labour Force Statistics: 1959-1970*, Paris: OECD, 1972. Call Number: HD 5764 A6 O58. Location: *Koerner Library, Level 1 Stacks*. (The same publication covering the years 1969-1989, 1980-2000 is in the same location). Thus data for all years 1960-2000 can be obtained at this location.

Please collect annual data for the following series:

- *Civilian employment* plus the *armed forces* which is *total employment*; total employment can also be calculated as the *total labour force* less the *unemployed*.
- *Wage earners and salaried employees* (or employees); these are the workers whose earnings are collected in Compensation of Employees in the cost components of the National Accounts;
- *Employers and persons working on their own account* (the self employed); these workers' compensation shows up in Operating Surplus;
- *Unpaid family workers*; these workers' compensation shows up in Operating Surplus.
- Population 15 to 64 years.

From the above sources, please collect data on the total population for your country as well as the number of persons from 15 to 64 years of age. Data for recent years is also available from the next source listed below.

**Source 5:** *Labour Force Statistics: 1982-2002*, Paris: OECD, 2002. Call Number: HD 5764 A6 O58. Location: *Koerner Library, Level 1 Stacks*.

There are several problems with taking the quantity of the labour input of employees to be proportional to the number of wage earners and salaried employees:

- these estimates make no allowance for changes in average hours worked by each employee in each year (over time, hours of work have tended to decline and more holidays and increased vacations have been offered to workers);
- these estimates make no allowance for the changing mix of full time and part time workers and
- each hour of work offered by workers of varying skills is regarded as being equal in its contribution to production.

In order to overcome the last problem listed above, the last version of the international system of national accounts (*System of National Accounts 1993*, Eurostat, IMF, OECD, UN and World Bank, Luxembourg, Washington, D.C., Paris, New York, and Washington, D.C.) recommended that countries construct a proper index number of wages (an employment cost index) but it is only in recent years that a few countries have actually implemented this suggestion.

Our suggested approach to solving the problem of obtaining an estimate of real labour input into the economy has been to use the number of workers as the estimator. However, another approach is to divide our estimate of the value of labour input by a price index, which leads to an implicit estimator for the quantity of labour input. We now pursue this second approach.

**Source 6:** *Main Economic Indicators: Historical Statistics 1960-1979*, Paris: OECD. Call Number: HC10 O68 H58. Location: *Koerner Library, Level 1 Stacks*. (The same publication covering the

years 1962-1991 is in the same location). *Main Economic Indicators: Historical Statistics 1969-1988*, Paris: OECD. Call Number: HC10 O68 H58. *Location: Main Library, Level 1 Stacks.*

These publications contain annual indexes of either:

- weekly wage rates (all activities or just manufacturing) (New Zealand);
- hourly rates in manufacturing (France, Germany);
- hourly rates: industry (France);
- hourly earnings in manufacturing (Germany, Canada);
- unit labour costs in mining and manufacturing (Germany);
- unit labour costs in manufacturing (Canada).

It can be seen that different countries have different wage indexes. Try to pick the most comprehensive (annual) one that is available. Unit labour costs are generally preferable to hourly or weekly wage rates because these series include fringe benefits and try to adjust for changes in the number of days worked each month or year. It may be necessary to link your best series for the years 1962-1991 with other series that are available for the remaining years. Again, there are problems associated with the use of these wage series as deflators for the compensation of employees:

- The wage series usually are not comprehensive; i.e., they cover only a portion of the economy (usually just manufacturing) and typically do not cover service industry wage rates.
- Usually, no adjustments are made (for the wage rate series) for changes in employee benefits (mainly pensions and medical coverage) or for changes in holidays or for changes in days paid for but not worked.

To complete your chosen wage rate series, use the following source:

**Source 7:** *Main Economic Indicators: (Monthly)*, Paris: OECD. Call Number: HC10 O68. *Location: Koerner Library, Level 1 Stacks.* (The same publication covering the earlier years is in the same location). Each of these monthly publications has the data only for 4 years so it will be necessary to use a number of these publications in order to complete your wage rate series for the years 1991-2008 (move forward only 3 years at a time so that you have an overlap year as you move forward). Note that the base year for these indexes (i.e., the year for which the index is set equal to 100) will vary from publication to publication. When the base year changes going from one publication to another, it will be necessary to collect data so that your series overlap each other for at least one year. Data for recent years are available from the following source:

**Source 8:** *Main Economic Indicators: (Monthly)*, Paris: OECD, March 2009. Call Number: HC10 O68. *Location: Koerner Library, Level 2 Reference.*

It should be noted that sources 6 to 8 can also be used to collect interest rates at the same time; see below.

An alternative source for an index of weekly wage rates (see the series: wages: hourly earnings) is:

**Source 9:** *International Financial Statistics*, Washington D.C.: The International Monetary Fund. Call Number: HG1 I55. *Location: Koerner Library, Level 1 Stacks.*

**Source 10:** *International Financial Statistics: Yearbook 2004*: Washington D.C.: The International Monetary Fund. Call Number: HG1 I552. *Location: Koerner Library, Level 2 Reference.*

## 4.4 Capital and Interest Rate Series

We will construct an annual series for capital input into the economy by using the data on investment or gross fixed capital formation collected in section 4.2 above (plus assumptions about the length of life of investment goods or assumptions about depreciation rates). However, we will also require a series on the price of financial capital or on the nominal interest rate that producers face.

It is somewhat difficult to find long time series on nominal interest rates that producers face so we will settle for a time series for government bond interest rates. Our source for this series is Sources 6 to 8 listed above. These publications contain a variety of interest rate series for various countries, including the following series:

- Yield of long term government bonds (New Zealand);
- Official discount rate (France, Germany);
- Call money rate (France, Germany);
- Bond yields: private corporations (France);
- Bond yields: issues guaranteed by the government (France);
- Public and semi public sector bonds (France);
- Treasury bill rate (these are very short term bonds of less than one year) (Germany);
- Yield of government bonds (Germany);
- 7-15 year public sector bonds (Germany);
- Federal government bonds (Canada);
- U.S. government bonds; composite over 10 years;
- Central government bonds (Japan);
- 10 year Commonwealth government bonds (Australia);
- Public sector bonds (Austria);
- Taxable public bonds; 3-6 years (Finland);
- Treasury bonds (Italy);
- Government bonds (more than two years) (Spain);
- Confederation bonds (Switzerland);
- 20 year government bonds (United Kingdom).

As can be seen from the above, there are a bewildering array of possible bond rates to choose from in some cases. Ideally, we would like to have a one year private sector bond rate but for most countries, such an ideal rate is not available going back to 1960. I would not use the official discount rate or Treasury bill rates; these series usually have quite different movements compared to other interest rates and hence these rates are not usually representative. From the remaining series, you will have to make your best judgement as to which series to use. It may also be necessary for you to link your earlier interest rate series taken from the *Main Economic Indicators: Historical Statistics* with a different series taken from the monthly publications. To see if two different series can be linked, check their behavior for the years that they overlap. If the two series are quite similar during the overlap years, then the two series can be linked into one composite series.

An alternative source for an interest rate series is Sources 9 and 10 listed above; see the “Government Bond Rate” series for your country. For Canada, for the early years, the listed series are:

- Bank rate (end of period);
- Treasury bill rate (these are short term government bonds);
- Government bond yield.

For the more recent years, for Canada, the listed series are:

- Bank rate (end of period);
- Money market rate;
- Corporate paper rate;
- Treasury bill rate;
- Deposit rate;
- Lending rate;
- Government bond yield, medium term;
- Government bond yield, long term.

My two preferred choices would be:

- The treasury bill rate or
- The government bond yield (early years) linked to the government bond yield, medium term.

## 4.5 Taxes and Tax Rates

The national accounts data on outputs that was collected in section 4.2 above includes all commodity taxes paid by the final demanders of the products and services. In particular, the price series include indirect commodity taxes that are paid for by purchasers. This is the correct treatment of commodity taxes from the viewpoint of a consumer, who faces the after tax prices on goods and services. However, from the viewpoint of producers, a price that includes commodity taxes that fall on outputs is *not* the price that producers face. The price of outputs that producers face should *not* include any taxes that are collected on the sales of outputs. Hence, when we are estimating production functions or profit functions (the producer side of our general equilibrium model), these indirect taxes that fall on outputs should be removed from output prices. In order to accomplish this removal, we require information on the amount of indirect taxes collected on outputs. A subsidy on the production of an output works in the opposite way to the way an output commodity tax works so what we require is information on indirect taxes on outputs less producer subsidies that fall on outputs. Now in section 4.2 above, we collected information on indirect taxes and subsidies so it would seem that the required information is readily at hand. However, there is a problem; namely, the series on indirect taxes includes not only the commodity taxes on outputs but it includes also certain commodity taxes that fall on *inputs* used by the production sector of the economy. Two examples of commodity taxes that fall on inputs are:

- property taxes on land and structures and
- gasoline taxes that fall within the integrated production sector of the economy; i.e., the energy taxes that trucks, railroads and aircraft pay while moving goods from industry to industry are taxes on the intermediate inputs used by the receiving sectors.

We will not model the second situation but it is necessary to remove property taxes from indirect taxes that fall on outputs. These property taxes can be better modeled as taxes that fall on the use of capital.

Another tax complication occurs with the treatment of labour. The wage rate that we have constructed in section 4.3 and the compensation of employees constructed in section 4.2 include (in theory) all of the income taxes that fall on labour income plus all of the payroll taxes paid by firms. These tax wedges should be removed in order to calculate the after tax wage rate faced by the consumer since it is the after tax wage rate that is relevant in modeling the labour supply decision. Thus while the price and quantity series that we have constructed for labour are the correct ones from the viewpoint of the firm, they are *not* the correct ones from the viewpoint of the consumer's utility maximization problem that generates commodity demand functions and labour supply functions.

The bottom line is that we need to allocate all of the taxes in the economy into 5 categories:

- taxes that fall on outputs;
- taxes that fall on exports;
- taxes that fall on imports;
- taxes that fall on labour and
- taxes that fall on capital.

There are two main sources that can help us with this tax allocation problem:

**Source 11:** *Revenue Statistics: 1965-1974*, Paris: OECD. Call Number: HJ2279 O75. *Location:* Koerner Library, Level 1 Stacks. At the same location, you can find the volumes for 1965-1980

(which has the data for 1965, and the years 1970-1979), 1965-1986 (which has the data for the years 1977-1985), 1965-1990 (which has the data for the years 1985-1990) and for 1965-1995 (which has the data for the years 1990-1994). The data for the more recent years, 1995-1999, can be found in *Revenue Statistics: 1965-2000*, Paris: OECD, 2001. Call Number: HJ2279 O75. *Location: Koerner Library, Level 1 Stacks*. Note that we cannot obtain data for the years 1960-1964 using these sources. However, aggregated data for 1960 can be found on page 185 of *Revenue Statistics: 1965-1990*, Paris: OECD. Call Number: HJ2279 O75. *Location: Koerner Library, Level 1 Stacks*. But OECD.Stat usually has all of the tax data except for the years 1961-1964. We will deal with this lack of data problem for these missing years in our econometric modeling by interpolating tax rates between 1960 and 1965. The data for the most recent years should be found in the following location (or online at OECD.Stat):

**Source 12:** *Revenue Statistics: 1965-200?*, Paris: OECD, 200?. Call Number: HJ2279 O75. *Location: Koerner Library, Level 2 Reference*.

However, all of these data can readily be downloaded using OECD.Stat.

**Source 13:** *Government Finance Statistics Yearbook, 1980*, Washington, D.C.: International Monetary Fund. Call Number: HJ101 G67 H45. *Location: Main Library, Level 2 Stacks or Koerner Level 1 Stacks*. This volume has data for the years 1971 and 1973-1978. At the same location, you can find the volumes for 1988 (which has the data for the years 1977-1985), 1996 (which has the data for the years 1985-1994).

**Source 14:** *Government Finance Statistics Yearbook, 2001*, Washington, D.C.: International Monetary Fund. Call Number: HJ101 G67 H45. *Location: Koerner Library, Level 1 Stacks*. This volume has data for the years 1990-1999. At the same location, you can find the volumes for the years 1997, 1998 and 1999. The *Yearbook* for the most recent year has data for the years prior to the year of publication.

Using sources 11 and 12, please table the following series:

- 1000, taxes on income, profits and capital gains;
- 1100, taxes on individuals (we will assume that these taxes fall on labour; we will assume that 1000-1100 falls on capital);
- 2000, social security contributions (we assume that these taxes fall on labour);
- 3000, taxes on payroll and workforce (we assume that these taxes fall on labour);
- 4000, taxes on property (we assume that these taxes fall on capital);
- 5000, taxes on goods and services (we assume that these taxes less customs and import duties fall on outputs);
- 5123, customs and import duties (we assume that these taxes fall on imports).

Using sources 13 and 14, please table the following series:

From Table A: Revenue and Grants, Consolidated Central Government:

- 1, tax on income, profits and capital gains;
- 1.1, individual (we will assume that these taxes fall on labour; we will assume that 1 minus 1.1 falls on capital);
- 2, social security contributions (we assume that these taxes fall on labour);
- 3, taxes on payroll and workforce (we assume that these taxes fall on labour);
- 4, taxes on property (we assume that these taxes fall on capital);
- 5, domestic taxes on goods and services (we assume that these taxes fall on outputs);
- 6, taxes on international trade (we assume that these taxes fall on imports).

Other levels of government (state, regional or provincial governments): Revenue and Grants:

- 1, tax on income, profits and capital gains;

- 1.1, individual (we will assume that these taxes fall on labour; we will assume that 1 minus 1.1 falls on capital);
- 2, social security contributions (we assume that these taxes fall on labour);
- 3, taxes on payroll and workforce (we assume that these taxes fall on labour);
- 4, taxes on property (we assume that these taxes fall on capital);
- 5, domestic taxes on goods and services (we assume that these taxes fall on outputs);
- 7, other taxes (assume that these fall on goods).

From Table L, Local Government, Part A, Revenue and Grants:

- 3, taxes on payroll and workforce (we assume that these taxes fall on labour);
- 4, taxes on property (we assume that these taxes fall on capital);
- 5, domestic taxes on goods and services (we assume that these taxes fall on outputs).

I have regarded export taxes as being equivalent to import taxes in the above. Most countries tax imports but do not tax exports so usually export taxes will be small.

It can be seen that the OECD *Revenue Statistics* are much simpler to work with than the IMF *Government Finance Statistics Yearbook*.

## Chapter 5

# Index Number Theory: Part I: Early Approaches

### 5.1 Index Number Purpose and Overview

“The answer to the question what is the *Mean* of a given set of magnitudes cannot in general be found, unless there is given also the object for the sake of which a mean value is required. There are as many kinds of average as there are purposes; and we may almost say in the matter of prices as many purposes as writers. Hence much vain controversy between persons who are literally at cross purposes.” F.Y. Edgeworth (1888; 347)[165].

The number of physically distinct goods and unique types of services that consumers can purchase is in the millions.\*<sup>1</sup> On the business or production side of the economy, there are even more commodities that are actively traded. This is because firms not only produce commodities for final consumption, they also produce exports and intermediate commodities that are demanded by other producers. Firms collectively also use millions of imported goods and services, thousands of different types of labor services and hundreds of thousands of specific types of capital. If we further distinguish physical commodities by their geographic location or by the season or time of day that they are produced or consumed, then there are *billions* of commodities that are traded within each year of any advanced economy. Yet most macroeconomic models have only half a dozen quantity variables and many have only three: output, labor and capital. The models used in applied microeconomics generally have less than 20 or so quantity variables. The question that this Chapter addresses is: *how exactly should the microeconomic information involving possibly millions of prices and quantities be aggregated into a smaller number of price and quantity variables?* This is the basic *index number problem*.

Note that we have posed the index number problem in the context of microeconomic theory; i.e., given that we wish to implement some economic model based on producer or consumer theory, what is the “best” method for constructing a set of aggregates for the model? However, when constructing aggregate prices or quantities, other points of view (that do not rely on economics) are possible. We will also consider these alternative points of view in Part I (this chapter) but the primary focus will be on economic approaches to index number theory, which is Part II (next chapter). Thus in sections 5.2 to 5.5 of Part I below, we consider some of the early noneconomic approaches to index number theory.

Section 5.2 will deal with the *levels approach to index number theory*, which will prove to be unsuc-

---

\*<sup>1</sup> The material for this Chapter is drawn from earlier versions of this Chapter and Diewert (2012)[134].

cessful. In the following sections, we focus on determining functional forms for bilateral indexes. A *bilateral index* attempts to determine the rate of aggregate price and quantity change rather than aggregate price and quantity levels. We will consider three alternative approaches to the determination of the functional form for a bilateral price or quantity index. Section 5.3 below will consider fixed basket approaches; section 5.4 will consider stochastic or descriptive statistics approaches and section 5.5 will consider the axiomatic approach to the determination of the index number formula. Finally, section 5.6 will consider when chained indexes should be used.

## 5.2 Setting the Stage and the Levels Approach to the Index Number Problem

It will be useful to set the stage for the subsequent discussion of alternative approaches by defining more precisely what the index number problem is.

We specify two accounting periods,  $t = 0, 1$  for which we have micro price and quantity data for  $N$  commodities pertaining to transactions by a consumer (or a well defined group of consumers). Denote the price and quantity of commodity  $n$  in period  $t$  by  $p_n^t$  and  $q_n^t$  respectively for  $n = 1, 2, \dots, N$  and  $t = 0, 1$ . Before proceeding further, we need to discuss the exact meaning of the microeconomic prices and quantities if there are *multiple* transactions for say commodity  $n$  within period  $t$ . In this case, it is natural to interpret  $q_n^t$  as the *total* amount of commodity  $n$  transacted within period  $t$ . In order to conserve the value of transactions, it is necessary that  $p_n^t$  be defined as a *unit value*<sup>\*2</sup>; i.e.,  $p_n^t$  must be equal to the value of transactions for commodity  $n$  during period  $t$  divided by the total quantity transacted,  $q_n^t$ . For  $t = 0, 1$ , define *the value of transactions in period  $t$*  as:

$$V^t \equiv \sum_{n=1}^N p_n^t q_n^t \equiv \mathbf{p}^t \cdot \mathbf{q}^t \quad (5.1)$$

where  $\mathbf{p}^t \equiv (p_1^t, \dots, p_N^t)$  is the period  $t$  price vector,  $\mathbf{q}^t \equiv (q_1^t, \dots, q_N^t)$  is the period  $t$  quantity vector and  $\mathbf{p}^t \cdot \mathbf{q}^t$  denotes the inner product of these two vectors.

Using the above notation, we can now state the following *levels version of the index number problem using the test or axiomatic approach*: for  $t = 0, 1$ , find scalar numbers  $P^t$  and  $Q^t$  such that

$$V^t = P^t Q^t. \quad (5.2)$$

The number  $P^t$  is interpreted as an aggregate period  $t$  price level while the number  $Q^t$  is interpreted as an aggregate period  $t$  quantity level. The aggregate price level  $P^t$  is allowed to be a function of the period  $t$  price vector,  $\mathbf{p}^t$  while the aggregate period  $t$  quantity level  $Q^t$  is allowed to be a function of the period  $t$  quantity vector,  $\mathbf{q}^t$ ; i.e., we have

$$P^t = c(\mathbf{p}^t) \text{ and } Q^t = f(\mathbf{q}^t); \quad t = 0, 1. \quad (5.3)$$

However, from the viewpoint of the *test approach* to index number theory, the levels approach to finding aggregate quantities and prices comes to an abrupt halt: Eichhorn (1978; 144)[170] showed that if the number of commodities  $N$  in the aggregate is equal to or greater than 2 and we restrict  $c(\mathbf{p}^t)$  and  $f(\mathbf{q}^t)$  to be positive if the micro prices and quantities  $p_n^t$  and  $q_n^t$  are positive, then there do not exist any functions  $c$  and  $f$  such that  $c(\mathbf{p}^t)f(\mathbf{q}^t) = \mathbf{p}^t \cdot \mathbf{q}^t$  for all  $\mathbf{p}^t \gg \mathbf{0}_N$  and  $\mathbf{q}^t \gg \mathbf{0}_N$ .<sup>\*3</sup>

<sup>\*2</sup> The early index number theorists Walsh (1901; 96)[389], Fisher (1922; 318)[187] and Davies (1924; 96)[66] (1932)[68] all suggested unit values as the prices that should be inserted into an index number formula. This advice is followed in the *Consumer Price Index Manual: Theory and Practice* with the proviso that the unit value be a narrowly defined one; see the ILO (2004; 356)[249].

<sup>\*3</sup> Notation:  $\mathbf{p} \gg \mathbf{0}_N$  means all components of  $\mathbf{p}$  are positive;  $\mathbf{p} \geq \mathbf{0}_N$  means all components of  $\mathbf{p}$  are nonnegative and  $\mathbf{p} > \mathbf{0}_N$  means  $\mathbf{p} \geq \mathbf{0}_N$  but  $\mathbf{p} \neq \mathbf{0}_N$ . Finally,  $\mathbf{p} \cdot \mathbf{q} \equiv \sum_{n=1}^N p_n q_n$ .

This negative result can be reversed if we take the *economic approach* to index number theory. This economic approach is due to Shephard (1953)[355] (1970)[358] and Samuelson and Swamy (1974)[348]. In this approach, we assume that the economic agent has a linearly homogeneous utility function,  $f(\mathbf{q})$ , and when facing the prices  $\mathbf{p}^t$  chooses  $\mathbf{q}^t$  to solve the following cost minimization problem:

$$\min_{\mathbf{q}} \{\mathbf{p}^t \cdot \mathbf{q} : \mathbf{p}^t \cdot \mathbf{q} = Y^t; \mathbf{q} \geq \mathbf{0}_N\}; \quad t = 0, 1 \quad (5.4)$$

where period  $t$  “income”  $Y^t$  is defined as  $\mathbf{p}^t \cdot \mathbf{q}^t$ . In this setup, it turns out that  $c(\mathbf{p})$  is the unit cost function that is dual\*<sup>4</sup> to the linearly homogeneous utility function  $f(\mathbf{q})$  and we can define  $P^t$  and  $Q^t$  as in (5.3) with  $P^t Q^t = c(\mathbf{p}^t) f(\mathbf{q}^t) = \mathbf{p}^t \cdot \mathbf{q}^t$  for  $t = 0, 1$ . Why does the economic approach work in the levels version of the index number problem whereas the test approach does not? In the test approach, both  $\mathbf{p}^t$  and  $\mathbf{q}^t$  are regarded as completely independent variables, whereas in the economic approach,  $\mathbf{p}^t$  can vary independently but  $\mathbf{q}^t$  cannot vary independently; it is a solution to the period  $t$  cost minimization problem (5.4).

Even though the economic approach to the index number problem as formulated above “works”, it is not a *practical* solution that statistical agencies can implement and provide suitable aggregates to the public. In order to implement this solution, the statistical agency would have to hire hundreds of econometricians in order to estimate cost functions for all relevant macroeconomic aggregates and it is simply not feasible to do this. Thus we turn to our second formulation of the index number problem and it is this formulation that was initiated Walsh (1901)[389] (1921a)[391] and Fisher (1911)[185] (1922)[187] in their books on index number theory.

In the second approach to index number theory, instead of trying to decompose the value of the aggregate into price and quantity components for a single period, we instead attempt to decompose a *value ratio* for the two periods under consideration into a *price change component*  $P$  times a *quantity change component*  $Q$ .\*<sup>5</sup> Thus we now look for two functions of  $4N$  variables,  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  and  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  such that:\*<sup>6</sup>

$$\frac{\mathbf{p}^1 \cdot \mathbf{q}^1}{\mathbf{p}^0 \cdot \mathbf{q}^0} = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1). \quad (5.5)$$

If we take the test approach, then we want equation (5.5) to hold for all positive price and quantity vectors pertaining to the two periods under consideration,  $\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1$ . If we take the economic approach, then only the price vectors  $\mathbf{p}^0$  and  $\mathbf{p}^1$  are regarded as independent variables while the quantity vectors,  $\mathbf{q}^0$  and  $\mathbf{q}^1$ , are regarded as dependent variables.

In this second approach to index number theory, the *price index*  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  and the *quantity index*  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  cannot be determined independently; i.e., if either one of these two functions is determined, then the remaining function is implicitly determined using equation (5.5). Historically, the focus has been on the determination of the price index but Fisher (1911; 388)[185] was the first to realize that once the price index was determined, then equation (5.5) could be used to determine the companion quantity index.\*<sup>7</sup>

\*<sup>4</sup> See Chapter 3 or Diewert (1974)[76] for materials and references to the literature on duality theory.

\*<sup>5</sup> If we use the economic approach,  $P$  can be interpreted to be the ratio of unit cost functions,  $c(\mathbf{p}^1)/c(\mathbf{p}^0)$ , and  $Q$  can be interpreted to be the utility ratio,  $f(\mathbf{q}^1)/f(\mathbf{q}^0)$ . Note that the linear homogeneity assumption on the utility function  $f$  effectively cardinalizes utility.

\*<sup>6</sup> If  $N = 1$ , then we define  $P(p_1^0, p_1^1, q_1^0, q_1^1) \equiv p_1^1/p_1^0$  and  $Q(p_1^0, p_1^1, q_1^0, q_1^1) \equiv q_1^1/q_1^0$ , the single price ratio and the single quantity ratio respectively. In the case of a general  $N$ , we think of  $P(p_1^0, p_1^1, q_1^0, q_1^1)$  as being a weighted average of the price ratios  $p_1^1/p_1^0, p_2^1/p_2^0, \dots, p_N^1/p_N^0$ . Thus we interpret  $P(p_1^0, p_1^1, q_1^0, q_1^1)$  as an aggregate price ratio,  $P^1/P^0$ , where  $P^t$  is the aggregate price level for period  $t$  for  $t = 0, 1$ .

\*<sup>7</sup> This approach to index number theory is due to Fisher (1911; 418)[185] who called the implicitly determined  $Q$ , the *correlative formula*. Frisch (1930; 399)[191] later called (5.5) the *product test*.

This value ratio decomposition approach to index number is called *bilateral index number theory* and its focus is the determination of “reasonable” functional forms for  $P$  and  $Q$ . Fisher’s 1911 and 1922 books address this functional form issue using the test approach.

Once the functional forms for  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  and  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  have been determined, price and quantity *levels* for the two periods under consideration can be determined as follows:

- The price and quantity levels for period 0 are determined as  $P^0 \equiv 1$  and  $Q^0 \equiv V^0 \equiv \mathbf{p}^0 \cdot \mathbf{q}^0$ ;
- The price level for period 1 is set equal to  $P^1 \equiv P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)P^0 = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ ;
- The quantity level for period 1 is set equal to  $Q^1 \equiv Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)Q^0 = Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)V^0$ .

We turn now to a discussion of the various approaches that have been used to determine the functional form for the bilateral price index,  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ .

### 5.3 Fixed Basket Approaches to Bilateral Index Number Theory

A very simple approach to the determination of a price index over a group of commodities is the *fixed basket approach*. In this approach, we are given a basket of commodities that is represented by the positive quantity vector  $\mathbf{q}$ . Given the price vectors for periods 0 and 1,  $\mathbf{p}^0$  and  $\mathbf{p}^1$  respectively, we can calculate the cost of purchasing this same basket in the two periods,  $\mathbf{p}^0 \cdot \mathbf{q}$  and  $\mathbf{p}^1 \cdot \mathbf{q}$ . Then the ratio of these costs is a very reasonable indicator of pure price change over the two periods under consideration, provided that the basket vector  $\mathbf{q}$  is “representative”. Thus define the *Lowe* (1823)[295] *price index*,  $P_{Lo}$ , as follows:<sup>\*8</sup>

$$P_{Lo}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}) \equiv \frac{\mathbf{p}^1 \cdot \mathbf{q}}{\mathbf{p}^0 \cdot \mathbf{q}}. \quad (5.6)$$

As time passed, economists and price statisticians demanded a bit more precision with respect to the specification of the basket vector  $\mathbf{q}$ . There are two natural choices for the reference basket: the period 0 commodity vector  $\mathbf{q}^0$  or the period 1 commodity vector  $\mathbf{q}^1$ . These two choices lead to the Laspeyres (1871)[285] price index  $P_L$  defined by (5.7) and the Paasche (1874)[325] price index  $P_P$  defined by (5.8):<sup>\*9</sup>

$$P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv \frac{\mathbf{p}^1 \cdot \mathbf{q}^0}{\mathbf{p}^0 \cdot \mathbf{q}^0} = \sum_{n=1}^N s_n^0 (p_n^1/p_n^0); \quad (5.7)$$

$$\begin{aligned} P_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) &= \frac{\mathbf{p}^1 \cdot \mathbf{q}^1}{\mathbf{p}^0 \cdot \mathbf{q}^1} \\ &= [\mathbf{p}^0 \cdot \mathbf{q}^1 / \mathbf{p}^1 \cdot \mathbf{q}^1]^{-1} \\ &= \left[ \sum_{n=1}^N p_n^0 q_n^1 / \mathbf{p}^1 \cdot \mathbf{q}^1 \right]^{-1} \\ &= \left[ \sum_{n=1}^N p_n^1 q_n^1 (p_n^0/p_n^1) / \mathbf{p}^1 \cdot \mathbf{q}^1 \right]^{-1} \\ &= \left[ \sum_{n=1}^N p_n^1 q_n^1 (p_n^1/p_n^0)^{-1} / \mathbf{p}^1 \cdot \mathbf{q}^1 \right]^{-1} \\ &= \left[ \sum_{n=1}^N s_n^1 (p_n^1/p_n^0)^{-1} / \mathbf{p}^1 \cdot \mathbf{q}^1 \right]^{-1} \end{aligned} \quad (5.8)$$

<sup>\*8</sup> See Ferger (1946)[180] for the early history of the Lowe index.

<sup>\*9</sup> Note that  $P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  does not actually depend on  $\mathbf{q}^1$  and  $P_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  does not actually depend on  $\mathbf{q}^0$ . However, it does no harm to include these vectors and the notation indicates that we are in the realm of bilateral index number theory.

where the period  $t$  expenditure share on commodity  $n$ ,  $s_n^t$ , is defined as  $p_n^t q_n^t / \mathbf{p}^t \cdot \mathbf{q}^t$  for  $n = 1, \dots, N$  and  $t = 0, 1$ . Thus the Laspeyres price index  $P_L$  can be written as a base period expenditure share weighted average of the  $N$  price ratios (or price relatives),  $p_n^1/p_n^0$ .<sup>\*10</sup> The last equation in (5.8) shows that the Paasche price index  $P_P$  can be written as a period 1 (or current period) expenditure share weighted *harmonic* average of the  $N$  price ratios.<sup>\*11</sup>

The problem with these index number formulae is that they are equally plausible but in general, they will give different answers. This suggests that if we require a single estimate for the price change between the two periods, then we should take some sort of evenly weighted average of the two indexes as our final estimate of price change between periods 0 and 1. Examples of such symmetric averages are the arithmetic mean, which leads to the Drobisch (1871)[162] Sidgwick (1883; 68)[359] Bowley (1901; 227)[41]<sup>\*12</sup> index,  $(1/2)P_L + (1/2)P_P$ , and the geometric mean, which leads to the Fisher (1922)[187] *ideal index*,  $P_F$ , defined as

$$P_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv [P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)P_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)]^{1/2}. \quad (5.9)$$

At this point, the fixed basket approach to index number theory has to draw on the *test approach* to index number theory; i.e., in order to determine which of these fixed basket indexes or which averages of them might be “best”, we need *criteria* or *tests* or *properties* that we would like our indexes to satisfy.

What is the “best” symmetric average of  $P_L$  and  $P_P$  to use as a point estimate for the theoretical cost of living index? It is very desirable for a price index formula that depends on the price and quantity vectors pertaining to the two periods under consideration to satisfy the *time reversal test*.<sup>\*13</sup> We say that the index number formula  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  satisfies this test if

$$P(\mathbf{p}^1, \mathbf{p}^0, \mathbf{q}^1, \mathbf{q}^0) = \frac{1}{P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)}; \quad (5.10)$$

i.e., if we interchange the period 0 and period 1 price and quantity data and evaluate the index, then this new index  $P(\mathbf{p}^1, \mathbf{p}^0, \mathbf{q}^1, \mathbf{q}^0)$  is equal to the reciprocal of the original index  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ .

Diewert (1997; 138)[115] showed that the Fisher ideal price index defined by (5.9) above is the *only* index that is a homogeneous symmetric mean of the Laspeyres and Paasche price indexes,  $P_L$  and  $P_P$ , and satisfies the time reversal test (5.10) above. Thus our first *symmetric basket approach* to bilateral index number theory leads to the Fisher index (5.9) as being “best” from the perspective of this approach.<sup>\*14</sup>

Instead of looking for a “best” average of the two fixed basket indexes that correspond to the baskets chosen in either of the two periods being compared, we could instead look for a “best” average basket of the two baskets represented by the vectors  $\mathbf{q}^0$  and  $\mathbf{q}^1$  and then use this average basket to compare

<sup>\*10</sup> This result is due to Walsh (1901; 428 and 539)[389].

<sup>\*11</sup> This expenditure share and price ratio representation of the Paasche index is described by Walsh (1901; 428)[389] and derived explicitly by Fisher (1911; 365)[185].

<sup>\*12</sup> See Diewert (1992)[101] (1993)[109] and Balk (2008)[21] for additional references to the early history of index number theory.

<sup>\*13</sup> The concept of this test is due to Pierson (1896; 128)[326], who was so upset with the fact that many of the commonly used index number formulae did not satisfy this test (and the commensurability test to be discussed later) that he proposed that the entire concept of an index number should be abandoned. More formal statements of the test were made by Walsh (1901; 324)[389] and Fisher (1922; 64)[187].

<sup>\*14</sup> Bowley was an early advocate of taking a symmetric average of the Paasche and Laspeyres indexes: “If [the Paasche index] and [the Laspeyres index] lie close together there is no further difficulty; if they differ by much they may be regarded as inferior and superior limits of the index number, which may be estimated as their arithmetic mean . . . as a first approximation.” Arthur L. Bowley (1901; 227)[41]. Fisher (1911; 418-419)[185] (1922)[187] considered taking the arithmetic, geometric and harmonic averages of the Paasche and Laspeyres indexes.

the price levels of periods 0 and 1.<sup>\*15</sup> Thus we ask that the  $n$ th quantity weight,  $q_n$ , be an average or *mean* of the base period quantity  $q_n^0$  and the period 1 quantity for commodity  $n$   $q_n^1$ , say  $m(q_n^0, q_n^1)$ , for  $n = 1, 2, \dots, N$ .<sup>\*16</sup> Price statisticians refer to this type of index as a *pure price index* and it corresponds to Knibbs' (1924; 43)[271] *unequivocal price index*. Under these assumptions, the pure price index can be defined as a member of the following class of index numbers:

$$P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv \frac{\sum_{n=1}^N p_n^1 m(q_n^0, q_n^1)}{\sum_{j=1}^N p_j^0 m(q_j^0, q_j^1)}. \quad (5.11)$$

In order to determine the functional form for the mean function  $m$ , it is necessary to impose some *tests* or *axioms* on the pure price index defined by (5.11). Suppose that we impose the time reversal test (5.10) and the following *invariance to proportional changes in current quantities test*:

$$P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \lambda \mathbf{q}^1) = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \text{ for all } \lambda > 0. \quad (5.12)$$

Diewert (2001; 207)[119] showed that these two tests determine the precise functional form for the pure price index  $P_K$  defined by (5.11) above: the pure price index  $P_K$  must be the *Walsh* (1901; 398)[389] (1921a; 97)[391] *price index*,  $P_W$ <sup>\*17</sup> defined by (5.13):

$$P_W(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv \frac{\sum_{n=1}^N p_n^1 (q_n^0 q_n^1)^{1/2}}{\sum_{j=1}^N p_j^0 (q_j^0 q_j^1)^{1/2}}. \quad (5.13)$$

Thus the fixed basket approach to bilateral index number theory starts out with the Laspeyres and Paasche price indexes. Some form of averaging of these two indexes is called for since both indexes are equally plausible. Averaging these two indexes directly leads to the Fisher ideal index  $P_F$  defined by (5.9) as being “best” while a direct averaging of the two quantity baskets  $\mathbf{q}^0$  and  $\mathbf{q}^1$  leads to the Walsh price index  $P_W$  defined by (5.13) as being “best”.

We turn now to another early approach to the index number problem.

## 5.4 Stochastic and Descriptive Statistics Approaches to Index Number Theory

The (unweighted) stochastic approach to the determination of the price index can be traced back to the work of Jevons (1865)[252] (1884)[253] and Edgeworth (1888)[165] (1896)[166] (1901)[167] over a hundred years ago<sup>\*18</sup>.

The basic idea behind the stochastic approach is that each price relative,  $p_n^1/p_n^0$  for  $n = 1, 2, \dots, N$ , can be regarded as an estimate of a common inflation rate  $\alpha$  between periods 0 and 1; i.e., Jevons and Edgeworth essentially assumed that

$$p_n^1/p_n^0 = \alpha + \varepsilon_n; \quad n = 1, 2, \dots, N \quad (5.14)$$

<sup>\*15</sup> Walsh (1901)[389] (1921a)[391] and Fisher (1922)[187] considered both averaging strategies in their classic studies on index numbers.

<sup>\*16</sup> Note that we have chosen the mean function  $m(q_n^0, q_n^1)$  to be the same for each commodity  $n$ .

<sup>\*17</sup> Walsh endorsed  $P_W$  as being the best index number formula: “We have seen reason to believe formula 6 better than formula 7. Perhaps formula 9 is the best of the rest, but between it and Nos. 6 and 8 it would be difficult to decide with assurance.” C.M. Walsh (1921a; 103)[391]. His formula 6 is  $P_W$  defined by (5.13) and his 9 is the Fisher ideal defined by (5.9) above. His formula 8 is the formula  $\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0 Q_W(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ , which is known as the implicit Walsh price index where  $Q_W(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is the Walsh quantity index defined by (5.13) except the role of prices and quantities is interchanged. Thus although Walsh thought that his Walsh price index was the best functional form, his implicit Walsh price index and the “Fisher” formula were not far behind.

<sup>\*18</sup> For additional references to the early literature, see Diewert (1993; 37-38)[109] (1995)[113] and Balk (2008; 32-36)[21].

where  $\alpha$  is the common inflation rate and the  $\varepsilon_n$  are random variables with mean 0 and variance  $\sigma^2$ . The least squares estimator for  $\alpha$  is the *Carli* (1804)[46] *price index*  $P_C$  defined as

$$P_C(\mathbf{p}^0, \mathbf{p}^1) \equiv \sum_{n=1}^N (1/N)(p_n^1/p_n^0). \quad (5.15)$$

Unfortunately,  $P_C$  does not satisfy the time reversal test, i.e.,  $P_C(\mathbf{p}^1, \mathbf{p}^0) \neq 1/P_C(\mathbf{p}^0, \mathbf{p}^1)$ <sup>\*19</sup>.

Now assume that the logarithm of each price relative,  $\ln(p_n^1/p_n^0)$ , is an independent unbiased estimate of the logarithm of the inflation rate between periods 0 and 1,  $\beta$  say. Thus we have:

$$\ln(p_n^1/p_n^0) = \beta + \varepsilon_n; \quad n = 1, 2, \dots, N \quad (5.16)$$

where  $\beta \equiv \ln \alpha$  and the  $\varepsilon_n$  are independently distributed random variables with mean 0 and variance  $\sigma^2$ . The least squares or maximum likelihood estimator for  $\beta$  is the logarithm of the geometric mean of the price relatives. Hence the corresponding estimate for the common inflation rate  $\alpha$  is the *Jevons* (1865)[252] *price index*  $P_J$  defined as:

$$P_J(\mathbf{p}^0, \mathbf{p}^1) \equiv \prod_{n=1}^N (p_n^1/p_n^0)^{1/N}. \quad (5.17)$$

The Jevons price index  $P_J$  does satisfy the time reversal test and hence is much more satisfactory than the Carli index  $P_C$ . However, both the Jevons and Carli price indexes suffer from a fatal flaw: each price relative  $p_n^1/p_n^0$  is regarded as being equally important and is given an equal weight in the index number formulae (5.15) and (5.17).<sup>\*20</sup> Keynes (1930; 76-81)[269] also criticized the unweighted stochastic approach to index number theory on two other grounds: (i) price relatives are not distributed independently and (ii) there is no single inflation rate that can be applied to all parts of an economy; e.g., Keynes demonstrated empirically that wage rates, wholesale prices and final consumption prices all had different rates of inflation. In order to overcome the Keynesian criticisms of the unweighted stochastic approach to index numbers, it is necessary to:

- Have a definite domain of definition for the index number and
- Weight the price relatives by their economic importance.

Theil (1967; 136-137)[371] proposed a solution to the lack of weighting in (5.15). He argued as follows. Suppose we draw price relatives at random in such a way that each dollar of expenditure in the base period has an equal chance of being selected. Then the probability that we will draw the  $n$ th price relative is equal to  $s_n^0 \equiv p_n^0 q_n^0 / \mathbf{p}^0 \cdot \mathbf{q}^0$ , the period 0 expenditure share for commodity  $n$ . Then the overall mean (period 0 weighted) logarithmic price change is  $\sum_{n=1}^N s_n^0 \ln(p_n^1/p_n^0)$ . Now repeat the above mental experiment and draw price relatives at random in such a way that each dollar of expenditure in period 1 has an equal probability of being selected. This leads to the overall mean (period 1 weighted) logarithmic price change of  $\sum_{n=1}^N s_n^1 \ln(p_n^1/p_n^0)$ . Each of these measures of overall logarithmic price change seems equally valid so we could argue for taking a symmetric average of the two measures in order to obtain a final single measure of overall logarithmic price change. Theil (1967; 137)[371] argued that a nice symmetric index number formula can be obtained

<sup>\*19</sup> In fact Fisher (1922; 66)[187] noted that  $P_C(\mathbf{p}^0, \mathbf{p}^1)P_C(\mathbf{p}^1, \mathbf{p}^0) \geq 1$  unless the period 1 price vector  $\mathbf{p}^1$  is proportional to the period 0 price vector  $\mathbf{p}^0$ ; i.e., Fisher showed that the Carli index has a definite upward bias. Walsh (1901; 327)[389] established this inequality for the case  $N = 2$ . Fisher urged users to abandon the use of the Carli index but his advice was generally ignored by statistical agencies until recently: "In fields other than index numbers it is often the best form of average to use. But we shall see that the simple arithmetic average produces one of the very worst of index numbers. And if this book has no other effect than to lead to the total abandonment of the simple arithmetic type of index number, it will have served a useful purpose." Irving Fisher (1922; 29-30)[187].

<sup>\*20</sup> Walsh (1901)[389] (1921a; 82-83)[391], Fisher (1922; 43)[187] and Keynes (1930; 76-77)[269] all objected to the lack of weighting in the unweighted stochastic approach to index number theory.

if we make the probability of selection for the  $n$ th price relative equal to the arithmetic average of the period 0 and 1 expenditure shares for commodity  $n$ . Using these probabilities of selection, Theil's final measure of overall logarithmic price change is

$$\ln P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv \sum_{n=1}^N (1/2)(s_n^0 + s_n^1) \ln(p_n^1/p_n^0). \quad (5.18)$$

It is possible to give a *descriptive statistics* interpretation of the right hand side of (5.18). Define the  $n$ th logarithmic price ratio  $r_n$  by:

$$r_n \equiv \ln(p_n^1/p_n^0) \quad \text{for } n = 1, \dots, N. \quad (5.19)$$

Now define the discrete random variable,  $R$  say, as the random variable which can take on the values  $r_n$  with probabilities  $\rho_n \equiv (1/2)(s_n^0 + s_n^1)$  for  $n = 1, \dots, N$ . Note that since each set of expenditure shares,  $s_n^0$  and  $s_n^1$ , sums to one, the probabilities  $\rho_n$  will also sum to one. It can be seen that the expected value of the discrete random variable  $R$  is  $\ln P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  as defined by the right hand side of (5.18). Thus the logarithm of the index  $P_T$  can be interpreted as *the expected value of the distribution of the logarithmic price ratios* in the domain of definition under consideration, where the  $N$  discrete price ratios in this domain of definition are weighted according to Theil's probability weights,  $\rho_n$ .

Taking antilogs of both sides of (5.18), we obtain the Törnqvist Theil price index;  $P_T$ .<sup>\*21</sup> This index number formula has a number of good properties. In particular,  $P_T$  satisfies the time reversal test (5.10) and the linear homogeneity test (5.12).<sup>\*22</sup>

Additional material on stochastic approaches to index number theory and references to the literature can be found in Selvanathan and Rao (1994)[354], Diewert (1995)[113] (2004)[123] (2005)[127], Wynne (1997)[406], Clements, Izan and Selvanathan (2006)[58] and Balk (2008; 32-36)[21].

## 5.5 Test Approaches to Index Number Theory

Recall equation (5.5) above, which set the value ratio,  $V^1/V^0$ , equal to the product of the price index,  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ , and the quantity index,  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ .<sup>\*23</sup> This is called the Product Test and we assume that it is satisfied. This equation means that as soon as the functional form for the price index  $P$  is determined, then (5.5) can be used to determine the functional form for the quantity index  $Q$ . However, a further advantage of assuming that the product test holds is that we can assume that the quantity index  $Q$  satisfies a "reasonable" property and then use (5.5) to translate this test on the quantity index into a corresponding test on the price index  $P$ .<sup>\*24</sup>

If  $N = 1$ , so that there is only one price and quantity to be aggregated, then a natural candidate for  $P$  is  $p_1^1/p_1^0$ , the single price ratio, and a natural candidate for  $Q$  is  $q_1^1/q_1^0$ , the single quantity ratio. When the number of commodities or items to be aggregated is greater than 1, then what index number theorists have done over the years is propose properties or tests that the price index  $P$  should satisfy. These properties are generally multi-dimensional analogues to the one good price index formula,  $p_1^1/p_1^0$ . Below, we list twenty-one tests that turn out to characterize the Fisher ideal price index.

<sup>\*21</sup> This index first appeared explicitly as formula 123 in Fisher (1922; 473)[187].  $P_T$  is generally attributed to Törnqvist (1936)[373] but this article did not have an explicit definition for  $P_T$ ; it was defined explicitly in Törnqvist and Törnqvist (1937)[374]; see Balk (2008; 26)[21].

<sup>\*22</sup> For a listing of some of the tests that  $P_T$ ,  $P_F$  and  $P_W$  satisfy, see Diewert (1992; 223)[101]. In Fisher (1922)[187], these indexes were listed as numbers 123, 353 and 1153 respectively.

<sup>\*23</sup> The material in this section is based on Diewert (1992)[101] where more detailed references to the literature on the origins of the various tests can be found.

<sup>\*24</sup> This observation was first made by Fisher (1911; 400-406)[185]. Vogt (1980)[383] also pursued this idea.

We shall assume that every component of each price and quantity vector is positive; i.e.,  $\mathbf{p}^t \gg \mathbf{0}_N$  and  $\mathbf{q}^t \gg \mathbf{0}_N$  for  $t = 0, 1$ . If we want to set  $\mathbf{q}^0 = \mathbf{q}^1$ , we call the common quantity vector  $\mathbf{q}$ ; if we want to set  $\mathbf{p}^0 = \mathbf{p}^1$ , we call the common price vector  $\mathbf{p}$ .

Our first two tests are not very controversial and so we will not discuss them.

T1: *Positivity*:  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) > 0$ .

T2: *Continuity*:  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is a continuous function of its arguments.

Our next two tests are somewhat more controversial.

T3: *Identity or Constant Prices Test*:  $P(\mathbf{p}, \mathbf{p}, \mathbf{q}^0, \mathbf{q}^1) = 1$ .

That is, if the price of every good is identical during the two periods, then the price index should equal unity, no matter what the quantity vectors are. The controversial part of this test is that the two quantity vectors are allowed to be different in the above test.\*<sup>25</sup>

T4: *Fixed Basket or Constant Quantities Test*:  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}, \mathbf{q}) = \sum_{i=1}^N p_i^1 q_i / \sum_{i=1}^N p_i^0 q_i$ .

That is, if quantities are constant during the two periods so that  $\mathbf{q}^0 = \mathbf{q}^1 \equiv \mathbf{q}$ , then the price index should equal the expenditure on the constant basket in period 1,  $\sum_{i=1}^N p_i^1 q_i$ , divided by the expenditure on the basket in period 0,  $\sum_{i=1}^N p_i^0 q_i$ .

The following four tests are *homogeneity tests* and they restrict the behavior of the price index  $P$  as the scale of any one of the four vectors  $\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1$  changes.

T5: *Proportionality in Current Prices*:  $P(\mathbf{p}^0, \lambda \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \lambda P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  for  $\lambda > 0$ .

That is, if all period 1 prices are multiplied by the positive number  $\lambda$ , then the new price index is  $\lambda$  times the old price index. Put another way, the price index function  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is (positively) homogeneous of degree one in the components of the period 1 price vector  $\mathbf{p}^1$ . Most index number theorists regard this property as a very fundamental one that the index number formula should satisfy.

Walsh (1901)[389] and Fisher (1911; 418)[185] (1922; 420)[187] proposed the related proportionality test  $P(\mathbf{p}, \lambda \mathbf{p}, \mathbf{q}^0, \mathbf{q}^1) = \lambda$ . This last test is a combination of T3 and T5; in fact Walsh (1901, 385)[389] noted that this last test implies the identity test, T3.

In our next test, instead of multiplying all period 1 prices by the same number, we multiply all period 0 prices by the number  $\lambda$ .

T6: *Inverse Proportionality in Base Period Prices*:  $P(\lambda \mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \lambda^{-1} P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  for  $\lambda > 0$ .

That is, if all period 0 prices are multiplied by the positive number  $\lambda$ , then the new price index is  $1/\lambda$  times the old price index. Put another way, the price index function  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is (positively) homogeneous of degree minus one in the components of the period 0 price vector  $\mathbf{p}^0$ .

The following two homogeneity tests can also be regarded as invariance tests.

T7: *Invariance to Proportional Changes in Current Quantities*:

$$P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \lambda \mathbf{q}^1) = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \text{ for all } \lambda > 0.$$

That is, if current period quantities are all multiplied by the number  $\lambda$ , then the price index remains unchanged. Put another way, the price index function  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is (positively) homogeneous of degree zero in the components of the period 1 quantity vector  $\mathbf{q}^1$ . Vogt (1980, 70)[383] was the first to propose this test and his derivation of the test is of some interest. Suppose the quantity index

\*<sup>25</sup> Usually, economists assume that given a price vector  $\mathbf{p}$ , the corresponding quantity vector  $\mathbf{q}$  is uniquely determined. Here, we have the same price vector but the corresponding quantity vectors are allowed to be different.

$Q$  satisfies the quantity analogue to the price test T5; i.e., suppose  $Q$  satisfies  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \lambda \mathbf{q}^1) = \lambda Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  for  $\lambda > 0$ . Then using the product test (5.5), we see that  $P$  must satisfy T7.

T8: *Invariance to Proportional Changes in Base Quantities:*

$$P(\mathbf{p}^0, \mathbf{p}^1, \lambda \mathbf{q}^0, \mathbf{q}^1) = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \text{ for all } \lambda > 0.$$

That is, if base period quantities are all multiplied by the number  $\lambda$ , then the price index remains unchanged. Put another way, the price index function  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is (positively) homogeneous of degree zero in the components of the period 0 quantity vector  $\mathbf{q}^0$ . If the quantity index  $Q$  satisfies the following counterpart to T8:  $Q(\mathbf{p}^0, \mathbf{p}^1, \lambda \mathbf{q}^0, \mathbf{q}^1) = \lambda^{-1} Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  for all  $\lambda > 0$ , then using (5.5), the corresponding price index  $P$  must satisfy T8. This argument provides some additional justification for assuming the validity of T8 for the price index function  $P$ .

T7 and T8 together impose the property that the price index  $P$  does not depend on the *absolute* magnitudes of the quantity vectors  $\mathbf{q}^0$  and  $\mathbf{q}^1$ .

The next five tests are *invariance* or *symmetry tests*. Fisher (1922; 62-63, 458-460)[187] and Walsh (1921b; 542)[392] seem to have been the first researchers to appreciate the significance of these kinds of tests. Fisher (1922, 62-63)[187] spoke of fairness but it is clear that he had symmetry properties in mind. It is perhaps unfortunate that he did not realize that there were more symmetry and invariance properties than the ones he proposed; if he had realized this, it is likely that he would have been able to provide an axiomatic characterization for his ideal price index. Our first invariance test is that the price index should remain unchanged if the *ordering* of the commodities is changed:

T9: *Commodity Reversal Test* (or invariance to changes in the ordering of commodities):

$$P(\mathbf{p}^{0*}, \mathbf{p}^{1*}, \mathbf{q}^{0*}, \mathbf{q}^{1*}) = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$$

where  $\mathbf{p}^{t*}$  denotes a permutation of the components of the vector  $\mathbf{p}^t$  and  $\mathbf{q}^{t*}$  denotes the same permutation of the components of  $\mathbf{q}^t$  for  $t = 0, 1$ . This test is due to Irving Fisher (1922)[187], and it is one of his three famous reversal tests. The other two are the time reversal test and the factor reversal test which will be considered below.

T10: *Invariance to Changes in the Units of Measurement* (commensurability test):

$$P(\alpha_1 p_1^0, \dots, \alpha_N p_N^0; \alpha_1 p_1^1, \dots, \alpha_N p_N^1; \alpha_1^{-1} q_1^0, \dots, \alpha_N^{-1} q_N^0; \alpha_1^{-1} q_1^1, \dots, \alpha_N^{-1} q_N^1) = P(p_1^0, \dots, p_N^0; p_1^1, \dots, p_N^1; q_1^0, \dots, q_N^0; q_1^1, \dots, q_N^1) \text{ for all } \alpha_1 > 0, \dots, \alpha_N > 0.$$

That is, the price index does not change if the units of measurement for each commodity are changed. The concept of this test was due to Jevons (1884; 23)[253] and the Dutch economist Pierson (1896; 131)[326], who criticized several index number formula for not satisfying this fundamental test. Fisher (1911; 411)[185] first called this test *the change of units test* and later, Fisher (1922; 420)[187] called it the *commensurability test*.

T11: *Time Reversal Test:*  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = 1/P(\mathbf{p}^1, \mathbf{p}^0, \mathbf{q}^1, \mathbf{q}^0)$ .

That is, if the data for periods 0 and 1 are interchanged, then the resulting price index should equal the reciprocal of the original price index. Obviously, in the one good case when the price index is simply the single price ratio; this test is satisfied (as are all of the other tests listed in this section). When the number of goods is greater than one, many commonly used price indexes fail this test; e.g., the Laspeyres (1871)[285] price index,  $P_L$  defined earlier by (5.7), and the Paasche (1874)[325] price index,  $P_P$  defined earlier by (5.8), both *fail* this fundamental test. The concept of the test was due to Pierson (1896; 128)[326], who was so upset with the fact that many of the commonly used index number formulae did not satisfy this test, that he proposed that the entire concept of an index number should be abandoned. More formal statements of the test were made by Walsh (1901; 368)[389] (1921b; 541)[392] and Fisher (1911; 534)[185] (1922; 64)[187].

Our next two tests are more controversial, since they are not necessarily consistent with the economic

approach to index number theory. However, these tests are quite consistent with the weighted stochastic approach to index number theory discussed earlier in section 5.4.

T12: *Quantity Reversal Test* (quantity weights symmetry test):

$$P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^1, \mathbf{q}^0).$$

That is, if the quantity vectors for the two periods are interchanged, then the price index remains invariant. This property means that if quantities are used to weight the prices in the index number formula, then the period 0 quantities  $\mathbf{q}^0$  and the period 1 quantities  $\mathbf{q}^1$  must enter the formula in a symmetric or even handed manner. Funke and Voeller (1978; 3)[194] introduced this test; they called it the *weight property*.

The next test is the analogue to T12 applied to quantity indexes:

T13: *Price Reversal Test* (price weights symmetry test):

$$\{\sum_{i=1}^N p_i^1 q_i^1 / \sum_{i=1}^N p_i^0 q_i^0\} / P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \{\sum_{i=1}^N p_i^0 q_i^1 / \sum_{i=1}^N p_i^1 q_i^0\} / P(\mathbf{p}^1, \mathbf{p}^0, \mathbf{q}^0, \mathbf{q}^1).$$

Thus if we use (5.5) to define the quantity index  $Q$  in terms of the price index  $P$ , then it can be seen that T13 is equivalent to the following property for the associated quantity index  $Q$ :  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = Q(\mathbf{p}^1, \mathbf{p}^0, \mathbf{q}^0, \mathbf{q}^1)$ . That is, if the price vectors for the two periods are interchanged, then the quantity index remains invariant. Thus if prices for the same good in the two periods are used to weight quantities in the construction of the quantity index, then property T13 implies that these prices enter the quantity index in a symmetric manner.

The next three tests are mean value tests.

T14: *Mean Value Test for Prices*:

$$\min_i(p_i^1/p_i^0 : i = 1, \dots, N) \leq P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \leq \max_i(p_i^1/p_i^0 : i = 1, \dots, N).$$

That is, the price index lies between the minimum price ratio and the maximum price ratio. Since the price index is supposed to be some sort of an average of the  $N$  price ratios,  $p_i^1/p_i^0$ , it seems essential that the price index  $P$  satisfy this test.

The next test is the analogue to T14 applied to quantity indexes:

T15: *Mean Value Test for Quantities*:

$$\min_i(q_i^1/q_i^0 : i = 1, \dots, N) \leq \{V^1/V^0\} / P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \leq \max_i(q_i^1/q_i^0 : i = 1, \dots, N)$$

where  $V^t$  is the period  $t$  value aggregate  $V^t \equiv \sum_{n=1}^N p_n^t q_n^t$  for  $t = 0, 1$ . Using (5.5) to define the quantity index  $Q$  in terms of the price index  $P$ , we see that T15 is equivalent to the following property for the associated quantity index  $Q$ :

$$\min_i(q_i^1/q_i^0 : i = 1, \dots, N) \leq Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \leq \max_i(q_i^1/q_i^0 : i = 1, \dots, N). \quad (5.20)$$

That is, the implicit quantity index  $Q$  defined by  $P$  lies between the minimum and maximum rates of growth  $q_i^1/q_i^0$  of the individual quantities.

In section 5.3, we argued that it was very reasonable to take an average of the Laspeyres and Paasche price indexes as a single “best” measure of overall price change. This point of view can be turned into a test:

T16: *Paasche and Laspeyres Bounding Test*: The price index  $P$  lies between the Laspeyres and Paasche indices,  $P_L$  and  $P_P$ , defined earlier by (5.7) and (5.8) above.

The final four tests are monotonicity tests; i.e., how should the price index  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  change as any component of the two price vectors  $\mathbf{p}^0$  and  $\mathbf{p}^1$  increases or as any component of the two quantity vectors  $\mathbf{q}^0$  and  $\mathbf{q}^1$  increases.

T17: *Monotonicity in Current Prices*:  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) < P(\mathbf{p}^0, \mathbf{p}^2, \mathbf{q}^0, \mathbf{q}^1)$  if  $\mathbf{p}^1 < \mathbf{p}^2$ .

That is, if some period 1 price increases, then the price index must increase, so that  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is increasing in the components of  $\mathbf{p}^1$ . This property was proposed by Eichhorn and Voeller (1976; 23)[171] and it is a very reasonable property for a price index to satisfy.

T18: *Monotonicity in Base Prices*:  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) > P(\mathbf{p}^2, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  if  $\mathbf{p}^0 < \mathbf{p}^2$ .

That is, if any period 0 price increases, then the price index must decrease, so that  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is decreasing in the components of  $\mathbf{p}^0$ . This very reasonable property was also proposed by Eichhorn and Voeller (1976; 23)[171].

T19: *Monotonicity in Current Quantities*: if  $\mathbf{q}^1 < \mathbf{q}^2$ , then

$$\{\sum_{i=1}^N p_i^1 q_i^1 / \sum_{i=1}^N p_i^0 q_i^0\} / P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) < \{\sum_{i=1}^N p_i^1 q_i^2 / \sum_{i=1}^N p_i^0 q_i^0\} / P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^2).$$

T20: *Monotonicity in Base Quantities*: if  $\mathbf{q}^0 < \mathbf{q}^2$ , then

$$\{\sum_{i=1}^N p_i^1 q_i^1 / \sum_{i=1}^N p_i^0 q_i^0\} / P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) > \{\sum_{i=1}^N p_i^1 q_i^1 / \sum_{i=1}^N p_i^0 q_i^2\} / P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^2, \mathbf{q}^1).$$

If we define the implicit quantity index  $Q$  that corresponds to  $P$  using (5.5), we find that T19 translates into the following inequality involving  $Q$ :

$$Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) < Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^2) \text{ if } \mathbf{q}^1 < \mathbf{q}^2. \quad (5.21)$$

That is, if any period 1 quantity increases, then the implicit quantity index  $Q$  that corresponds to the price index  $P$  must increase. Similarly, we find that T20 translates into:

$$Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) > Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^2, \mathbf{q}^1) \text{ if } \mathbf{q}^0 < \mathbf{q}^2. \quad (5.22)$$

That is, if any period 0 quantity increases, then the implicit quantity index  $Q$  must decrease. Tests T19 and T20 are due to Vogt (1980, 70)[383].

The final test is Irving Fisher's (1921; 534)[186] (1922; 72-81)[187] third reversal test (the other two being T9 and T11):

T21: *Factor Reversal Test* (functional form symmetry test):

$$P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) P(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^0, \mathbf{p}^1) = \sum_{i=1}^N p_i^1 q_i^1 / \sum_{i=1}^N p_i^0 q_i^0 = V^1 / V^0.$$

A justification for this test is the following one: if  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is a good functional form for the price index, then if we reverse the roles of prices and quantities,  $P(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^0, \mathbf{p}^1)$  ought to be a good functional form for a quantity index (which seems to be a correct argument) and thus the product of the price index  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  and the quantity index  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = P(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^0, \mathbf{p}^1)$  ought to equal the value ratio,  $V^1/V^0$ . The second part of this argument does not seem to be valid and thus many researchers over the years have objected to the factor reversal test.

It is straightforward to show that the Fisher ideal price index  $P_F$  defined earlier by (5.9) satisfies all 21 tests. Is this the only index number formula that satisfies all of these tests? The answer is yes: Funke and Voeller (1978; 180)[194] showed that the only index number function  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  which satisfies T1 (positivity), T11 (time reversal test), T12 (quantity reversal test) and T21 (factor reversal test) is the Fisher ideal index  $P_F$  defined by (5.9). Diewert (1992; 221)[101] proved a similar result: namely that if  $P$  satisfied T1 and the three reversal tests T11-T13, then  $P$  must equal  $P_F$ . We will provide a proof of Diewert's result.

It is relatively straightforward to show that the Fisher index satisfies all of the above 21 tests. The more difficult part of the proof is to show that it is the *only* index number formula which satisfies these tests. This part of the proof follows from the fact that if  $P$  satisfies the positivity test T1 and the three reversal tests, T11-T13, then  $P$  must equal  $P_F$ . To see this, rearrange the terms in the

statement of test T13 into the following equation:

$$\begin{aligned}
& \left\{ \frac{\sum_{i=1}^N p_i^1 q_i^1}{\sum_{i=1}^N p_i^0 q_i^0} \right\} / \left\{ \frac{\sum_{i=1}^N p_i^0 q_i^1}{\sum_{i=1}^N p_i^1 q_i^0} \right\} \\
& = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) / P(\mathbf{p}^1, \mathbf{p}^0, \mathbf{q}^0, \mathbf{q}^1) \\
& = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) / P(\mathbf{p}^1, \mathbf{p}^0, \mathbf{q}^1, \mathbf{q}^0) \quad \text{using T12, the quantity reversal test} \\
& = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \quad \text{using T11, the time reversal test.} \quad (5.23)
\end{aligned}$$

Now take positive square roots on both sides of (5.23) and we see that the left hand side of the equation is the Fisher index  $P_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  defined by (5.9) and the right hand side is  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ . Thus if  $P$  satisfies T1, T11, T12 and T13, it must equal the Fisher ideal index  $P_F$ .

Thus it seems that from the perspective of the above test approach to index number theory, the Fisher ideal index satisfies more “reasonable” tests than competing indexes and hence can be regarded as “best” from the viewpoint of this perspective.

There is another perspective to the test approach to index number theory. The above approach looked at axioms or tests that pertained to situations where the price index was a function of the two price vectors,  $\mathbf{p}^0$  and  $\mathbf{p}^1$ , and the two matching quantity vectors,  $\mathbf{q}^0$  and  $\mathbf{q}^1$ . In this framework, the two quantity vectors essentially act as weights for the prices. However, there is an alternative framework where the price index, say  $P^*(\mathbf{p}^0, \mathbf{p}^1, \mathbf{e}^0, \mathbf{e}^1)$ , is regarded as a function of the two price vectors,  $\mathbf{p}^0$  and  $\mathbf{p}^1$ , and the two matching *expenditure vectors*,  $\mathbf{e}^0$  and  $\mathbf{e}^1$ .<sup>\*26</sup> An axiomatic approach to the determination of the functional form for indexes of this type is developed in the ILO (2004; 307-309)[249] and the Törnqvist index defined earlier by (5.18) emerges as “best” from the perspective of this second test approach to index number theory. Thus both the Fisher and Törnqvist indexes can be given strong axiomatic justifications.

There is one final important test that should be added to the above list of tests and that is the following *circularity test*<sup>\*27</sup> which involves looking at the prices and quantities that pertain to three periods:

$$\text{T22: } \textit{Circularity Test: } P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) P(\mathbf{p}^1, \mathbf{p}^2, \mathbf{q}^1, \mathbf{q}^2) = P(\mathbf{p}^0, \mathbf{p}^2, \mathbf{q}^0, \mathbf{q}^2).$$

If this test is satisfied, then the rate of price change going from period 0 to 1,  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ , times the rate of price change going from period 1 to 2,  $P(\mathbf{p}^1, \mathbf{p}^2, \mathbf{q}^1, \mathbf{q}^2)$ , is equal to the rate of price change going from period 0 to 2 directly,  $P(\mathbf{p}^0, \mathbf{p}^2, \mathbf{q}^0, \mathbf{q}^2)$ . If there is only one commodity in the aggregate, then the price index  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  just becomes the single price ratio,  $p_1^1/p_1^0$ , and the circularity test T22 becomes the equation  $[p_1^1/p_1^0][p_1^2/p_1^1] = [p_1^2/p_1^0]$ , which is obviously satisfied. The equation in the circularity test illustrates the difference between chained index numbers and fixed base index numbers. The left hand side of T22 uses the *chain principle* to construct the overall inflation between periods 0 and 2 whereas the right hand side uses the *fixed base principle* to construct an estimate of the overall price change between periods 0 and 1.<sup>\*28</sup>

It would be good if our preferred index number formulae, the Fisher, Walsh and Törnqvist indexes ( $P_F, P_W$  and  $P_T$ ), satisfied the circularity test but unfortunately, none of these indexes satisfy T22. An interesting problem is to determine exactly what class of indexes does satisfy the circularity test.

<sup>\*26</sup> Component  $n$  of the period  $t$  expenditure vector  $\mathbf{e}^t$  is defined as  $e_n^t \equiv p_n^t q_n^t$  for  $n = 1, \dots, N$  and  $t = 0, 1$ . Thus if the price components  $p_n^t$  are known, then a knowledge of either the quantity components  $q_n^t$  or the expenditure components  $e_n^t$  will determine prices, quantities and expenditures in both periods.

<sup>\*27</sup> The test name is due to Fisher (1922; 413)[187] and the concept was originally due to Westergaard (1890; 218-219)[396].

<sup>\*28</sup> Thus when the chain principle is used, the price index  $P(\mathbf{p}^t, \mathbf{p}^{t+1}, \mathbf{q}^t, \mathbf{q}^{t+1})$  is used to update the period  $t$  index level to construct the period  $t+1$  index level, whereas the fixed base system constructs the period  $t+1$  index level relative to period 0 directly as  $P(\mathbf{p}^0, \mathbf{p}^{t+1}, \mathbf{q}^0, \mathbf{q}^{t+1})$ , where the period 0 level is set equal to 1. Fisher (1911; 203)[185] introduced this fixed base and chain terminology. The concept of chaining is due to Lehr (1885)[288] and Marshall (1887; 373)[303].

The following proposition, due essentially to Eichhorn (1978; 167-168)[170], helps to answer this question.

**Proposition** Assume that the index number formula  $P$  satisfies the following tests: T1 (positivity), T2 (continuity), T3 (identity), T5 (proportionality in current prices), T10 (commensurability) and T17 (monotonicity in current prices) in addition to the circularity test above. Then  $P$  must have the following functional form due originally to Konüs and Byushgens<sup>\*29</sup> (1926; 163-166)[282]:<sup>\*30</sup>

$$P_{KB}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv \prod_{i=1}^N [p_i^1/p_i^0]^{\alpha_i} \quad (5.24)$$

where the  $N$  constants  $\alpha_i$  satisfy the following restrictions:

$$\sum_{i=1}^N \alpha_i = 1 \text{ and } \alpha_i > 0 \text{ for } i = 1, \dots, N. \quad (5.25)$$

**Proof.** Rewrite the circularity test T22 in the following form:

$$P(\mathbf{p}^*, \mathbf{p}, \mathbf{q}^*, \mathbf{q}) = P(\mathbf{p}^*, \mathbf{p}^0, \mathbf{q}^*, \mathbf{q}^0)P(\mathbf{p}^0, \mathbf{p}, \mathbf{q}^0, \mathbf{q}). \quad (5.26)$$

Using T1, we can rewrite (5.26) as follows:

$$P(\mathbf{p}^0, \mathbf{p}, \mathbf{q}^0, \mathbf{q}) = P(\mathbf{p}^*, \mathbf{p}, \mathbf{q}^*, \mathbf{q})/P(\mathbf{p}^*, \mathbf{p}^0, \mathbf{q}^*, \mathbf{q}^0). \quad (5.27)$$

Now hold  $\mathbf{p}^*$  and  $\mathbf{q}^*$  constant at some fixed values and define the function  $f(\mathbf{p}, \mathbf{q})$  as follows:

$$f(\mathbf{p}, \mathbf{q}) \equiv P(\mathbf{p}^*, \mathbf{p}, \mathbf{q}^*, \mathbf{q}) > 0 \quad \text{for all } \mathbf{p} \gg \mathbf{0}_N \text{ and } \mathbf{q} \gg \mathbf{0}_N \quad (5.28)$$

where the positivity of  $f(\mathbf{p}, \mathbf{q})$  follows from T1. Substituting definition (5.28) back into (5.27) gives us the following representation for  $P(\mathbf{p}^0, \mathbf{p}, \mathbf{q}^0, \mathbf{q})$ :

$$P(\mathbf{p}^0, \mathbf{p}, \mathbf{q}^0, \mathbf{q}) = f(\mathbf{p}, \mathbf{q})/f(\mathbf{p}^0, \mathbf{q}^0). \quad (5.29)$$

Now let  $\mathbf{p}^0 = \mathbf{p}$  in (5.29) and apply the identity test T3 to the resulting equation. We obtain:

$$1 = P(\mathbf{p}, \mathbf{p}, \mathbf{q}^0, \mathbf{q}) = f(\mathbf{p}, \mathbf{q})/f(\mathbf{p}, \mathbf{q}^0); \quad \mathbf{p} \gg \mathbf{0}_N; \mathbf{q} \gg \mathbf{0}_N; \mathbf{q}^0 \gg \mathbf{0}_N. \quad (5.30)$$

Define the function  $g(\mathbf{p})$  as

$$g(\mathbf{p}) \equiv f(\mathbf{p}, \mathbf{1}_N) > 0 \quad \mathbf{p} \gg \mathbf{0}_N. \quad (5.31)$$

Now set  $\mathbf{q}^0$  in (5.30) equal to a vector of ones,  $\mathbf{1}_N$ , and (5.30) becomes:

$$\begin{aligned} f(\mathbf{p}, \mathbf{q}) &= f(\mathbf{p}, \mathbf{1}_N) \\ &= g(\mathbf{p}) \quad \text{using definition (5.31)}. \end{aligned} \quad (5.32)$$

Thus  $f(\mathbf{p}, \mathbf{q})$  cannot depend on  $\mathbf{q}$ . Now substitute (5.32) back into (5.29) and we find that  $P$  must have the following representation if  $P$  satisfies the circularity test and the tests T1 and T3:

$$P(\mathbf{p}^0, \mathbf{p}, \mathbf{q}^0, \mathbf{q}) = g(\mathbf{p})/g(\mathbf{p}^0); \quad \mathbf{p} \gg \mathbf{0}_N; \mathbf{p}^0 \gg \mathbf{0}_N; \mathbf{q} \gg \mathbf{0}_N; \mathbf{q}^0 \gg \mathbf{0}_N. \quad (5.33)$$

<sup>\*29</sup> Konüs and Byushgens show that the index defined by (5.24) is exact for Cobb-Douglas (1928)[59] preferences; see also Pollak (1989; 23)[332]. The concept of an exact index number formula will be explained when we study the economic approach to index number theory in Part II (next chapter).

<sup>\*30</sup> See also Eichhorn (1978; 167-168)[170] and Vogt and Barta (1997; 47)[384]. Proofs of related results can be found in Funke, Hacker and Voeller (1979)[193] and Balk (1995)[19].

Now apply the commensurability test, T10, to the  $P$  that is defined by (5.33) where we set  $\alpha_i = (p_i^0)^{-1}$  for  $i = 1, \dots, N$ . Using the representation for  $P$  given by (5.33), we find that  $g$  must satisfy the following functional equation:

$$g(\mathbf{p}^1)/g(\mathbf{p}^0) = g(p_1^1/p_1^0, p_2^1/p_2^0, \dots, p_N^1/p_N^0)/g(\mathbf{1}_N); \quad \mathbf{p}^0 \gg \mathbf{0}_N; \mathbf{p}^1 \gg \mathbf{0}_N. \quad (5.34)$$

Define  $h(\mathbf{p})$  as follows:

$$h(\mathbf{p}) \equiv g(\mathbf{p})/g(\mathbf{1}_N) > 0; \quad \mathbf{p} \gg \mathbf{0}_N \quad (5.35)$$

where the positivity of  $h$  follows from the positivity of  $g$ . Using definition (5.35), we have:

$$\begin{aligned} h(p_1^1/p_1^0, p_2^1/p_2^0, \dots, p_N^1/p_N^0) &= g(p_1^1/p_1^0, p_2^1/p_2^0, \dots, p_N^1/p_N^0)/g(\mathbf{1}_N) \quad \mathbf{p}^0 \gg \mathbf{0}_N; \mathbf{p}^1 \gg \mathbf{0}_N \\ &= g(\mathbf{p}^1)/g(\mathbf{p}^0) \quad \text{using (5.34)} \\ &= [g(\mathbf{p}^1)/g(\mathbf{1}_N)]/[g(\mathbf{p}^0)/g(\mathbf{1}_N)] \quad \text{using T1} \\ &= h(\mathbf{p}^1)/h(\mathbf{p}^0) \quad \text{using (5.35) twice.} \end{aligned} \quad (5.36)$$

Thus  $h$  must satisfy the following functional equation:

$$h(\mathbf{p}^0)h(p_1^1/p_1^0, p_2^1/p_2^0, \dots, p_N^1/p_N^0) = h(\mathbf{p}^1); \quad \mathbf{p}^0 \gg \mathbf{0}_N; \mathbf{p}^1 \gg \mathbf{0}_N. \quad (5.37)$$

Define the vector  $\mathbf{x}$  as the vector  $\mathbf{p}^0$  and the vector  $\mathbf{y}$  as  $p_1^1/p_1^0, p_2^1/p_2^0, \dots, p_N^1/p_N^0$ . Hence the product of the  $i$ th components of  $\mathbf{x}$  and  $\mathbf{y}$  is equal to the  $i$ th component of the vector  $\mathbf{p}^1$  and it can be seen that the functional equation (5.37) is equivalent to the following functional equation:

$$h(x_1y_1, x_2y_2, \dots, x_Ny_N) = h(x_1, x_2, \dots, x_N)h(y_1, y_2, \dots, y_N); \quad \mathbf{x} \gg \mathbf{0}_N; \mathbf{y} \gg \mathbf{0}_N. \quad (5.38)$$

Equation (5.38) becomes the following equation if we allow  $x_1$  and  $y_1$  to vary freely but fix all  $x_i$  and  $y_i$  at 1 for  $i = 2, 3, \dots, N$ :

$$h(x_1y_1, 1, \dots, 1) = h(x_1, 1, \dots, 1)h(y_1, 1, \dots, 1); \quad x_1 > 0; y_1 > 0. \quad (5.39)$$

But (5.39) is an example of *Cauchy's* (1821)[48] *fourth functional equation*.<sup>\*31</sup> Using the T1 (positivity) and T2 (continuity) properties of  $P$ , which carry over to  $h$ , we see that the solution to (5.39) is:

$$h(x_1, 1, \dots, 1) = x_1^{c(1)} \quad (5.40)$$

where  $c(1)$  is an arbitrary constant. In a similar fashion, (5.38) becomes the following equation if we allow  $x_2$  and  $y_2$  to vary freely but fix all other  $x_i$  and  $y_i$  at 1:

$$h(1, x_2y_2, 1, \dots, 1) = h(1, x_2, 1, \dots, 1)h(1, y_2, 1, \dots, 1); \quad x_2 > 0; y_2 > 0. \quad (5.41)$$

The solution to (5.41) is:

$$h(1, x_2, 1, \dots, 1) = x_2^{c(2)} \quad (5.42)$$

where  $c(2)$  is an arbitrary constant. In a similar fashion, we find that

$$h(1, 1, x_3, 1, \dots, 1) = x_3^{c(3)}; \dots; h(1, 1, \dots, 1, x_N) = x_N^{c(N)} \quad (5.43)$$

<sup>\*31</sup> See Eichhorn (1978)[170] for material on Cauchy's four fundamental functional equations.

where the  $c(i)$  are arbitrary constants. Using (5.38) repeatedly, we can show:

$$\begin{aligned}
h(x_1, x_2, \dots, x_N) &= h(x_1, 1, \dots, 1)h(1, x_2, \dots, x_N) \\
&= h(x_1, 1, \dots, 1)h(1, x_2, 1, \dots, 1)h(1, 1, x_3, \dots, x_N) \\
&= h(x_1, 1, \dots, 1)h(1, x_2, 1, \dots, 1)h(1, 1, x_3, 1, \dots, 1)h(1, 1, 1, x_4, \dots, x_N) \\
&\dots \\
&= h(x_1, 1, \dots, 1)h(1, x_2, 1, \dots, 1)h(1, 1, x_3, 1, \dots, 1) \cdots h(1, 1, 1, \dots, 1, x_N) \\
&= \prod_{i=1}^N x_i^{c(i)} \quad \text{using (5.40), (5.42) and (5.43)}. \tag{5.44}
\end{aligned}$$

Thus we have determined the functional form for the function  $h$ . Now use (5.35) to determine the function  $g(\mathbf{p})$  in terms of  $h(\mathbf{p})$ :

$$\begin{aligned}
g(\mathbf{p}) &= g(\mathbf{1}_N)h(\mathbf{p}) \\
&= g(\mathbf{1}_N)\prod_{i=1}^N p_i^{c(i)}. \tag{5.45}
\end{aligned}$$

Using (5.33), we can express  $P$  in terms of  $g$  as follows:

$$\begin{aligned}
P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) &= g(\mathbf{p}^1)/g(\mathbf{p}^0) \\
&= g(\mathbf{1}_N)\prod_{i=1}^N (p_i^1)^{c(i)} / g(\mathbf{1}_N)\prod_{i=1}^N (p_i^0)^{c(i)} \quad \text{using (5.45)} \\
&= \prod_{i=1}^N (p_i^1/p_i^0)^{c(i)}. \tag{5.46}
\end{aligned}$$

Now apply test T5, proportionality in current prices, to the  $P$  defined by (5.46). It is easy to see that this test implies that the constants  $c(i)$  must sum to 1.

Finally, apply test T17, monotonicity in current prices, to conclude that the constants  $c(i)$  must be positive. Hence we can set the  $c(i)$  equal to the  $\alpha_i$  and we have proved the Proposition. ■

Thus under fairly weak regularity conditions, *the only price index satisfying the circularity test is a weighted geometric average of all the individual price ratios*, the weights being constant through time. This is a somewhat discouraging result!

Looking at the above proof, it is interesting to note that we arrive at the representation for the price index given by (5.33) using only the circularity test T22 and the two tests, T1 (Positivity) and T3 (Identity), which are rather weak tests. The representation (5.33) can be rewritten as follows:

$$P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = g(\mathbf{p}^1)/g(\mathbf{p}^0) \tag{5.47}$$

where  $g(\mathbf{p})$  is a continuous function of  $\mathbf{p}$ , assuming that  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  satisfies T2 (Continuity). Looking at (5.47), *it can be seen that the index number formula does not depend on the quantity vectors  $\mathbf{q}^0$  and  $\mathbf{q}^1$* . Thus any weighting of the prices (by quantities or expenditure shares) must be constant or nonexistent. Thus if the same formula is applied over long periods of time where relative quantities or expenditures are changing, the formula must lose its relevance for at least part of the sample period.

Fisher realized this difficulty with the circularity test as can be seen from the following quotation:

“The only formulae which conform perfectly to the circular test are index numbers which have *constant weights*; i.e., weights which are the same for all sides of the ‘triangle’ or segments of the ‘circle’; i.e., for every pair of times or places compared. ... But, clearly, constant weighting is not theoretically correct. If we compare 1913 with 1914, we need one set of weights; if we compare 1913 with 1915 we need, theoretically at least, another set of weights. In the former

case we need weights involving the quantities of the two years concerned, 1913 and 1914; in the second case we need weights involving the (somewhat different) quantities of the two years, 1913 and 1915. We cannot justify using the same weights for comparing the price level of 1913, not only with 1914 and 1915, but with 1969, 1776, 1492 and the times of Diocletian, Rameses II, and the Stone Age!” Irving Fisher (1922; 274-275)[187].

Fisher did not have at his disposal a knowledge of functional equations so he was not able to prove the above result but his intuition was quite correct.

Thus far, the results of sections 5.3-5.5 have suggested that the “best” bilateral index number formulae for the price index are the Fisher or the Walsh (fixed basket approaches), the Törnqvist Theil (the stochastic or descriptive statistics approach) or the Fisher (the test approach). None of these indexes satisfy the circularity test and so there will be a difference if fixed base or chained indexes are used in empirical applications. The question now arises: should the sequence of index values be computed using fixed base indexes or chained indexes? We will address this question in the following section.

## 5.6 Fixed Base versus Chained Indexes

Fixed base indexes cannot be used for long periods of time in today’s dynamic economy where new commodities appear and older ones become obsolete. Under these conditions, it becomes increasingly difficult to match commodity prices over long periods of time and index number theory is dependent on a high degree of matching of the prices between the two periods being compared. However, this possible lack of matching does not rule out using fixed base indexes for shorter periods of time, say over a year or two.

The main advantage of using chained indexes is that if prices and quantities are trending relatively smoothly, chaining will reduce the spread between the Paasche and Laspeyres indexes.\*<sup>32</sup> These two indexes each provide an asymmetric perspective on the amount of price change that has occurred between the two periods under consideration and it could be expected that a single point estimate of the aggregate price change should lie between these two estimates. Thus the use of either a chained Paasche or Laspeyres index will usually lead to a smaller difference between the two and hence to estimates that are closer to the “truth”. Since annual data generally has smooth trends, the use of chained indexes is generally appropriate at this level of aggregation; see Hill (1988)[238].

However, the story is different at subannual levels; i.e., if the index is to be produced at monthly or quarterly frequencies. Hill (1993; 388)[239], drawing on the earlier research of Szulc (1983)[368] and Hill (1988; 136-137)[238], noted that it is not appropriate to use the chain system when prices oscillate or “bounce” to use Szulc’s (1983; 548)[368] term. This phenomenon can occur in the context of regular seasonal fluctuations or in the context of sales. The *price bouncing problem* or the problem of *chain drift* can be illustrated if we make use of the following test due to Walsh (1901; 389)[389], (1921b; 540)[392] (1924; 506)[393].\*<sup>33</sup>

T23: *Multiperiod Identity Test*:  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)P(\mathbf{p}^1, \mathbf{p}^2, \mathbf{q}^1, \mathbf{q}^2)P(\mathbf{p}^2, \mathbf{p}^0, \mathbf{q}^2, \mathbf{q}^0) = 1$ .

Thus price change is calculated over consecutive periods but an artificial final period is introduced where the prices and quantities revert back to the prices and quantities in the very first period. The Walsh test T23 asks that the product of all of these price changes should equal unity. If prices have no definite trends but are simply bouncing up and down in a range, then the above test can be used to evaluate the amount of chain drift that occurs if chained indexes are used under these conditions.

\*<sup>32</sup> See Diewert (1978; 895)[85] and Hill (1988)[238] (1993; 387-388)[239]. Chaining under these conditions will also reduce the spread between fixed base and chained indexes using  $P_F$ ,  $P_W$  or  $P_T$  as the basic bilateral formula.

\*<sup>33</sup> This is Diewert’s (1993; 40)[109] term for the test. Walsh did not limit himself to just three periods as in T23; he considered an indefinite number of periods. If tests T3 and T22 are satisfied, then T23 will also be satisfied.

*Chain drift* occurs when an index does not return to unity when prices in the current period return to their levels in the base period; see the ILO (2004; 445)[249]. Fixed base indexes operating under these conditions will not be subject to chain drift.

It is possible to be a bit more precise under what conditions one should chain or not chain. Basically, one should chain if the prices and quantities pertaining to adjacent periods are *more similar* than the prices and quantities of more distant periods, since this strategy will lead to a narrowing of the spread between the Paasche and Laspeyres indexes at each link. Of course, one needs a measure of how similar are the prices and quantities pertaining to two periods. A practical problem with this *similarity linking approach* is: exactly how should the measure of price or quantity similarity be measured? \*<sup>34</sup> For *annual* time series data, it turns out that for various “reasonable” similarity measures, chained indexes are generally consistent with the similarity approach to linking observations. However, for subannual data, it is generally better to use fixed base indexes in order to eliminate the problem of chain drift.

We conclude this subsection with a discussion on how well our best indexes,  $P_F$ ,  $P_W$  and  $P_T$  defined by (5.9), (5.13) and (5.18) above, satisfy the circularity test, T22. Fisher (1922; 277)[187] found that for his annual data set, the Fisher ideal index  $P_F$  satisfied circularity to a reasonably high degree of approximation. It turns out that this result generally holds using annual data for  $P_W$  and  $P_T$  as well. It is possible to give a theoretical explanation for the approximate satisfaction of the circularity test for these three indexes. Alterman, Diewert and Feenstra (1999; 61)[7] showed that if the logarithmic price ratios  $\ln(p_n^t/p_n^{t-1})$  trend linearly with time  $t$  and the expenditure shares  $s_i^t$  also trend linearly with time, then the Törnqvist index  $P_T$  will satisfy the circularity test *exactly*. \*<sup>35</sup> Since many economic time series on prices and quantities satisfy these assumptions approximately, the above exactness result will imply that the Törnqvist index  $P_T$  will satisfy the circularity test approximately. But Diewert (1978; 888)[85] showed that  $P_T$ ,  $P_F$  and  $P_W$  numerically approximate each other to the second order around an equal price and quantity point \*<sup>36</sup> and so these three indexes will generally be very close to each other using annual time series data. Hence since  $P_T$  will generally satisfy the circularity test to some degree of approximation,  $P_F$  and  $P_W$  will also satisfy circularity approximately in the time series context using annual data. Thus for *annual* economic time series,  $P_F$ ,  $P_T$  and  $P_W$  will generally satisfy the circularity test to a high enough degree of approximation so that it will not matter whether we use the fixed base or chain principle. However, this same conclusion does *not* hold for *subannual* data that has substantial period to period fluctuations in prices. For fluctuating subannual data, chained indexes can give very unsatisfactory results; i.e., Walsh’s multiperiod identity test can be far from being satisfied. \*<sup>37</sup> Under these conditions, fixed base indexes or multilateral methods should be used. \*<sup>38</sup>

**Problem 1** Let  $a$  and  $b$  be positive numbers and define the following *means* of  $a$  and  $b$ :

\*<sup>34</sup> This similarity approach to linking bilateral comparisons into a complete set of comparisons across all observations has been pioneered by Robert Hill (1999a)[231] (1999b)[232] (2001)[233] (2004)[234] (2009)[236]. For an axiomatic approach to similarity measures, see Diewert (2009)[133].

\*<sup>35</sup> This exactness result can be extended to cover the case when there are monthly proportional variations in prices and the expenditure shares have constant seasonal effects in addition to linear trends; see Alterman, Diewert and Feenstra (1999; 65)[7].

\*<sup>36</sup> See problems 8 and 9 below.

\*<sup>37</sup> See Szulc (1983)[368] (1987)[369], Feenstra and Shapiro (2003)[177], Ivancic, Diewert and Fox (2011)[251], de Haan and van der Grient (2011)[69], de Haan and Krsinich (2012)[70] and Diewert (2013)[135] for evidence of chain drift using subannual data.

\*<sup>38</sup> See Szulc (1983)[368] (1987)[369] and Hill (1988)[238] on this point. For solutions to the chain drift problem using subannual data, see and Ivancic, Diewert and Fox (2009)[250] (2011)[251], de Haan and van der Grient (2011)[69] and de Haan and Krsinich (2012)[70].

- (a)  $m_A(a, b) \equiv (1/2)a + (1/2)b$  (the arithmetic mean);  
 (b)  $m_G(a, b) \equiv (ab)^{1/2}$  (the geometric mean);  
 (c)  $m_H(a, b) \equiv [(1/2)a^{-1} + (1/2)b^{-1}]^{-1}$  (the harmonic mean).

Using elementary arguments, prove that:

- (d)  $m_H(a, b) \leq m_G(a, b) \leq m_A(a, b)$ .

Under what conditions on  $a$  and  $b$  will strict inequalities hold in (d)?

**Problem 2** Instead of taking the arithmetic average of the expenditure shares in Theil's weighted stochastic approach, consider taking the geometric or harmonic average of these shares. Check whether the resulting indexes,  $P_{TG}$  and  $P_{TH}$ , and the Theil index  $P_T$  satisfy Tests T3, T4, T5, T8, T10 and T11. Provide proofs.

**Problem 3** Prove that the Fisher index satisfies Tests T3-T18.

**Problem 4** Which tests from Tests T3-T18 does the Walsh index  $P_W$  satisfy? Provide proofs.

**Problem 5** If the price index  $P$  satisfies Test T4 (the Fixed Basket Test) and  $P$  and  $Q$  jointly satisfy the product test, (5.5) above, then show<sup>\*39</sup> that  $Q$  must satisfy the identity test  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}, \mathbf{q}) = 1$  for all strictly positive vectors  $\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}$ . This *constant quantities test* for  $Q$  is somewhat controversial since  $\mathbf{p}^0$  and  $\mathbf{p}^1$  are allowed to be different.

**Problem 6** Consider a test where the implicit quantity index  $Q$  that corresponds to  $P$  via (5.5) is to lie between the Laspeyres and Paasche quantity indexes,  $Q_L$  and  $Q_P$ , defined by (a) and (b) below:

- (a)  $Q_L \equiv \mathbf{p}^0 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0$ ;  
 (b)  $Q_P \equiv \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^1 \cdot \mathbf{q}^0$ .

Show that the resulting test turns out to be *equivalent* to test T16 on  $P$ .

**Problem 7** Consider the following bilateral price index:

- (a)  $P_\alpha(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv \mathbf{p}^1 \cdot \boldsymbol{\alpha} / \mathbf{p}^0 \cdot \boldsymbol{\alpha}$

where  $\boldsymbol{\alpha} \gg \mathbf{0}_N$  is vector of positive constants.

(i) Which of the tests T3-T18 and T22 does  $P_\alpha$  satisfy? (Provide proofs).

(ii) Note that the index defined by (a) is a simple index number formula that satisfies the Circularity Test, T22, but yet it is not equal to the Cobb-Douglas index defined by (5.46) above. Explain why this equality does not occur.

*Comment:* If the vector  $\boldsymbol{\alpha}$  is chosen to be an annual quantity vector pertaining to some year prior to period 0, then  $P_\alpha$  becomes the Lowe index, which is widely used by statistical agencies as their target CPI index.

**Problem 8** Consider the Laspeyres, Paasche, Fisher, Törnqvist and Walsh price indexes,  $P_L, P_P, P_F, P_T$  and  $P_W$  as functions of the four sets of variables,  $\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1$ . Show that all of the  $4N$  first order partial derivatives of each of these 5 indexes are equal when evaluated at a point where the two price vectors are equal (so that  $\mathbf{p}^0 = \mathbf{p}^1 \equiv \mathbf{p}$ ) and where the two quantity vectors are equal (so that  $\mathbf{q}^0 = \mathbf{q}^1 \equiv \mathbf{q}$ ); i.e., show that

$$\nabla P_L(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = \nabla P_P(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = \nabla P_F(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = \nabla P_T(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = \nabla P_W(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}). \quad (\text{a})$$

<sup>\*39</sup> See Vogt (1980; 70)[383].

*Hint:* You need to calculate 20 vectors of first order partial derivatives and show that they are equal when evaluated at an equal price and equal quantity point. We calculate these derivatives for the case of the Laspeyres index to show what is required.

$$P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \mathbf{p}^{1T} \mathbf{q}^0 / \mathbf{p}^{0T} \mathbf{q}^0. \quad (\text{b})$$

Differentiate  $P_L$  with respect to the components of  $\mathbf{p}^0$  and get the following vector of first order partial derivatives:

$$\nabla_{\mathbf{p}^0} P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = -\mathbf{p}^{1T} \mathbf{q}^0 (\mathbf{p}^{0T} \mathbf{q}^0)^{-2} \mathbf{q}^0. \quad (\text{c})$$

Now evaluate these derivatives at  $\mathbf{p}^0 = \mathbf{p}^1 \equiv \mathbf{p}$  and  $\mathbf{q}^0 = \mathbf{q}^1 \equiv \mathbf{q}$  and we get the following expression:

$$\nabla_{\mathbf{p}^0} P_L(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = -\mathbf{q} / \mathbf{p}^T \mathbf{q} \quad (\text{d})$$

which is the answer for the first block of derivatives. Now differentiate  $P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  with respect to the components of  $\mathbf{p}^1$  and get the following vector of first order derivatives:

$$\nabla_{\mathbf{p}^1} P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \mathbf{q}^0 / \mathbf{p}^{0T} \mathbf{q}^0 = \mathbf{q} / \mathbf{p}^T \mathbf{q} \quad \text{when } \mathbf{p}^0 = \mathbf{p}^1 \equiv \mathbf{p} \text{ and } \mathbf{q}^0 = \mathbf{q}^1 \equiv \mathbf{q}. \quad (\text{e})$$

Similarly:

$$\begin{aligned} \nabla_{\mathbf{q}^0} P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) &= [\mathbf{p}^0 / \mathbf{p}^{0T} \mathbf{q}^0] - \mathbf{p}^{1T} \mathbf{q}^0 (\mathbf{p}^{0T} \mathbf{q}^0)^{-2} \mathbf{p}^0 = \mathbf{0}_N \\ &\quad \text{when } \mathbf{p}^0 = \mathbf{p}^1 \equiv \mathbf{p} \text{ and } \mathbf{q}^0 = \mathbf{q}^1 \equiv \mathbf{q}; \end{aligned} \quad (\text{f})$$

$$\nabla_{\mathbf{q}^1} P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \mathbf{0}_N. \quad (\text{g})$$

Calculating the derivatives of the Törnqvist Theil index is much more difficult. Note that the log of  $P_T$  is defined as

$$\ln P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv (1/2) \sum_{n=1}^N (s_n^0 + s_n^1) \ln(p_n^1 / p_n^0). \quad (\text{h})$$

Letting  $z$  be equal to any one of the components in the vectors  $\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0$  and  $\mathbf{q}^1$ , we have:

$$\partial P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) / \partial z = P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \partial \ln P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) / \partial z. \quad (\text{i})$$

Now let  $z = p_n^0$  and use formula (i) above along with definition (h) (and remembering that  $s_n^t \equiv p_n^t q_n^t / \mathbf{p}^t \cdot \mathbf{q}^t$  for  $t = 0, 1$  and  $n = 1, \dots, N$ ):

$$\begin{aligned} \partial P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) / \partial p_n^0 &= P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \partial \ln P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) / \partial p_n^0 \\ &= P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) (1/2) [q_n^0 (\mathbf{p}^{0T} \mathbf{q}^0)^{-1}] \ln(p_n^1 / p_n^0) - P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) (1/2) [s_n^0 + s_n^1] (p_n^0)^{-1} \\ &\quad - P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) (1/2) [\sum_{j=1}^N p_j^0 q_j^0 q_j^0 (\mathbf{p}^{0T} \mathbf{q}^0)^{-2}] \ln(p_j^1 / p_j^0) \\ &= 0 - 1(1/2) [(p_n q_n / \mathbf{p}^T \mathbf{q}) + (p_n q_n / \mathbf{p}^T \mathbf{q})] (p_n)^{-1} - 0 \quad \text{when } \mathbf{p}^0 = \mathbf{p}^1 \equiv \mathbf{p} \text{ and } \mathbf{q}^0 = \mathbf{q}^1 \equiv \mathbf{q} \\ &= -q_n / \mathbf{p}^T \mathbf{q} \quad \text{for } n = 1, \dots, N. \end{aligned} \quad (\text{j})$$

Thus equations (j) are equivalent to the equations  $\nabla_{\mathbf{p}^0} P_T(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = -\mathbf{q} / \mathbf{p}^T \mathbf{q}$  which is the same result that we obtained for the Laspeyres price index in equations (d) above.

*Comment:* It is easy to show that

$$P_L(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = P_P(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = P_F(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = P_T(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = P_W(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = 1. \quad (\text{k})$$

Equations (a) and (k) show that the Laspeyres, Paasche, Fisher, Törnqvist and Walsh indexes all approximate each other to the first order around an equal price and quantity point.

**Problem 9** Now consider forming second order Taylor series approximations to the 5 indexes defined in the previous problem.

(a) Show that  $\nabla^2 P_L(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) \neq \nabla^2 P_P(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q})$  and hence the Laspeyres and Paasche indexes do *not* approximate each other to the second order around an equal price and quantity point; i.e., their  $4N \times 4N$  matrices of second order partial derivatives are not all equal when evaluated at an equal price and quantity point.

(b) Show that  $\nabla^2 P_L(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) \neq \nabla^2 P_F(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q})$  and hence the Laspeyres and Fisher indexes do *not* approximate each other to the second order around an equal price and quantity point.

(c) Show that  $\nabla^2 P_P(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) \neq \nabla^2 P_F(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q})$  and hence the Paasche and Fisher indexes do *not* approximate each other to the second order around an equal price and quantity point.

(d) Show that  $\nabla^2 P_F(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = \nabla^2 P_W(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q})$  and hence the Fisher and Walsh indexes *do* approximate each other to the second order around an equal price and quantity point.

(e) Show that  $\nabla^2 P_F(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q}) = \nabla^2 P_T(\mathbf{p}, \mathbf{p}, \mathbf{q}, \mathbf{q})$  and hence the Fisher and Törnqvist indexes *do* approximate each other to the second order around an equal price and quantity point.

*Comment:* Problems 8 and 9 show that the Fisher, Törnqvist and Walsh price indexes all approximate each other to the second order around an equal price and quantity point and hence these indexes are likely to be numerically very close to each other provided prices and quantities do not change “too much” between the two periods under consideration. These problems also show that the Paasche and Laspeyres indexes do not approximate the other three indexes to the second order around an equal price and quantity point.

**Problem 10** Let  $\mathbf{q}$  and  $\mathbf{b}$  be  $N$  dimensional column vectors, let  $a$  be a scalar and let  $\mathbf{C}$  be an  $N \times N$  symmetric matrix. Define the quadratic function of  $\mathbf{q}$ ,  $f(\mathbf{q})$  as follows:

(a)  $f(\mathbf{q}) \equiv a + \mathbf{b}^T \mathbf{q} + (1/2) \mathbf{q}^T \mathbf{C} \mathbf{q}$ .

The second order Taylor series approximation to  $f(\mathbf{q})$  around the point  $\mathbf{q}^0$  is defined as:

(b)  $F(\mathbf{q}) \equiv f(\mathbf{q}^0) + \nabla f(\mathbf{q}^0)^T (\mathbf{q} - \mathbf{q}^0) + (1/2) (\mathbf{q} - \mathbf{q}^0)^T \nabla^2 f(\mathbf{q}^0) (\mathbf{q} - \mathbf{q}^0)$ .

(c) Show that  $f(\mathbf{q}) = F(\mathbf{q})$  for all  $\mathbf{q}$  when  $f(\mathbf{q})$  is defined by (a).

(d) Show that the following identity holds if  $f(\mathbf{q})$  is defined by (a):

(e)  $f(\mathbf{q}^1) - f(\mathbf{q}^0) = (1/2) [\nabla f(\mathbf{q}^0) + \nabla f(\mathbf{q}^1)]^T [\mathbf{q}^1 - \mathbf{q}^0]$  for all  $\mathbf{q}^0$  and  $\mathbf{q}^1$ .

(f) Consider the following two first order approximations:

(g)  $f(\mathbf{q}^1) - f(\mathbf{q}^0) \approx \nabla f(\mathbf{q}^0)^T [\mathbf{q}^1 - \mathbf{q}^0]$  and

(h)  $f(\mathbf{q}^0) - f(\mathbf{q}^1) \approx \nabla f(\mathbf{q}^1)^T [\mathbf{q}^0 - \mathbf{q}^1]$ .

(i) Show that the right hand side of (e) is related to the two first order approximations in (g) and (h).

## 5.7 References

- Alterman, W.F., W.E. Diewert and R.C. Feenstra (1999), *International Trade Price Indexes and Seasonal Commodities*, Bureau of Labor Statistics, Washington D.C.
- Balk, B.M. (1995), “Axiomatic Price Index Theory: A Survey”, *International Statistical Review* 63, 69-93.

- Balk, B.M. (2008), *Price and Quantity Index Numbers*, New York: Cambridge University Press.
- Bowley, A.L. (1901), *Elements of Statistics*, Westminster: P.S. King and Son.
- Bowley, A.L. (1919), "The Measurement of Changes in the Cost of Living", *Journal of the Royal Statistical Society* 82, 343-372.
- Carli, Gian-Rinaldo, (1804), "Del valore e della proporzione de' metalli monetati", pp. 297-366 in *Scrittori classici italiani di economia politica*, Volume 13, Milano: G.G. Destefanis (originally published in 1764).
- Cauchy, A.L. (1821), *Cours d'analyse de l'École Polytechnique*, Volume 1, *Analyse algébrique*, Paris.
- Clements, K.W., H.Y. Izan and E.A. Selvanathan (2006), "Stochastic Index Numbers: A Review", *International Statistical Review* 74, 235-270.
- Cobb, C. and P.H. Douglas (1928), "A Theory of Production", *American Economic Review* 18, 139-165.
- Davies, G.R. (1924), "The Problem of a Standard Index Number Formula", *Journal of the American Statistical Association* 19, 180-188.
- Davies, G.R. (1932), "Index Numbers in Mathematical Economics", *Journal of the American Statistical Association* 27, 58-64.
- de Haan, J. and H.A. van der Grient (2011), "Eliminating Chain drift in Price Indexes Based on Scanner Data", *Journal of Econometrics* 161, 36-46.
- de Haan, J. and F. Krsinich (2012), "The Treatment of Unmatched Items in Rolling Year GEKS Price Indexes: Evidence from New Zealand Scanner Data", paper presented at the Meeting of Groups of Experts on Consumer Price Indices Organized jointly by UNECE and ILO at the United Nations Palais des Nations, Geneva Switzerland, May 30-June 1, 2012.
- Diewert, W.E. (1974), "Applications of Duality Theory", pp. 106-171 in M.D. Intriligator and D.A. Kendrick (ed.), *Frontiers of Quantitative Economics*, Vol. II, Amsterdam: North-Holland.
- Diewert, W.E. (1978), "Superlative Index Numbers and Consistency in Aggregation", *Econometrica* 46, 883-900.
- Diewert, W.E. (1992), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *Journal of Productivity Analysis* 3, 211-248.
- Diewert, W.E. (1993), "The Early History of Price Index Research", pp. 33-65 in *Essays in Index Number Theory*, Volume 1, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North-Holland.
- Diewert, W.E. (1995), "On the Stochastic Approach to Index Numbers", Discussion Paper 95-31, Department of Economics, University of British Columbia, Vancouver, Canada.
- Diewert, W.E. (1997), "Commentary on Mathew D. Shapiro and David W. Wilcox: Alternative Strategies for Aggregating Price in the CPI", *The Federal Reserve Bank of St. Louis Review*, Vol. 79:3, 127-137.
- Diewert, W.E. (2001), "The Consumer Price Index and Index Number Purpose", *Journal of Economic and Social Measurement* 27, 167-248.
- Diewert, W.E. (2004), "On the Stochastic Approach to Linking the Regions in the ICP", Department of Economics, Discussion Paper 04-16, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z1.
- Diewert, W.E. (2005), "Weighted Country Product Dummy Variable Regressions and Index Number Formulae", *The Review of Income and Wealth* 51:4, 561-571.
- Diewert, W.E. (2009), "Similarity Indexes and Criteria for Spatial Linking", pp. 183-216 in *Purchasing Power Parities of Currencies: Recent Advances in Methods and Applications*, D.S. Prasada Rao (ed.), Cheltenham UK: Edward Elgar.

- Diewert, W.E. (2012), *Consumer Price Statistics in the UK*, Government Buildings, Cardiff Road, Newport, UK, NP10 8XG: Office for National Statistics.  
<http://www.ons.gov.uk/ons/guide-method/userguidance/prices/cpi-and-rpi/index.html>
- Diewert, W.E. (2013), "An Empirical Illustration of Index Construction using Israeli Data on Vegetables", paper presented at the 13th Meeting of the Ottawa Group On Prices at Copenhagen, Denmark. May 2.
- Drobisch, M. W. (1871), "Ueber die Berechnung der Veränderungen der Waarenpreise und des Geldwerths", *Jahrbücher für Nationalökonomie und Statistik* 16, 143-156.
- Edgeworth, F.Y. (1888), "Some New Methods of Measuring Variation in General Prices", *Journal of the Royal Statistical Society* 51, 346-368.
- Edgeworth, F.Y. (1896), "A Defense of Index Numbers", *Economic Journal* 6, 132-142.
- Edgeworth, F.Y. (1901), "Mr. Walsh on the Measurement of General Exchange Value", *Economic Journal* 11, 404-416.
- Eichhorn, W. (1978), *Functional Equations in Economics*, London: Addison-Wesley.
- Eichhorn, W. and J. Voeller (1976), *Theory of the Price Index*, Lecture Notes in Economics and Mathematical Systems, Vol. 140, Berlin: Springer-Verlag.
- Feenstra, Robert C. and Matthew D. Shapiro (2003), "High-Frequency Substitution and the Measurement of Price Indexes", pp. 123-146 in *Scanner Data and Price Indexes*, Robert C. Feenstra and Matthew D. Shapiro (eds.), Studies in Income and Wealth Volume 64, Chicago: The University of Chicago Press.
- Ferger, W.F. (1946), "Historical Note on the Purchasing Power Concept and Index Numbers", *Journal of the American Statistical Association* 41, 53-57.
- Fisher, I. (1911), *The Purchasing Power of Money*, London: Macmillan.
- Fisher, I. (1921), "The Best Form of Index Number", *Quarterly Publication of the American Statistical Association* 17, 533-537.
- Fisher, I. (1922), *The Making of Index Numbers*, Boston: Houghton-Mifflin.
- Frisch, R. (1930), "Necessary and Sufficient Conditions Regarding the Form of an Index Number which Shall Meet Certain of Fisher's Tests", *American Statistical Association Journal* 25, 397-406.
- Funke, H., G. Hacker and J. Voeller (1979), "Fisher's Circular Test Reconsidered", *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 115, 677-687.
- Funke, H. and J. Voeller (1978), "A Note on the Characterization of Fisher's Ideal Index," pp. 177-181 in *Theory and Applications of Economic Indices*, W. Eichhorn, R. Henn, O. Opitz and R.W. Shephard (eds.), W · zburg: Physica Verlag.
- Hill, R.J. (1999a), "Comparing Price Levels across Countries Using Minimum Spanning Trees", *The Review of Economics and Statistics* 81, 135-142.
- Hill, R.J. (1999b), "International Comparisons using Spanning Trees", pp. 109-120 in *International and Interarea Comparisons of Income, Output and Prices*, A. Heston and R.E. Lipsey (eds.), Studies in Income and Wealth Volume 61, NBER, Chicago: The University of Chicago Press.
- Hill, R.J. (2001), "Measuring Inflation and Growth Using Spanning Trees", *International Economic Review* 42, 167-185.
- Hill, R.J. (2004), "Constructing Price Indexes Across Space and Time: The Case of the European Union", *American Economic Review* 94, 1379-1410.
- Hill, R.J. (2009), "Comparing Per Capita Income Levels Across Countries Using Spanning Trees: Robustness, Prior Restrictions, Hybrids and Hierarchies", pp. 217-244 in *Purchasing Power Parities of Currencies: Recent Advances in Methods and Applications*, D.S. Prasada Rao (ed.), Cheltenham UK: Edward Elgar.

- Hill, T.P. (1988), "Recent Developments in Index Number Theory and Practice", *OECD Economic Studies* 10, 123-148.
- Hill, T.P. (1993), "Price and Volume Measures", pp. 379-406 in *System of National Accounts 1993*, Eurostat, IMF, OECD, UN and World Bank, Luxembourg, Washington, D.C., Paris, New York, and Washington, D.C.
- ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004), *Consumer Price Index Manual: Theory and Practice*, Peter Hill (ed.), Geneva: International Labour Office.
- Ivancic, L., W.E. Diewert and K.J. Fox (2009), "Scanner Data, Time Aggregation and the Construction of Price Indexes", Discussion Paper 09-09, Department of Economics, University of British Columbia, Vancouver, Canada.
- Ivancic, L., W.E. Diewert and K.J. Fox (2011), "Scanner Data, Time Aggregation and the Construction of Price Indexes", *Journal of Econometrics* 161, 24-35.
- Jevons, W.S., (1865), "The Variation of Prices and the Value of the Currency since 1782", *Journal of the Statistical Society of London* 28, 294-320; reprinted in *Investigations in Currency and Finance* (1884), London: Macmillan and Co., 119-150.
- Jevons, W.S., (1884), "A Serious Fall in the Value of Gold Ascertained and its Social Effects Set Forth (1863)", pp. 13-118 in *Investigations in Currency and Finance*, London: Macmillan and Co.
- Keynes, J.M. (1930), *Treatise on Money*, Vol. 1, London: Macmillan.
- Knibbs, Sir G.H. (1924), "The Nature of an Unequivocal Price Index and Quantity Index", *Journal of the American Statistical Association* 19, 42-60 and 196-205.
- Konüs, A.A. and S.S. Byushgens (1926), "K probleme pokupatelnoi cili deneg", *Voprosi Konyunkturi* 2, 151-172.
- Laspeyres, E. (1871), "Die Berechnung einer mittleren Waarenpreissteigerung", *Jahrbücher für Nationalökonomie und Statistik* 16, 296-314.
- Lehr, J. (1885), *Beiträge zur Statistik der Preise*, Frankfurt: J.D. Sauerlander.
- Lowe, J. (1823), *The Present State of England in Regard to Agriculture, Trade and Finance*, second edition, London: Longman, Hurst, Rees, Orme and Brown.
- Marshall, A. (1887), "Remedies for Fluctuations of General Prices", *Contemporary Review* 51, 355-375.
- Paasche, H. (1874), "Über die Preisentwicklung der letzten Jahre nach den Hamburger Borsennotirungen", *Jahrbücher für Nationalökonomie und Statistik* 12, 168-178.
- Pierson, N.G. (1896), "Further Considerations on Index-Numbers," *Economic Journal* 6, 127-131.
- Pollak, R.A. (1989), *The Theory of the Cost-of-Living Index*, Oxford: Oxford University Press.
- Samuelson, P.A. and S. Swamy (1974), "Invariant Economic Index Numbers and Canonical Duality: Survey and Synthesis", *American Economic Review* 64, 566-593.
- Selvanathan, E.A. and D.S. Prasada Rao (1994), *Index Numbers: A Stochastic Approach*, Ann Arbor: The University of Michigan Press.
- Shephard, R.W. (1953), *Cost and Production Functions*, Princeton: Princeton University Press.
- Shephard, R.W. (1970), *Theory of Cost and Production Functions*, Princeton: Princeton University Press.
- Sidgwick, H. (1883), *The Principles of Political Economy*, London: Macmillan.
- Szulc, B.J. (1983), "Linking Price Index Numbers," pp. 537-566 in *Price Level Measurement*, W.E. Diewert and C. Montmarquette (eds.), Ottawa: Statistics Canada.
- Szulc, B.J. (1987), "Price Indices below the Basic Aggregation Level", *Bulletin of Labour Statistics* 2, 9-16.

- Theil, H. (1967), *Economics and Information Theory*, Amsterdam: North-Holland Publishing.
- Törnqvist, L. (1936), "The Bank of Finland's Consumption Price Index", *Bank of Finland Monthly Bulletin* 10, 1-8.
- Törnqvist, L. and E. Törnqvist (1937), "Vilket är förhållandet mellan finska markens och svenska kronans köpkraft?", *Ekonomiska Samfundets Tidskrift* 39, 1-39 reprinted as pp. 121-160 in *Collected Scientific Papers of Leo Törnqvist*, Helsinki: The Research Institute of the Finnish Economy, 1981.
- Vogt, A. (1980), "Der Zeit und der Faktorurnkehrtest als 'Finders of Tests'", *Statistische Hefte* 21, 66-71.
- Vogt, A. and J. Barta (1997), *The Making of Tests for Index Numbers*, Heidelberg: Physica-Verlag.
- Walsh, C.M. (1901), *The Measurement of General Exchange Value*, New York: Macmillan and Co.
- Walsh, C.M. (1921a), *The Problem of Estimation*, London: P.S. King & Son.
- Walsh, C. M. (1921b), "Discussion", *Journal of the American Statistical Association* 17, 537-544.
- Walsh, C.M. (1924), "Professor Edgeworth's Views on Index Numbers", *Quarterly Journal of Economics* 38, 500-519.
- Westergaard, H. (1890), *Die Grundzüge der Theorie der Statistik*, Jena: Fischer.
- Wynne, M.A. (1997), "Commentary on Measuring short Run Inflation for Central Bankers", *Federal Reserve Bank of St. Louis Review* 79:3, 161-167.



## Chapter 6

# Index Number Theory: Part II: The Economic Approach to Index Number Theory

### 6.1 Introduction

This part explains the economic approach to index number theory in more detail than was done in Part I (Chapter 5).<sup>\*1</sup>

In section 6.2, we outline the theory of the cost of living index that was first developed by the Russian economist, Konüs (1939)[281]. The approach in this section is completely nonparametric but it sets the stage for later developments.

In section 6.3, we specialize the general theory developed in section 6.2 to the case where the consumer's preferences are homothetic; i.e., they can be represented by a linearly homogeneous utility function. At first glance, it may seem that this restriction is not very interesting from an empirical point of view since Engel's Law demonstrates that overall consumer preferences are not homothetic. However, there are too many commodities in the real world; it is necessary to aggregate similar commodities into subaggregates in order to model the economy. In forming subaggregates, it is very useful to assume the existence of a linearly homogeneous subaggregator function so that we obtain a subaggregate price index that is independent of quantities.

In section 6.4, we establish Shephard's Lemma and Wold's Identity. These results will prove to be very useful in the subsequent sections.

In sections 6.5-6.7, we establish various exact index number formulae in the case where the consumer's preferences are homothetic or where the subaggregator function is linearly homogeneous. These formulae can be evaluated using observable price and quantity data pertaining to the two periods under consideration and they are exactly equal to a corresponding theoretical index, provided that the consumer's preferences can be represented by certain functional forms. We restrict our analysis to the case where the underlying functional form for the preference function can provide a second order approximation to an arbitrary preference function of the type under consideration; i.e., we restrict ourselves to *flexible functional forms* for functions that represent preferences.

In section 6.8, we consider price indexes or cost of living indexes in the case where preferences are general; i.e., we drop the homotheticity assumption in this section and in section 6.9, where we consider quantity indexes in the nonhomothetic case. The situation is much more complicated in the

---

<sup>\*1</sup> There is some duplication with the material that was developed in Part I.

case of nonhomothetic preferences but the results presented in sections 6.8 and 6.9 are reasonably powerful.

Section 6.10 offers a short conclusion.

## 6.2 Konüs True Cost of Living Indexes

In this section, we will outline the theory of the cost of living index for a single consumer (or household) that was first developed by the Russian economist, A. A. Konüs (1939)[281]. This theory relies on the assumption of *optimizing behavior* on the part of the consumer. Thus given a vector of commodity or input prices  $\mathbf{p}^t$  that the consumer faces in a given time period  $t$ , it is assumed that the corresponding observed quantity vector  $\mathbf{q}^t$  is the solution to a cost minimization problem that involves the consumer's preference or utility function  $f$ .

We assume that "the" consumer has well defined *preferences* over different combinations of the  $N$  consumer commodities or items.\*<sup>2</sup> Each combination of items can be represented by a nonnegative vector  $\mathbf{q} \equiv [q_1, \dots, q_N]$ . The consumer's preferences over alternative possible consumption vectors  $\mathbf{q}$  are assumed to be representable by a nonnegative, continuous, increasing, and quasiconcave utility function  $f$ , which is defined over the nonnegative orthant. Thus if  $f(\mathbf{q}^1) > f(\mathbf{q}^0)$ , then the consumer prefers the consumption vector  $\mathbf{q}^1$  to  $\mathbf{q}^0$ . We further assume that the consumer minimizes the cost of achieving the period  $t$  utility level  $u^t \equiv f(\mathbf{q}^t)$  for periods  $t = 0, 1$ . Thus we assume that the observed period  $t$  consumption vector  $\mathbf{q}^t$  solves the following *period  $t$  cost minimization problem*:\*<sup>3</sup>

$$C(u^t, \mathbf{p}^t) \equiv \min_{\mathbf{q}} \{\mathbf{p}^t \cdot \mathbf{q} : f(\mathbf{q}) = u^t\} = \mathbf{p}^t \cdot \mathbf{q}^t; \quad t = 0, 1. \quad (6.1)$$

The period  $t$  price vector for the  $n$  commodities under consideration that the consumer faces is  $\mathbf{p}^t$ . Note that the solution to the cost or expenditure minimization problem (6.1) for a general utility level  $u$  and general vector of commodity prices  $\mathbf{p}$  defines the *consumer's cost or expenditure function*,  $C(u, \mathbf{p})$ . It can be shown\*<sup>4</sup> that  $C(u, \mathbf{p})$  will have the following properties: (i)  $C(u, \mathbf{p})$  is jointly continuous in  $u, \mathbf{p}$  for  $\mathbf{p} \gg \mathbf{0}_N$  and  $u \in U$  where  $U$  is the range of  $f$  and is a nonnegative function over this domain of definition set; (ii)  $C(u, \mathbf{p})$  is increasing in  $u$  for each fixed  $\mathbf{p}$  and (iii)  $C(u, \mathbf{p})$  is nondecreasing, linearly homogeneous and concave function of  $\mathbf{p}$  for each  $u \in U$ . Conversely, if a cost function is given and satisfies the above properties, then the utility function  $f$  that is dual to  $C$  can be recovered using duality theory.\*<sup>5</sup> We shall use the cost function in order to define the consumer's cost of living price index.

The Konüs (1939)[281] family of *true cost of living indexes* pertaining to two periods where the consumer faces the strictly positive price vectors  $\mathbf{p}^0 \equiv (p_1^0, \dots, p_N^0)$  and  $\mathbf{p}^1 \equiv (p_1^1, \dots, p_N^1)$  in periods 0 and 1 respectively is defined as the ratio of the minimum costs of achieving the same utility level  $u \equiv f(\mathbf{q})$  where  $\mathbf{q}$  is a positive reference quantity vector:

$$P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}) \equiv C[f(\mathbf{q}), \mathbf{p}^1] / C[f(\mathbf{q}), \mathbf{p}^0]. \quad (6.2)$$

We say that definition (6.2) defines a *family* of price indexes because there is one such index for each reference quantity vector  $\mathbf{q}$  chosen.

It is natural to choose two specific reference quantity vectors  $\mathbf{q}$  in definition (6.2): the observed base period quantity vector  $\mathbf{q}^0$  and the current period quantity vector  $\mathbf{q}^1$ . The first of these two choices

\*<sup>2</sup> In this section, these preferences are assumed to be invariant over time. In section 6.8 when we introduce environmental variables, this assumption will be relaxed.

\*<sup>3</sup> Notation:  $\mathbf{p}^t \cdot \mathbf{q} \equiv \sum_{n=1}^N p_n^t q_n$ .

\*<sup>4</sup> See Diewert (1993b; 124)[111].

\*<sup>5</sup> See Diewert (1974; 119)[76] (1993b; 129)[111] and Blackorby and Diewert (1979)[38] for the details and for references to various duality theorems.

leads to the following Laspeyres-Konüs true cost of living index:

$$\begin{aligned}
P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0) &\equiv C[f(\mathbf{q}^0), \mathbf{p}^1] / C[f(\mathbf{q}^0), \mathbf{p}^0] \\
&= C[f(\mathbf{q}^0), \mathbf{p}^1] / \mathbf{p}^0 \cdot \mathbf{q}^0 \quad \text{using (6.1) for } t = 0 \\
&= \min_{\mathbf{q}} \{ \mathbf{p}^1 \cdot \mathbf{q} : f(\mathbf{q}) = f(\mathbf{q}^0) \} / \mathbf{p}^0 \cdot \mathbf{q}^0 \quad \text{using the definition of } C[f(\mathbf{q}^0), \mathbf{p}^1] \\
&\leq \mathbf{p}^1 \cdot \mathbf{q}^0 / \mathbf{p}^0 \cdot \mathbf{q}^0 \quad \text{since } \mathbf{q}^0 \equiv (q_1^0, \dots, q_N^0) \text{ is feasible} \\
&\equiv P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)
\end{aligned} \tag{6.3}$$

where  $P_L$  is the observable Laspeyres price index. Thus the (unobservable) Laspeyres-Konüs true cost of living index is bounded from above by the observable Laspeyres price index.\*6

The second of the two natural choices for a reference quantity vector  $\mathbf{q}$  in definition (6.2) leads to the following Paasche-Konüs true cost of living index:

$$\begin{aligned}
P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^1) &\equiv C[f(\mathbf{q}^1), \mathbf{p}^1] / C[f(\mathbf{q}^1), \mathbf{p}^0] \\
&= \mathbf{p}^1 \cdot \mathbf{q}^1 / C[f(\mathbf{q}^1), \mathbf{p}^0] \quad \text{using (6.1) for } t = 1 \\
&= \mathbf{p}^1 \cdot \mathbf{q}^1 / \min_{\mathbf{q}} \{ \mathbf{p}^0 \cdot \mathbf{q} : f(\mathbf{q}) = f(\mathbf{q}^1) \} \quad \text{using the definition of } C[f(\mathbf{q}^1), \mathbf{p}^0] \\
&\geq \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^1 \quad \text{since } \mathbf{q}^1 \equiv (q_1^1, \dots, q_N^1) \text{ is feasible and thus} \\
&\quad C[f(\mathbf{q}^1), \mathbf{p}^0] \leq \mathbf{p}^0 \cdot \mathbf{q}^1 \text{ and } 1/C[f(\mathbf{q}^1), \mathbf{p}^0] \geq 1/\mathbf{p}^0 \cdot \mathbf{q}^1 \\
&\equiv P_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)
\end{aligned} \tag{6.4}$$

where  $P_P$  is the observable Paasche price index. Thus the (unobservable) Paasche-Konüs true cost of living index is bounded from below by the observable Paasche price index.\*7

The bound (6.3) on the Laspeyres-Konüs true cost of living  $P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0)$  using the base period level of utility as the living standard is *one sided* as is the bound (6.4) on the Paasche-Konüs true cost of living  $P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^1)$  using the *current period* level of utility as the living standard. In a remarkable result, Konüs (1939; 20)[281] showed that there exists an intermediate consumption vector  $\mathbf{q}^*$  that is on the straight line joining the base period consumption vector  $\mathbf{q}^0$  and the current period consumption vector  $\mathbf{q}^1$  such that the corresponding (unobservable) true cost of living index  $P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^*)$  is between the observable Laspeyres and Paasche indexes,  $P_L$  and  $P_P$ .\*8 Thus we have:\*9

**Proposition 1** There exists a number  $\lambda^*$  between 0 and 1 such that

$$P_L \leq P_K(\mathbf{p}^0, \mathbf{p}^1, (1 - \lambda^*)\mathbf{q}^0 + \lambda^*\mathbf{q}^1) \leq P_P \quad \text{or} \quad P_P \leq P_K(\mathbf{p}^0, \mathbf{p}^1, (1 - \lambda^*)\mathbf{q}^0 + \lambda^*\mathbf{q}^1) \leq P_L. \tag{6.5}$$

**Proof.** Define  $g(\lambda)$  for  $0 \leq \lambda \leq 1$  by  $g(\lambda) \equiv P_K(\mathbf{p}^0, \mathbf{p}^1, (1 - \lambda)\mathbf{q}^0 + \lambda\mathbf{q}^1)$ . Note that  $g(0) = P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0)$  and  $g(1) = P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^1)$ . There are 24 = (4)(3)(2)(1) possible a priori inequality relations that are possible between the four numbers  $g(0), g(1), P_L$  and  $P_P$ . However, the inequalities (6.3) and (6.4) above imply that  $g(0) \leq P_L$  and  $P_P \leq g(1)$ . This means that there are only six

\*6 This inequality was first obtained by Konüs (1939; 17)[281]. See also Pollak (1983)[331].

\*7 This inequality is also due to Konüs (1939; 19)[281]. See also Pollak (1983)[331].

\*8 For more recent applications of the Konüs method of proof, see Diewert (1983a;191)[95] (2001; 173)[119] for applications in the consumer context and Diewert (1983b; 1059-1061)[96] for an application in the producer context.

\*9 For a generalization of this single consumer result to the case of many consumers, see Diewert (2001; 173)[119].

possible inequalities between the four numbers:

$$g(0) \leq P_L \leq P_P \leq g(1); \quad (6.6)$$

$$g(0) \leq P_P \leq P_L \leq g(1); \quad (6.7)$$

$$g(0) \leq P_P \leq g(1) \leq P_L; \quad (6.8)$$

$$P_P \leq g(0) \leq P_L \leq g(1); \quad (6.9)$$

$$P_P \leq g(1) \leq g(0) \leq P_L; \quad (6.10)$$

$$P_P \leq g(0) \leq g(1) \leq P_L. \quad (6.11)$$

Using the assumptions that: (a) the consumer's utility function  $f$  is continuous over its domain of definition; (b) the utility function is increasing in the components of  $\mathbf{q}$  and hence is subject to local nonsatiation and (c) the price vectors  $\mathbf{p}^t$  have strictly positive components, it is possible to use Debreu's (1959; 19)[71] Maximum Theorem (see also Diewert (1993b; 112-113)[111] for a statement of the Theorem) to show that the consumer's cost function  $C(f(\mathbf{q}), \mathbf{p}^t)$  will be continuous in the components of  $\mathbf{q}$ . Thus using definition (6.2), it can be seen that  $P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q})$  will also be continuous in the components of the vector  $\mathbf{q}$ . Hence  $g(\lambda)$  is a continuous function of  $\lambda$  and assumes all intermediate values between  $g(0)$  and  $g(1)$ . By inspecting the inequalities (6.6)-(6.11) above, it can be seen that we can choose  $\lambda$  between 0 and 1,  $\lambda^*$  say, such that  $P_L \leq g(\lambda^*) \leq P_P$  for case (6.6) or such that  $P_P \leq g(\lambda^*) \leq P_L$  for cases (6.7) to (6.11). Thus at least one of the two inequalities in (6.5) holds. ■

The above inequalities are of some practical importance. If the observable (in principle) Paasche and Laspeyres indexes are not too far apart, then taking a symmetric average of these indexes should provide a good approximation to a true cost of living index where the reference standard of living is somewhere between the base and current period living standards. Note that the theory thus far is completely nonparametric; i.e., we do not have to make any specific assumptions about the functional form of  $f$  or  $C$ .

If we require a single estimate for the price change between the two periods under consideration, then it is natural to take some sort of evenly weighted average of the two bounding indexes which appear in (6.5) as our final estimate of price change between periods 0 and 1. This averaging of the Paasche and Laspeyres strategy is due to Bowley:

“If [the Paasche index] and [the Laspeyres index] lie close together there is no further difficulty; if they differ by much they may be regarded as inferior and superior limits of the index number, which may be estimated as their arithmetic mean . . . as a first approximation.” A. L. Bowley (1901; 227)[?].

“When estimating the factor necessary for the correction of a change found in money wages to obtain the change in real wages, statisticians have not been content to follow Method II only [to calculate a Laspeyres price index], but have worked the problem backwards [to calculate a Paasche price index] as well as forwards. . . . They have then taken the arithmetic, geometric or harmonic mean of the two numbers so found.” A. L. Bowley (1919; 348)[42].\*<sup>10</sup>

Examples of such symmetric averages\*<sup>11</sup> are the arithmetic mean, which leads to the Sidgwick (1883; 68)[359] Bowley (1901; 227)[?]\*<sup>12</sup> index:

$$P_{SB}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv (1/2)P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) + (1/2)P_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \quad (6.12)$$

\*<sup>10</sup> Fisher (1911; 417-418)[185] (1922)[187] also considered the arithmetic, geometric and harmonic averages of the Paasche and Laspeyres indexes.

\*<sup>11</sup> For a discussion of the properties of symmetric averages, see Diewert (1993c)[112].

\*<sup>12</sup> See Diewert (1993a; 36)[110] for additional references to the early history of index number theory.

or the geometric mean, which leads to the Fisher (1922)[187] ideal index:

$$P_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv [P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)P_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)]^{1/2}. \quad (6.13)$$

In order to determine which average of the Laspeyres and Paasche indexes might be “best”, we need *criteria* or *tests* or *properties* that we would like our indexes to satisfy. We will conclude this section by suggesting one possible approach to picking the “best” average.

At this point, it is convenient to define exactly what we mean by a symmetric average of two numbers. Thus let  $a$  and  $b$  be two positive numbers. Diewert (1993c; 361)[112] defined a *symmetric mean* of  $a$  and  $b$  as a function  $m(a, b)$  that has the following properties:

$$m(a, a) = a \text{ for all } a > 0; \quad (\text{mean property}); \quad (6.14)$$

$$m(a, b) = m(b, a) \text{ for all } a > 0, b > 0; \quad (\text{symmetry property}); \quad (6.15)$$

$$m(a, b) \text{ is a continuous function for } a > 0, b > 0; \quad (\text{continuity property}); \quad (6.16)$$

$$m(a, b) \text{ is a strictly increasing function}; \quad (\text{increasingness property}). \quad (6.17)$$

It can be shown that if  $m(a, b)$  satisfies the above properties, then it also satisfies the following property:<sup>\*13</sup>

$$\min\{a, b\} \leq m(a, b) \leq \max\{a, b\}; \quad (\text{min-max property}); \quad (6.18)$$

i.e., the mean of  $a$  and  $b$ ,  $m(a, b)$ , lies between the maximum and minimum of the numbers  $a$  and  $b$ . Since we have restricted the domain of definition of  $a$  and  $b$  to be positive numbers, it can be seen that an implication of (6.18) is that  $m$  also satisfies the following property:

$$m(a, b) > 0 \text{ for all } a > 0, b > 0; \quad (\text{positivity property}); \quad (6.19)$$

If in addition,  $m$  satisfies the following property, then we say that  $m$  is a *homogeneous symmetric mean*:

$$m(\lambda a, \lambda b) = \lambda m(a, b) \text{ for all } \lambda > 0, a > 0, b > 0. \quad (6.20)$$

What is the “best” symmetric average of  $P_L$  and  $P_P$  to use as a point estimate for the theoretical cost of living index? It is very desirable for a price index formula that depends on the price and quantity vectors pertaining to the two periods under consideration to satisfy the *time reversal test*<sup>\*14</sup>. We say that the index number formula  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  satisfies this test if

$$P(\mathbf{p}^1, \mathbf{p}^0, \mathbf{q}^1, \mathbf{q}^0) = 1/P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1); \quad (6.21)$$

i.e., if we interchange the period 0 and period 1 price and quantity data and evaluate the index, then this new index  $P(\mathbf{p}^1, \mathbf{p}^0, \mathbf{q}^1, \mathbf{q}^0)$  is equal to the reciprocal of the original index  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ .

Now we are ready to look for a homogeneous symmetric mean of the Laspeyres and Paasche price indexes that satisfies the time reversal test (6.21).

**Proposition 2**<sup>\*15</sup> The Fisher Ideal price index defined by (6.13) above is the *only* index that is a homogeneous symmetric average of the Laspeyres and Paasche price indexes,  $P_L$  and  $P_P$ , and satisfies the time reversal test (6.21) above.

<sup>\*13</sup> To prove this, use the technique of proof used by Eichhorn and Voeller (1976; 10)[171].

<sup>\*14</sup> See Diewert (1992a; 218)[102] for early references to this test.

<sup>\*15</sup> This result was established by Diewert (1997; 138)[115]

**Proof.** In order to prove this proposition, we only require the homogeneous mean function to satisfy the positivity and homogeneity properties, (6.19) and (6.20) above.

We define the mean price index  $P$  using the function  $m$  as follows:

$$P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv m(P_L, P_P) = m(\mathbf{p}^1 \cdot \mathbf{q}^0 / \mathbf{p}^0 \cdot \mathbf{q}^0, \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^1) \quad (6.22)$$

where we have used the definitions of  $P_L$  and  $P_P$  which are in (6.3) and (6.4) above. Since  $P$  is supposed to satisfy the time reversal test, we can substitute definition (6.22) into (6.21) in order to obtain the following equation:

$$m(\mathbf{p}^0 \cdot \mathbf{q}^1 / \mathbf{p}^1 \cdot \mathbf{q}^1, \mathbf{p}^0 \cdot \mathbf{q}^0 / \mathbf{p}^1 \cdot \mathbf{q}^0) = 1 / m(\mathbf{p}^1 \cdot \mathbf{q}^0 / \mathbf{p}^0 \cdot \mathbf{q}^0, \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^1). \quad (6.23)$$

Letting  $a \equiv \mathbf{p}^1 \cdot \mathbf{q}^0 / \mathbf{p}^0 \cdot \mathbf{q}^0$  and  $b \equiv \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^1$ , we see that equation (6.23) can be rewritten as:

$$m(b^{-1}, a^{-1}) = 1 / m(a, b). \quad (6.24)$$

Equation (6.24) can be rewritten as:

$$\begin{aligned} 1 &= m(a, b) m(b^{-1}, a^{-1}) \\ &= a m(1, b/a) a^{-1} m(a/b, 1) \quad \text{using property (6.20) of } m \\ &= m(1, x) m(x^{-1}, 1) \quad \text{letting } x \equiv b/a \\ &= m(1, x) x^{-1} m(1, x) \quad \text{using property (6.20) of } m. \end{aligned} \quad (6.25)$$

Equation (6.25) can be rewritten as:

$$x = [m(1, x)]^2. \quad (6.26)$$

Thus using (6.19), we can take the positive square root of both sides of (6.26) and obtain

$$m(1, x) = x^{1/2}. \quad (6.27)$$

Using property (6.20) of  $m$  again, we have

$$\begin{aligned} m(a, b) &= a m(1, b/a) \\ &= a [b/a]^{1/2} \quad \text{using (6.27)} \\ &= a^{1/2} b^{1/2}. \end{aligned} \quad (6.28)$$

Now substitute (6.28) into (6.22) and we obtain the Fisher Index. ■

The bounds (6.3)-(6.5) are the best bounds that we can obtain on true cost of living indexes without making further assumptions. In the following sections, we will make further assumptions on the class of utility functions that describe the consumer's tastes for the  $N$  commodities under consideration. With these extra assumptions, we are able to determine the consumer's true cost of living exactly. However, before we can implement this strategy, we require some preliminary theoretical material, which will be developed in the following two sections.

### 6.3 The True Cost of Living Index when Preferences are Homothetic

Up to now, the consumer's preference function  $f$  did not have to satisfy any particular homogeneity assumption. In this section, we assume that  $f$  is (positively) *linearly homogeneous*<sup>\*16</sup>; i.e., we assume

<sup>\*16</sup> This assumption is fairly restrictive in the consumer context. It implies that each indifference curve is a radial projection of the unit utility indifference curve. It also implies that all income elasticities of demand are unity, which is contradicted by empirical evidence.

that the consumer's utility function has the following property:

$$f(\lambda \mathbf{q}) = \lambda f(\mathbf{q}) \text{ for all } \lambda > 0 \text{ and all } \mathbf{q} \geq \mathbf{0}_N. \quad (6.29)$$

Given the continuity of  $f$ , it can be seen that property (6.29) implies that  $f(\mathbf{0}_N) = 0$  so that the lower bound to the range of  $f$  is 0. Furthermore,  $f$  also satisfies  $f(\mathbf{q}) > 0$  if  $\mathbf{q} > \mathbf{0}_N$ .

In the economics literature, assumption (6.29) is known as the assumption of *homothetic preferences*.<sup>\*17</sup> Although this assumption is generally not justified when we consider the consumer's overall cost of living index, it can be justified in the context of a *subaggregate* if we assume that the consumer has a separable subaggregator function,  $f(\mathbf{q})$ , which is linearly homogeneous. In this case,  $\mathbf{q}$  is no longer interpreted as the entire consumption vector, but refers only to a subaggregate such as "food" or "clothing" or some more narrowly defined aggregate.<sup>\*18</sup> Under this assumption, the consumer's subaggregate expenditure or cost function,  $C(u, \mathbf{p})$  defined by (6.1) above (with a new interpretation), decomposes as follows. For a positive subaggregate price vector  $\mathbf{p} \gg \mathbf{0}_N$  and a positive subaggregate utility level  $u$ , we have the following decomposition of  $C$ :

$$\begin{aligned} C(u, \mathbf{p}) &\equiv \min_{\mathbf{q}} \{ \mathbf{p} \cdot \mathbf{q} : f(\mathbf{q}) \geq u \} \\ &= \min_{\mathbf{q}} \{ \mathbf{p} \cdot \mathbf{q} : (1/u)f(\mathbf{q}) \geq 1 \} && \text{dividing by } u > 0 \\ &= \min_{\mathbf{q}} \{ \mathbf{p} \cdot \mathbf{q} : f(\mathbf{q}/u) \geq 1 \} && \text{using the linear homogeneity of } f \\ &= u \min_{\mathbf{q}} \{ \mathbf{p} \cdot \mathbf{q}/u : f(\mathbf{q}/u) \geq 1 \} \\ &= u \min_{\mathbf{z}} \{ \mathbf{p} \cdot \mathbf{z} : f(\mathbf{z}) \geq 1 \} && \text{letting } \mathbf{z} = \mathbf{q}/u \\ &= uC(1, \mathbf{p}) && \text{using definition (6.1) with } u = 1 \\ &= uc(\mathbf{p}) \end{aligned} \quad (6.30)$$

where  $c(\mathbf{p}) \equiv C(1, \mathbf{p})$  is the *unit cost function* that corresponds to  $f$ .<sup>\*19</sup> It can be shown that the unit cost function  $c(\mathbf{p})$  satisfies the same regularity conditions that  $f$  satisfied; i.e.,  $c(\mathbf{p})$  is positive, concave and (positively) linearly homogeneous for positive price vectors.<sup>\*20</sup> Substituting (6.30) into (6.1) and using  $u^t = f(\mathbf{q}^t)$  leads to the following equations:

$$\mathbf{p}^t \cdot \mathbf{q}^t = c(\mathbf{p}^t)f(\mathbf{q}^t) \quad \text{for } t = 0, 1. \quad (6.31)$$

Thus under the linear homogeneity assumption on the utility function  $f$ , observed period  $t$  expenditure on the  $n$  commodities (the left hand side of (6.31) above) is equal to the period  $t$  unit cost  $c(\mathbf{p}^t)$  of achieving one unit of utility times the period  $t$  utility level,  $f(\mathbf{q}^t)$ , (the right hand side of (6.31) above). Obviously, we can identify the period  $t$  unit cost,  $c(\mathbf{p}^t)$ , as the period  $t$  price level  $P^t$  and the period  $t$  level of utility,  $f(\mathbf{q}^t)$ , as the period  $t$  quantity level  $Q^t$ .

<sup>\*17</sup> More precisely, Shephard (1953)[355] defined a homothetic function to be a monotonic transformation of a linearly homogeneous function. However, if a consumer's utility function is homothetic, we can always rescale it to be linearly homogeneous without changing consumer behavior. Hence, we simply identify the homothetic preferences assumption with the linear homogeneity assumption.

<sup>\*18</sup> This particular branch of the economic approach to index number theory is due to Shephard (1953)[355] (1970)[358] and Samuelson and Swamy (1974)[348]. Shephard in particular realized the importance of the homotheticity assumption in conjunction with separability assumptions in justifying the existence of subindexes of the overall cost of living index.

<sup>\*19</sup> Economists will recognize the producer theory counterpart to the result  $C(u, \mathbf{p}) = uc(\mathbf{p})$ : if a producer's production function  $f$  is subject to constant returns to scale, then the corresponding total cost function  $C(u, \mathbf{p})$  is equal to the product of the output level  $u$  times the unit cost  $c(\mathbf{p})$ .

<sup>\*20</sup> Obviously, the utility function  $f$  determines the consumer's cost function  $C(u, \mathbf{p})$  as the solution to the cost minimization problem in the first line of (6.13). Then the unit cost function  $c(\mathbf{p})$  is defined as  $C(1, \mathbf{p})$ . Thus  $f$  determines  $c$ . But we can also use  $c$  to determine  $f$  under appropriate regularity conditions. In the economics literature, this is known as *duality theory*. For additional material on duality theory and the properties of  $f$  and  $c$ , see Samuelson (1953)[342], Shephard (1953)[355] and Diewert (1974)[76] (1993b; 107-123)[111].

The linear homogeneity assumption on the consumer's preference function  $f$  leads to a simplification for the family of Konüs true cost of living indices,  $P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q})$ , defined by (6.2) above. Using this definition for an arbitrary reference quantity vector  $\mathbf{q}$ , we have:

$$\begin{aligned} P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}) &\equiv C[f(\mathbf{q}), \mathbf{p}^1]/C[f(\mathbf{q}), \mathbf{p}^0] \\ &= c(\mathbf{p}^1)f(\mathbf{q})/c(\mathbf{p}^0)f(\mathbf{q}) \quad \text{using (6.30) twice} \\ &= c(\mathbf{p}^1)/c(\mathbf{p}^0). \end{aligned} \quad (6.32)$$

Thus under the homothetic preferences assumption, the entire family of Konüs true cost of living indexes collapses to a single index,  $c(\mathbf{p}^1)/c(\mathbf{p}^0)$ , the ratio of the minimum costs of achieving unit utility level when the consumer faces period 1 and 0 prices respectively. Put another way, *under the homothetic preferences assumption,  $P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q})$  is independent of the reference quantity vector  $\mathbf{q}$ .*

If we use the Konüs true cost of living index defined by the right hand side of (6.32) as our price index concept, then the corresponding implicit quantity index can be defined as the subaggregate value ratio divided by the Konüs price index:

$$\begin{aligned} Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1, \mathbf{q}) &\equiv \mathbf{p}^1 \cdot \mathbf{q}^1 / \{\mathbf{p}^0 \cdot \mathbf{q}^0 P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q})\} \\ &= c(\mathbf{p}^1)f(\mathbf{q}^1) / \{c(\mathbf{p}^0)f(\mathbf{q}^0)P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q})\} \quad \text{using (6.31) twice} \\ &= c(\mathbf{p}^1)f(\mathbf{q}^1) / \{c(\mathbf{p}^0)f(\mathbf{q}^0)[c(\mathbf{p}^1)/c(\mathbf{p}^0)]\} \quad \text{using (6.32)} \\ &= f(\mathbf{q}^1)/f(\mathbf{q}^0). \end{aligned} \quad (6.33)$$

Thus under the homothetic preferences assumption, the *implicit quantity index* that corresponds to the true cost of living price index  $c(\mathbf{p}^1)/c(\mathbf{p}^0)$  is the *utility ratio*  $f(\mathbf{q}^1)/f(\mathbf{q}^0)$ . Since the utility function is assumed to be homogeneous of degree one, this is the natural definition for a quantity index.

## 6.4 Wold's Identity and Shephard's Lemma

In subsequent sections, we will need two additional results from economic theory: Wold's Identity and Shephard's Lemma.

*Wold's* (1944; 69-71)[403] (1953; 145)[404] *Identity* is the following result. Assuming that the consumer satisfies the cost minimization assumptions (6.1) for periods 0 and 1 and that the utility function  $f$  is differentiable at the observed quantity vectors  $\mathbf{q}^0 \gg \mathbf{0}_N$  and  $\mathbf{q}^1 \gg \mathbf{0}_N$  it can be shown\*<sup>21</sup> that the following equations hold:

$$\mathbf{p}^t / \mathbf{p}^t \cdot \mathbf{q}^t = \nabla f(\mathbf{q}^t) / \mathbf{q}^t \cdot \nabla f(\mathbf{q}^t); \quad t = 0, 1. \quad (6.34)$$

If we assume that the utility function is linearly homogeneous, then *Wold's Identity* (6.34) simplifies into the following equations which will prove to be very useful.\*<sup>22</sup>

$$\mathbf{p}^t / \mathbf{p}^t \cdot \mathbf{q}^t = \nabla f(\mathbf{q}^t) / f(\mathbf{q}^t); \quad t = 0, 1. \quad (6.35)$$

\*<sup>21</sup> To prove this, consider the first order necessary conditions for the strictly positive vector  $\mathbf{q}^t$  to solve the period  $t$  cost minimization problem. The conditions of Lagrange with respect to the vector of  $\mathbf{q}$  variables are:  $\mathbf{p}^t = \lambda^t \nabla f(\mathbf{q}^t)$  where  $\lambda^t$  is the optimal Lagrange multiplier and  $\nabla f(\mathbf{q}^t)$  is the vector of first order partial derivatives of  $f$  evaluated at  $\mathbf{q}^t$ . Note that this system of equations is the price equals a constant times marginal utility equations that are familiar to economists. Now take the inner product of both sides of this equation with respect to the period  $t$  quantity vector  $\mathbf{q}^t$  and solve the resulting equation for  $\lambda^t$ . Substitute this solution back into the vector equation  $\mathbf{p}^t = \lambda^t \nabla f(\mathbf{q}^t)$  and we obtain (6.34).

\*<sup>22</sup> Differentiate both sides of the equation  $f(\lambda \mathbf{q}) = \lambda f(\mathbf{q})$  with respect to  $\lambda$  and then evaluate the resulting equation at  $\lambda = 1$ . We obtain the equation  $\sum_{i=1}^N f_i(\mathbf{q})q_i = f(\mathbf{q})$  where  $f_i(\mathbf{q}) \equiv \partial f(\mathbf{q})/\partial q_i$ .

*Shephard's* (1953; 11)[355] *Lemma* is the following result. Consider the period  $t$  cost minimization problem defined by (6.1) above. If the cost function  $C(u^t, \mathbf{p}^t)$  is differentiable with respect to the components of the price vector  $\mathbf{p}$ , then the period  $t$  quantity vector  $\mathbf{q}^t$  is equal to the vector of first order partial derivatives of the cost function with respect to the components of  $\mathbf{p}$ ; i.e., we have

$$\mathbf{q}^t = \nabla_{\mathbf{p}} C(u^t, \mathbf{p}^t); \quad t = 0, 1. \quad (6.36)$$

To explain why (6.36) holds, consider the following argument. Because we are assuming that the observed period  $t$  quantity vector  $\mathbf{q}^t$  solves the cost minimization problem defined by  $C(u^t, \mathbf{p}^t)$ , then  $\mathbf{q}^t$  must be feasible for this problem so we must have  $f(\mathbf{q}^t) = u^t$ . Thus  $\mathbf{q}^t$  is a feasible solution for the following cost minimization problem where the general price vector  $\mathbf{p}$  has replaced the specific period  $t$  price vector  $\mathbf{p}^t$ :

$$C(u^t, \mathbf{p}) \equiv \min_{\mathbf{q}} \{\mathbf{p} \cdot \mathbf{q} : f(\mathbf{q}) \geq u^t\} \leq \mathbf{p} \cdot \mathbf{q}^t \quad \text{for all } \mathbf{p} \gg \mathbf{0}_N \quad (6.37)$$

where the inequality follows from the fact that  $\mathbf{q}^t$  is a feasible (but usually not optimal) solution for the cost minimization problem in (6.37). Now define for each strictly positive price vector  $\mathbf{p}$  the function  $g(\mathbf{p})$  as follows:

$$g(\mathbf{p}) \equiv \mathbf{p} \cdot \mathbf{q}^t - C(u^t, \mathbf{p}). \quad (6.38)$$

Using (6.1) and (6.37), it can be seen that  $g(\mathbf{p})$  is minimized (over all strictly positive price vectors  $\mathbf{p}$ ) at  $\mathbf{p} = \mathbf{p}^t$ . Thus the first order necessary conditions for minimizing a differentiable function of  $N$  variables hold, which simplify to equations (6.36).

If we assume that the utility function is linearly homogeneous, then using (6.30), Shephard's Lemma (6.36) becomes:

$$\mathbf{q}^t = u^t \nabla_{\mathbf{p}} c(\mathbf{p}^t); \quad t = 0, 1. \quad (6.39)$$

Equations (6.31) can be rewritten as follows:

$$\mathbf{p}^t \cdot \mathbf{q}^t = c(\mathbf{p}^t) f(\mathbf{q}^t) = c(\mathbf{p}^t) u^t; \quad t = 0, 1. \quad (6.40)$$

Dividing equations (6.39) by equation (6.40), we obtain the following system of equations:

$$\mathbf{q}^t / \mathbf{p}^t \cdot \mathbf{q}^t = \nabla c(\mathbf{p}^t) / c(\mathbf{p}^t); \quad t = 0, 1. \quad (6.41)$$

Note the symmetry of equations (6.35) with equations (6.41). It is these two sets of equations that we shall use in sections 6.5-6.7 below.

**Problem 1** Suppose the consumer's cost function is  $C(u, \mathbf{p})$  and assume that  $C$  is twice continuously differentiable with respect to its arguments. The consumer's vector of Hicksian demand functions is  $\mathbf{q}(u, \mathbf{p}) \equiv \nabla_{\mathbf{p}} C(u, \mathbf{p})$ . Assume that the consumer's dual utility function is linearly homogeneous and calculate the consumer's vector of Hicksian income elasticities of demand,  $\partial \ln q_n(u, \mathbf{p}) / \partial \ln u$  for  $n = 1, \dots, N$ . Simplify your answer. You can assume that  $q_n(u, \mathbf{p}) > 0$  for each  $n$  and  $u > 0$ .

## 6.5 Superlative Indexes I: The Fisher Ideal Index

Recall that the Fisher price index,  $P_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ , was defined by (6.13). The companion *Fisher quantity index*,  $Q_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ , can be defined as the expenditure ratio for the two periods,  $\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0$ , divided by the price index,  $P_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ :<sup>\*23</sup>

$$Q_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv [\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0] / P_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = [\mathbf{p}^0 \cdot \mathbf{q}^1 \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0 \mathbf{p}^1 \cdot \mathbf{q}^0]^{1/2}. \quad (6.42)$$

<sup>\*23</sup> Given either a price index  $P$  or a quantity index  $Q$ , then a matching index can be defined using the equation  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0$ . Frisch (1930; 399)[191] called this equation the *product test*.

Suppose the consumer has the following utility function:

$$f(\mathbf{q}) \equiv [\mathbf{q}^T \mathbf{A} \mathbf{q}]^{1/2}; \quad \mathbf{A} = \mathbf{A}^T; \mathbf{q} \in S \quad (6.43)$$

where  $\mathbf{A} \equiv [a_{ij}]$  is an  $N \times N$  symmetric matrix that has one positive eigenvalue (that has a strictly positive eigenvector) and the remaining  $N - 1$  eigenvalues are zero or negative. The set  $S$  is the *region of regularity* where the function  $f$  is positive, concave and increasing and hence  $f$  can provide a valid representation of preferences over this region. It can be shown<sup>\*24</sup> that the region of regularity can be defined as follows:

$$S \equiv \{\mathbf{q} : \mathbf{A} \mathbf{q} \gg \mathbf{0}_N; \mathbf{q} \gg \mathbf{0}_N\}. \quad (6.44)$$

Differentiating the  $f(\mathbf{q})$  defined by (6.43) for  $\mathbf{q} \in S$  leads to the following vector of first order partial derivatives:

$$\nabla f(\mathbf{q}) = \mathbf{A} \mathbf{q} / [\mathbf{q}^T \mathbf{A} \mathbf{q}]^{1/2} = \mathbf{A} \mathbf{q} / f(\mathbf{q}) \quad (6.45)$$

where the second equation in (6.45) follows using (6.43). We assume that the consumer minimizes the cost of achieving the utility level  $u^t = f(\mathbf{q}^t)$  for periods  $t = 0, 1$  and the observed period  $t$  quantity vector  $\mathbf{q}^t$  belongs to the regularity region  $S$  for both periods. Evaluate (6.45) at  $\mathbf{q} = \mathbf{q}^t$  and divide both sides of the resulting equation by  $f(\mathbf{q}^t)$ . We obtain the following equations:

$$\nabla f(\mathbf{q}^t) / f(\mathbf{q}^t) = \mathbf{A} \mathbf{q}^t / f(\mathbf{q}^t)^2 = \mathbf{p}^t / \mathbf{p}^t \cdot \mathbf{q}^t; \quad t = 0, 1 \quad (6.46)$$

where the second set of equations in (6.46) follows using Wold's Identity, (6.35).

Now use definition (6.42) for the Fisher ideal quantity index,  $Q_F$ :

$$\begin{aligned} Q_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) &= [\mathbf{p}^0 \cdot \mathbf{q}^1 \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0 \mathbf{p}^1 \cdot \mathbf{q}^0]^{1/2} \\ &= [(\mathbf{p}^0 / \mathbf{p}^0 \cdot \mathbf{q}^0) \cdot \mathbf{q}^1 / (\mathbf{p}^1 / \mathbf{p}^1 \cdot \mathbf{q}^1) \cdot \mathbf{q}^0]^{1/2} \\ &= \{[\mathbf{q}^{0T} \mathbf{A}^T \mathbf{q}^1 / f(\mathbf{q}^0)^2] / [\mathbf{q}^{1T} \mathbf{A}^T \mathbf{q}^0 / f(\mathbf{q}^1)^2]\}^{1/2} \quad \text{using (6.46)} \\ &= [f(\mathbf{q}^1)^2 / f(\mathbf{q}^0)^2]^{1/2} \quad \text{using } \mathbf{A} = \mathbf{A}^T \\ &= f(\mathbf{q}^1) / f(\mathbf{q}^0). \end{aligned} \quad (6.47)$$

Thus under the assumption that the consumer engages in cost minimizing behavior during periods 0 and 1 and has preferences over the  $N$  commodities that correspond to the utility function defined by (6.43), the Fisher ideal quantity index  $Q_F$  is *exactly* equal to the true quantity index,  $f(\mathbf{q}^1) / f(\mathbf{q}^0)$ .<sup>\*25</sup>

Let  $c(\mathbf{p})$  be the unit cost function that corresponds to the homogeneous quadratic utility function  $f$  defined by (6.43). Then using (6.31) and (6.47), it can be seen that

$$\begin{aligned} P_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) &\equiv [\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0] / Q_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \\ &= [\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0] / f(\mathbf{q}^1) / f(\mathbf{q}^0) \quad \text{using (6.47)} \\ &= [c(\mathbf{p}^1) f(\mathbf{q}^1) / c(\mathbf{p}^0) f(\mathbf{q}^0)] / [f(\mathbf{q}^1) / f(\mathbf{q}^0)] \quad \text{using (6.31)} \\ &= c(\mathbf{p}^1) / c(\mathbf{p}^0). \end{aligned} \quad (6.48)$$

Thus under the assumption that the consumer engages in cost minimizing behavior during periods 0 and 1 and has preferences over the  $N$  commodities that correspond to the utility function defined by (6.43), the Fisher ideal price index  $P_F$  is *exactly* equal to the true price index,  $c(\mathbf{p}^1) / c(\mathbf{p}^0)$ .<sup>\*26</sup>

<sup>\*24</sup> See Diewert and Hill (2009)[140].

<sup>\*25</sup> This result was first derived by Konüs and Byushgens (1926)[282]. For the early history of this result, see Diewert (1976; 116)[82].

<sup>\*26</sup> We also require the assumption that  $\mathbf{q}^0$  and  $\mathbf{q}^1$  belong to the regularity region  $S$  defined by (6.44).

A twice continuously differentiable function  $f(\mathbf{q})$  of  $N$  variables  $\mathbf{q}$  can provide a *second order approximation* to another such function  $f^*(\mathbf{q})$  around the point  $\mathbf{q}^*$  if the level and all of the first and second order partial derivatives of the two functions coincide at  $\mathbf{q}^*$ . It can be shown<sup>\*27</sup> that the homogeneous quadratic function  $f$  defined by (6.43) can provide a second order approximation to an arbitrary  $f^*$  around any (strictly positive) point  $\mathbf{q}^*$  in the class of twice continuously differentiable linearly homogeneous functions. Thus the homogeneous quadratic functional form defined by (6.43) is a *flexible functional form*.<sup>\*28</sup>

Diewert (1976; 117)[82] termed an index number formula  $Q_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  that was *exactly* equal to the true quantity index  $f(\mathbf{q}^1)/f(\mathbf{q}^0)$  (where  $f$  is a flexible functional form) a *superlative index number formula*.<sup>\*29</sup> Equation (6.47) and the fact that the homogeneous quadratic function  $f$  defined by (6.43) is a flexible functional form shows that the Fisher ideal quantity index  $Q_F$  is a superlative index number formula. Since the Fisher ideal price index  $P_F$  also satisfies (6.48) where  $c(\mathbf{p})$  is the unit cost function that is generated by the homogeneous quadratic utility function, we also call  $P_F$  a superlative index number formula.

It is possible to show that the Fisher ideal price index is a superlative index number formula by a different route. Instead of starting with the assumption that the consumer's utility function is the homogeneous quadratic function defined by (6.43), we can start with the assumption that the consumer's unit cost function is a homogeneous quadratic. Thus we suppose that the consumer has the following unit cost function:

$$c(\mathbf{p}) \equiv [\mathbf{p}^T \mathbf{B} \mathbf{p}]^{1/2}; \quad \mathbf{B} = \mathbf{B}^T; \mathbf{p} \in S^* \quad (6.49)$$

where  $\mathbf{B} \equiv [b_{ij}]$  is an  $N \times N$  symmetric matrix that has one positive eigenvalue (that has a strictly positive eigenvector) and the remaining  $N - 1$  eigenvalues are zero or negative. The set  $S^*$  is the *price region of regularity* where the function  $c$  is positive, concave and increasing and hence  $c$  can provide a valid representation of preferences over this region. It can be shown that the region of regularity can be defined as follows:<sup>\*30</sup>

$$S^* \equiv \{\mathbf{p} : \mathbf{B} \mathbf{p} \gg \mathbf{0}_N; \mathbf{p} \gg \mathbf{0}_N\}. \quad (6.50)$$

Differentiating the  $c(\mathbf{p})$  defined by (6.49) for  $\mathbf{p} \in S^*$  leads to the following vector of first order partial derivatives:

$$\nabla c(\mathbf{p}) = \mathbf{B} \mathbf{q} / [\mathbf{p}^T \mathbf{B} \mathbf{p}]^{1/2} = \mathbf{B} \mathbf{p} / c(\mathbf{p}) \quad (6.51)$$

where the second equation in (6.51) follows using (6.49). We assume that  $\mathbf{p}^0$  and  $\mathbf{p}^1$  both belong to the regularity region of prices defined by (6.50). Now evaluate the second equation in (6.51) at the observed period  $t$  price vector  $\mathbf{p}^t$  and divide both sides of the resulting equation by  $c(\mathbf{p}^t)$ . We obtain the following equations:

$$\nabla c(\mathbf{p}^t) / c(\mathbf{p}^t) = \mathbf{B} \mathbf{p}^t / c(\mathbf{p}^t)^2 = \mathbf{q}^t / \mathbf{p}^t \cdot \mathbf{q}^t; \quad t = 0, 1 \quad (6.52)$$

where the second set of equations in (6.52) follows using Shephard's Lemma, equations (6.41). Now

<sup>\*27</sup> See Diewert (1976; 130)[82] and let the parameter  $r$  equal 2.

<sup>\*28</sup> Diewert (1974; 133)[76] introduced this term to the economics literature.

<sup>\*29</sup> Fisher (1922; 247)[187] used the term superlative to describe the Fisher ideal price index. Thus Diewert adopted Fisher's terminology but attempted to give some precision to Fisher's definition of superlativeness. Fisher defined an index number formula to be superlative if it approximated the corresponding Fisher ideal results using his data set.

<sup>\*30</sup> See Diewert and Hill (2009)[140] for the details and see Blackorby and Diewert (1979)[38] for local duality theorems.

recall the definition of the Fisher ideal price index,  $P_F$ , given by (6.13) above:

$$\begin{aligned}
P_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) &= [\mathbf{p}^1 \cdot \mathbf{q}^0 \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0 \mathbf{p}^0 \cdot \mathbf{q}^1]^{1/2} \\
&= [\mathbf{p}^1 \cdot (\mathbf{q}^0 / \mathbf{p}^0 \cdot \mathbf{q}^0) / \mathbf{p}^0 \cdot (\mathbf{q}^1 / \mathbf{p}^1 \cdot \mathbf{q}^1)]^{1/2} \\
&= [\mathbf{p}^1 \cdot \{\mathbf{B}\mathbf{p}^0 / c(\mathbf{p}^0)^2\} / \mathbf{p}^0 \cdot \{\mathbf{B}\mathbf{p}^1 / c(\mathbf{p}^1)^2\}]^{1/2} \quad \text{using (6.52)} \\
&= [c(\mathbf{p}^1)^2 / c(\mathbf{p}^0)^2]^{1/2} \quad \text{using } \mathbf{B} = \mathbf{B}^T \\
&= c(\mathbf{p}^1) / c(\mathbf{p}^0). \tag{6.53}
\end{aligned}$$

Thus under the assumption that the consumer engages in cost minimizing behavior during periods 0 and 1 and has preferences over the  $N$  commodities that correspond to the unit cost function defined by (6.49), the Fisher ideal price index  $P_F$  is *exactly* equal to the true price index,  $c(\mathbf{p}^1) / c(\mathbf{p}^0)$ .<sup>\*31</sup>

Since the homogeneous quadratic unit cost function  $c(\mathbf{p})$  defined by (6.49) is also a flexible functional form, the fact that the Fisher ideal price index  $P_F$  exactly equals the true price index  $c(\mathbf{p}^1) / c(\mathbf{p}^0)$  means that  $P_F$  is a *superlative index number formula*.<sup>\*32</sup>

Suppose that the  $\mathbf{B}$  matrix in (6.49) is equal to the following matrix of rank 1:

$$\mathbf{B} \equiv \mathbf{b}\mathbf{b}^T; \quad \mathbf{b} \gg \mathbf{0}_N \tag{6.54}$$

where  $\mathbf{b}$  is an  $N \times 1$  vector with strictly positive components. In this case, it can be verified that the region of regularity is the entire positive orthant. Note that the cost function defined by (6.49) simplifies in this case:

$$c(\mathbf{p}) \equiv [\mathbf{p}^T \mathbf{B}\mathbf{p}]^{1/2} = [\mathbf{p}^T \mathbf{b}\mathbf{b}^T \mathbf{p}]^{1/2} = \mathbf{b}^T \mathbf{p} = \mathbf{b} \cdot \mathbf{p}. \tag{6.55}$$

Substituting (6.55) into Shephard's Lemma (6.39) yields the following expressions for the period  $t$  quantity vectors,  $\mathbf{q}^t$ :

$$\mathbf{q}^t = u^t \nabla_{\mathbf{p}} c(\mathbf{p}^t) = \mathbf{b}u^t; \quad t = 0, 1. \tag{6.56}$$

Thus if the consumer has the preferences that correspond to the unit cost function defined by (6.49) where  $\mathbf{B}$  satisfies the restrictions (6.54), then the period 0 and 1 quantity vectors are equal to a multiple of the vector  $\mathbf{b}$ ; i.e.,  $\mathbf{q}^0 = \mathbf{b}u^0$  and  $\mathbf{q}^1 = \mathbf{b}u^1$ . Under these assumptions, the Fisher, Paasche and Laspeyres indices,  $P_F, P_P$  and  $P_L$ , *all coincide*. However, the (Leontief fixed coefficient) preferences which correspond to the unit cost function defined by (6.49) and (6.54) are not consistent with normal consumer behavior since they imply that the consumer will not substitute away from more expensive commodities to cheaper commodities if relative prices change going from period 0 to 1.

**Problem 2** Prove that  $f(\mathbf{q}) \equiv (\mathbf{q}^T \mathbf{A}\mathbf{q})^{1/2}$  is a flexible functional form in the class of functions that are positively linearly homogeneous; i.e., for  $\mathbf{q}^* \gg \mathbf{0}_N$ , find a symmetric  $\mathbf{A}$  matrix such that the following equations are satisfied:

$$f(\mathbf{q}^*) = f^*(\mathbf{q}^*); \tag{i}$$

$$\nabla f(\mathbf{q}^*) = \nabla f^*(\mathbf{q}^*); \tag{ii}$$

$$\nabla^2 f(\mathbf{q}^*) = \nabla^2 f^*(\mathbf{q}^*) \tag{iii}$$

<sup>\*31</sup> This result was obtained by Diewert (1976; 133-134)[82]. We also require the assumption that  $\mathbf{p}^0$  and  $\mathbf{p}^1$  belong to the regularity region  $S^*$ .

<sup>\*32</sup> Note that we have shown that the Fisher index  $P_F$  is exact for the preferences defined by (6.43) as well as the preferences that are dual to the unit cost function defined by (6.49). These two classes of preferences do not coincide in general. However, if the  $N \times N$  symmetric matrix  $\mathbf{A}$  has an inverse, then it can be shown the corresponding unit cost function is equal to  $c(\mathbf{p}) \equiv (\mathbf{p}^T \mathbf{A}^{-1} \mathbf{p})^{1/2} = (\mathbf{p}^T \mathbf{B}\mathbf{p})^{1/2}$  where  $\mathbf{B} \equiv \mathbf{A}^{-1}$ .

where  $f^*(\mathbf{q})$  is an arbitrary twice continuously differentiable function of  $\mathbf{q}$  that is positively linearly homogeneous.

*Hint:* You can assume that  $\mathbf{A}$  can be written in the following form:

$$\mathbf{A} = \mathbf{a}\mathbf{a}^T + \mathbf{B} \quad (\text{iv})$$

where  $\mathbf{a}^T \equiv [a_1, \dots, a_N]$  is an  $N$  dimensional row vector of parameters and  $\mathbf{B}$  is an  $N \times N$  symmetric matrix which satisfies the following  $N$  linear restrictions:

$$\mathbf{B}\mathbf{q}^* = \mathbf{0}_N. \quad (\text{v})$$

Thus  $\mathbf{B}$  is an  $N \times N$  symmetric matrix that has only  $N(N - 1)/2$  linearly independent parameters in its elements. As a further hint, remember Euler's Theorems on homogeneous functions.

## 6.6 Superlative Indexes II: Quadratic Mean of Order $r$ Indexes

It turns out that there are many other superlative index number formulae; i.e., there exist many quantity indexes  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  that are exactly equal to  $f(\mathbf{q}^1)/f(\mathbf{q}^0)$  and many price indexes  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  that are exactly equal to  $c(\mathbf{p}^1)/c(\mathbf{p}^0)$  where the aggregator function  $f$  or the unit cost function  $c$  is a flexible functional form. We will define two families of superlative indexes below. Suppose that the consumer has the following quadratic mean of order  $r$  utility function:<sup>\*33</sup>

$$f^r(q_1, \dots, q_N) \equiv \left[ \sum_{i=1}^N \sum_{k=1}^N a_{ik} q_i^{r/2} q_k^{r/2} \right]^{1/r} \quad (6.57)$$

where the parameters  $a_{ik}$  satisfy the symmetry conditions  $a_{ik} = a_{ki}$  for all  $i$  and  $k$  and the parameter  $r$  satisfies the restriction  $r \neq 0$ . The regularity region where  $f^r$  is positive, concave and increasing is defined as follows:

$$S \equiv \{ \mathbf{q} : \mathbf{q} \gg \mathbf{0}_N; \nabla f^r(\mathbf{q}) \gg \mathbf{0}_N; \nabla^2 f^r(\mathbf{q}) \text{ is negative semidefinite} \} \quad (6.58)$$

where  $\nabla^2 f^r(\mathbf{q})$  is the matrix of second order partial derivatives of  $f^r$  evaluated at  $\mathbf{q}$ . Diewert (1976; 130)[82] showed that the utility function  $f^r$  defined by (6.57) is a flexible functional form; i.e., it can approximate an arbitrary twice continuously differentiable linearly homogeneous functional form to the second order.<sup>\*34</sup> Note that when  $r = 2$ ,  $f^r$  equals the homogeneous quadratic function defined by (6.43) above.

Define the quadratic mean of order  $r$  quantity index  $Q^r$  by:

$$Q^r(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv \left\{ \sum_{i=1}^N s_i^0 (q_i^1/q_i^0)^{r/2} \right\}^{1/r} \left\{ \sum_{i=1}^N s_i^1 (q_i^1/q_i^0)^{-r/2} \right\}^{-1/r} \quad (6.59)$$

where  $s_i^t \equiv p_i^t q_i^t / \sum_{k=1}^N p_k^t q_k^t$  is the period  $t$  expenditure share for commodity  $i$ . It can be verified that when  $r = 2$ ,  $Q^r$  simplifies into  $Q_F$ , the Fisher ideal quantity index.

Using exactly the same techniques as were used in section 6.5 above, it can be shown that  $Q^r$  is exact for the aggregator function  $f^r$  defined by (6.57); i.e., we have

$$Q^r(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = f^r(\mathbf{q}^1)/f^r(\mathbf{q}^0). \quad (6.60)$$

<sup>\*33</sup> This terminology is due to Diewert (1976; 129)[82].

<sup>\*34</sup> This result holds for any predetermined  $r \neq 0$ ; i.e., we require only the  $N(N + 1)/2$  independent  $a_{ik}$  parameters in order to establish the flexibility of  $f^r$  in the class of linearly homogeneous aggregator functions.

Thus under the assumption that the consumer engages in cost minimizing behavior during periods 0 and 1 and has preferences over the  $N$  commodities that correspond to the utility function defined by (6.57),<sup>\*35</sup> the quadratic mean of order  $r$  quantity index  $Q_F$  is *exactly* equal to the true quantity index,  $f^r(\mathbf{q}^1)/f^r(\mathbf{q}^0)$ .<sup>\*36</sup> Since  $Q^r$  is exact for  $f^r$  and  $f^r$  is a flexible functional form, we see that the quadratic mean of order  $r$  quantity index  $Q^r$  is a *superlative index* for each  $r \neq 0$ . Thus there are an infinite number of superlative quantity indexes.

For each quantity index  $Q^r$ , we can use the counterpart to (6.42) (that the product of the price and quantity index must equal the value ratio) in order to define the corresponding *implicit quadratic mean of order  $r$  price index*  $P^{r*}$ :

$$\begin{aligned} P^{r*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) &\equiv \mathbf{p}^1 \cdot \mathbf{q}^1 / \{\mathbf{p}^0 \cdot \mathbf{q}^0 Q^r(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)\} \\ &= c^{r*}(\mathbf{p}^1) / c^{r*}(\mathbf{p}^0) \end{aligned} \quad (6.61)$$

where  $c^{r*}$  is the unit cost function that corresponds to the aggregator function  $f^r$  defined by (6.57) above. For each  $r \neq 0$ , the implicit quadratic mean of order  $r$  price index  $P^{r*}$  is also a superlative index.

When  $r = 2$ ,  $Q^r$  defined by (6.59) simplifies to  $Q_F$ , the Fisher ideal quantity index and  $P^{r*}$  defined by (6.61) simplifies to  $P_F$ , the Fisher ideal price index. When  $r = 1$ ,  $Q^r$  defined by (6.59) simplifies to:

$$\begin{aligned} Q^1(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) &\equiv \left\{ \sum_{i=1}^N s_i^0 (q_i^1 / q_i^0)^{1/2} \right\} / \left\{ \sum_{i=1}^N s_i^1 (q_i^1 / q_i^0)^{-1/2} \right\} \\ &= \left\{ \sum_{i=1}^N [p_i^0 q_i^0 / \mathbf{p}^0 \cdot \mathbf{q}^0] (q_i^1 / q_i^0)^{1/2} \right\} / \left\{ \sum_{i=1}^N [p_i^1 q_i^1 / \mathbf{p}^1 \cdot \mathbf{q}^1] (q_i^1 / q_i^0)^{-1/2} \right\} \\ &= \left\{ \sum_{i=1}^N p_i^0 (q_i^0 q_i^1)^{1/2} / \mathbf{p}^0 \cdot \mathbf{q}^0 \right\} / \left\{ \sum_{i=1}^N p_i^1 (q_i^0 q_i^1)^{1/2} / \mathbf{p}^1 \cdot \mathbf{q}^1 \right\} \\ &= [\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0] P_W(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \end{aligned} \quad (6.62)$$

where  $P_W$  is the *Walsh* (1901; 398)[389] (1921; 97)[390] *price index*. Thus  $P^{1*}$  is equal to  $P_W$ , the *Walsh price index*, and hence it is also a superlative price index.

Suppose the consumer has the *following quadratic mean of order  $r$  unit cost function*:<sup>\*37</sup>

$$c^r(p_1, \dots, p_N) \equiv \left[ \sum_{i=1}^N \sum_{k=1}^N b_{ik} p_i^{r/2} p_k^{r/2} \right]^{1/r} \quad (6.63)$$

where the parameters  $b_{ik}$  satisfy the symmetry conditions  $b_{ik} = b_{ki}$  for all  $i$  and  $k$  and the parameter  $r$  satisfies the restriction  $r \neq 0$ . Diewert (1976; 130)[82] showed that the unit cost function  $c^r$  defined by (6.63) is a flexible functional form; i.e., it can approximate an arbitrary twice continuously differentiable linearly homogeneous functional form to the second order. Note that when  $r = 2$ ,  $c^r$  equals the homogeneous quadratic unit cost function defined by (6.49) above. The price regularity region for  $c^r$  is defined as follows:

$$S^* \equiv \{ \mathbf{p} : \mathbf{p} \gg \mathbf{0}_N; \nabla c^r(\mathbf{p}) \gg \mathbf{0}_N; \nabla^2 c^r(\mathbf{p}) \text{ is negative semidefinite} \}. \quad (6.64)$$

Define the *quadratic mean of order  $r$  price index*  $P^r$  by:

$$P^r(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv \left\{ \sum_{i=1}^N s_i^0 (p_i^1 / p_i^0)^{r/2} \right\}^{1/r} \left\{ \sum_{i=1}^N s_i^1 (p_i^1 / p_i^0)^{-r/2} \right\}^{-1/r} \quad (6.65)$$

<sup>\*35</sup> We also require that  $\mathbf{q}^0$  and  $\mathbf{q}^1$  belong to the regularity region  $S$  defined by (6.58).

<sup>\*36</sup> See Diewert (1976; 130)[82].

<sup>\*37</sup> This terminology is due to Diewert (1976; 130)[82]. This unit cost function was first defined by Denny (1974)[72].

where  $s_i^t \equiv p_i^t q_i^t / \sum_{k=1}^N p_k^t q_k^t$  is the period  $t$  expenditure share for commodity  $i$  as usual. It can be verified that when  $r = 2$ ,  $P^r$  simplifies into  $P_F$ , the Fisher ideal quantity index.

Using exactly the same techniques as were used in section 6.5 above and using the counterparts to (6.52) and (6.53), it can be shown that  $P^r$  is exact for the unit cost function  $c^r$  defined by (6.63); i.e., we have

$$P^r(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = c^r(\mathbf{p}^1) / c^r(\mathbf{p}^0). \quad (6.66)$$

Thus under the assumption that the consumer engages in cost minimizing behavior during periods 0 and 1 and has preferences over the  $N$  commodities that are dual to the unit cost function defined by (6.63), the quadratic mean of order  $r$  price index  $P^r$  is *exactly* equal to the true price index,  $c^r(\mathbf{p}^1) / c^r(\mathbf{p}^0)$ .<sup>\*38</sup> Since  $P^r$  is exact for  $c^r$  and  $c^r$  is a flexible functional form, we see that the quadratic mean of order  $r$  price index  $P^r$  is a *superlative index* for each  $r \neq 0$ . Thus there are an infinite number of superlative price indexes.

For each price index  $P^r$ , we can use the product test in order to define the corresponding *implicit quadratic mean of order  $r$  quantity index*  $Q^{r*}$ :

$$\begin{aligned} Q^{r*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) &\equiv \mathbf{p}^1 \cdot \mathbf{q}^1 / \{\mathbf{p}^0 \cdot \mathbf{q}^0 P^r(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)\} \\ &= f^{r*}(\mathbf{q}^1) / f^{r*}(\mathbf{q}^0) \end{aligned} \quad (6.67)$$

where  $f^{r*}$  is the aggregator function that is dual to the unit cost function  $c^r$  defined by (6.63) above. For each  $r \neq 0$ , the implicit quadratic mean of order  $r$  quantity index  $Q^{r*}$  is also a superlative index. In this section, we have exhibited two families of superlative price and quantity indexes,  $Q^r$  and  $P^{r*}$  defined by (6.59) and (6.61), and  $P^r$  and  $Q^{r*}$  defined by (6.65) and (6.67) for each  $r \neq 0$ . A natural question to ask at this point is: how different will these indexes be? It is possible to show that all of the price indexes  $P^r$  and  $P^{r*}$  approximate each other to the second order around any point where the price vectors  $\mathbf{p}^0$  and  $\mathbf{p}^1$  are equal and where the quantity vectors  $\mathbf{q}^0$  and  $\mathbf{q}^1$  are equal; i.e., we have the following equalities if the derivatives are evaluated at  $\mathbf{p}^0 = \mathbf{p}^1$  and  $\mathbf{q}^0 = \mathbf{q}^1$  for any  $r$  and  $s$  not equal to 0:<sup>\*39</sup>

$$P^r(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = P^s(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = P^{r*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = P^{s*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1); \quad (6.68)$$

$$\nabla P^r(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \nabla P^s(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \nabla P^{r*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \nabla P^{s*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1); \quad (6.69)$$

$$\nabla^2 P^r(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \nabla^2 P^s(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \nabla^2 P^{r*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) = \nabla^2 P^{s*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1). \quad (6.70)$$

A similar set of equalities holds for the companion quantity indexes,  $Q^r$  and  $Q^{s*}$  for any  $r$  and  $s$  not equal to 0. The implication of the above equalities is that if prices and quantities do not change much over the two periods being compared, then all of the mean of order  $r$  price indexes will give much the same answer and so will all of the mean of order  $r$  quantity indexes.

For an empirical comparisons of some of the above indexes, see Diewert (1978; 894-895)[85] and Hill (2006)[235]. Unfortunately, Hill (2006)[235] showed that the second order approximation property of the mean of order  $r$  indexes breaks down as  $r$  approaches plus or minus infinity. However, in most empirical applications, we generally choose  $r$  equal to 2 (the Fisher case) or 1 (the Walsh indexes). For these cases, the resulting indexes generally approximate each other very closely.<sup>\*40</sup>

<sup>\*38</sup> See Diewert (1976; 133-134)[82].

<sup>\*39</sup> The proof is a straightforward differentiation exercise; see Diewert (1978; 889)[85]. In fact, these derivative equalities are still true provided that  $\mathbf{p}^1 = \lambda \mathbf{p}^0$  and  $\mathbf{q}^1 = \mu \mathbf{q}^0$  for any numbers  $\lambda > 0$  and  $\mu > 0$ .

<sup>\*40</sup> The approximations will be close if we are using annual time series data where price and quantity changes are generally smooth. However, if we are making international comparisons or using panel data or using subannual time series data, then the approximations may not be close.

**Problem 3** Prove (6.60).

**Problem 4** Prove (6.66).

## 6.7 Superlative Indexes III: Normalized Quadratic Indexes

In addition to the family of quadratic means of order  $r$  indexes, there is another family of superlative indexes which we will exhibit in the present section.

Suppose that a consumer has preferences that are dual to the *normalized quadratic unit cost function* defined as follows:<sup>\*41</sup>

$$c(\mathbf{p}) \equiv \mathbf{p}^T \mathbf{b} + (1/2) \mathbf{p}^T \mathbf{A} \mathbf{p} / \alpha^T \mathbf{p}; \quad \mathbf{p} \gg \mathbf{0}_N; \alpha > \mathbf{0}_N; \mathbf{A} = \mathbf{A}^T; \quad (6.71)$$

$$\mathbf{A} \text{ is negative semidefinite}^{*42}; \quad (6.72)$$

where  $\mathbf{p}$  is a positive vector of commodity prices that the consumer faces and the vectors  $\mathbf{b}$  and  $\alpha$  are parameter vectors and the symmetric matrix  $\mathbf{A}$  is a matrix of parameters.

Let  $\mathbf{p}^* \gg \mathbf{0}_N$  be a reference commodity price vector. In addition to the restrictions in (6.71) and (6.72), we can impose the following restrictions on  $c$ :

$$\mathbf{A} \mathbf{p}^* = \mathbf{0}_N. \quad (6.73)$$

If the restrictions on  $\mathbf{A}$  given by (6.73) are satisfied, then it is straightforward to show that we have the following expressions for the first and second order partial derivatives of  $c$  evaluated at  $\mathbf{p} = \mathbf{p}^*$ :

$$\nabla c(\mathbf{p}^*) = \mathbf{b}; \quad (6.74)$$

$$\nabla^2 c(\mathbf{p}^*) = \mathbf{A} / \alpha^T \mathbf{p}^*. \quad (6.75)$$

**Proposition 3** Let  $\alpha$  be an arbitrary predetermined vector which satisfies  $\alpha > \mathbf{0}_N$ . Conditional on this predetermined  $\alpha$ , the  $c(\mathbf{p})$  defined by (6.71), (6.72) and (6.73) is *flexible* at the point of approximation  $\mathbf{p}^*$ ; i.e., there exists a  $\mathbf{b}$  vector and an  $\mathbf{A}$  matrix satisfying (6.73) such that the following equations are satisfied:

$$c(\mathbf{p}^*) = c^*(\mathbf{p}^*); \quad (6.76)$$

$$\nabla c(\mathbf{p}^*) = \nabla c^*(\mathbf{p}^*); \quad (6.77)$$

$$\nabla^2 c(\mathbf{p}^*) = \nabla^2 c^*(\mathbf{p}^*) \quad (6.78)$$

where  $c^*(\mathbf{p})$  is an arbitrary twice continuously differentiable, linearly homogeneous, increasing and concave function of  $\mathbf{p}$  defined for  $\mathbf{p} \gg \mathbf{0}_N$ .

**Proof.** Substitute (6.75) into (6.78) and solve the resulting equation for  $\mathbf{A}$ :

$$\mathbf{A} = \alpha^T \mathbf{p}^* \nabla^2 c^*(\mathbf{p}^*). \quad (6.79)$$

Note that  $\alpha > \mathbf{0}_N$  and  $\mathbf{p}^* \gg \mathbf{0}_N$  implies  $\alpha^T \mathbf{p}^* > 0$ . Since  $c^*$  is concave, it must be the case that  $\nabla^2 c^*(\mathbf{p}^*)$  is a negative semidefinite symmetric matrix. Also, the linear homogeneity of  $c^*$  implies via Euler's Theorem on homogeneous functions that the following restrictions are satisfied:

$$\nabla^2 c^*(\mathbf{p}^*) \mathbf{p}^* = \mathbf{0}_N. \quad (6.80)$$

<sup>\*41</sup> This function was introduced in the producer context by Diewert and Wales (1987; 53)[153] and applied by Diewert and Wales (1992)[156] and Diewert and Lawrence (2002)[143] in this context and by Diewert and Wales (1988a)[154] (1988b)[155] (1993)[157] in the consumer context. The advantages of this flexible functional form are explained in Diewert and Wales (1993)[157].

<sup>\*42</sup> Diewert and Wales (1987; 66)[153] show that this condition is necessary and sufficient for  $c(\mathbf{p})$  to be concave in  $\mathbf{p}$ .

Thus the  $\mathbf{A}$  defined by (6.79) is negative semidefinite and satisfies the restrictions (6.73). Now substitute (6.74) into (6.77) and we obtain the following equation:

$$\mathbf{b} = \nabla c^*(\mathbf{p}^*). \quad (6.81)$$

(6.79) and (6.81) determine  $\mathbf{A}$  and  $\mathbf{b}$  and it can be seen that equations (6.77) and (6.78) are satisfied. The final equation that we need to satisfy to prove the flexibility of  $c(\mathbf{p})$  is (6.76) but this equation is implied by (6.77) and another Euler Theorem on homogeneous functions:

$$c(\mathbf{p}^*) = \mathbf{p}^* \cdot \nabla c(\mathbf{p}^*) \text{ and } c^*(\mathbf{p}^*) = \mathbf{p}^* \cdot \nabla c^*(\mathbf{p}^*). \quad (6.82)$$

■

We note that there are  $N$  free  $b_n$  parameters in the  $\mathbf{b}$  vector and  $N(N-1)/2$  free  $a_{ij}$  parameters in the  $\mathbf{A}$  matrix, taking into account the symmetry restrictions on  $\mathbf{A}$  and the restrictions (6.73). This is a total of  $N(N+1)/2$  free parameters, which is the minimal number of free parameters that is required for a linearly homogeneous  $c(\mathbf{p})$  to be flexible. Thus the normalized quadratic unit cost function defined by (6.71)-(6.73) is a *parsimonious flexible functional form*. In what follows, we do not need to impose the restrictions (6.73).

The region of regularity for the normalized quadratic unit cost function is the following region:

$$S^* \equiv \{\mathbf{p} : \mathbf{p} \gg \mathbf{0}_N; \nabla c(\mathbf{p}) = \mathbf{b} + (\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{A} \mathbf{p} - (\boldsymbol{\alpha}^T \mathbf{p})^{-2} \mathbf{A} \mathbf{p} \mathbf{p}^T \mathbf{A} \gg \mathbf{0}_N\}. \quad (6.83)$$

Suppose that a consumer has preferences that can be represented by a normalized quadratic expenditure function,  $C(u, \mathbf{p})$  equal to  $uc(\mathbf{p})$  where  $c(\mathbf{p})$  is defined by (6.71) and (6.72). Suppose further that the prices that the consumer faces in periods 0 and 1,  $\mathbf{p}^0$  and  $\mathbf{p}^1$ , are in the regularity region defined by (6.83) and the corresponding quantity vectors,  $\mathbf{q}^t$ , are equal to  $\nabla_p C(u^t, \mathbf{p}^t)$  for  $t = 0, 1$  (Shephard's Lemma) where  $u^0 > 0$  and  $u^1 > 0$  are the utility levels that the consumer attains for the two periods. Then Shephard's Lemma gives us the following two equations:

$$\mathbf{q}^0 = [\mathbf{b} + (\boldsymbol{\alpha}^T \mathbf{p}^0)^{-1} \mathbf{A} \mathbf{p}^0 - (1/2)(\boldsymbol{\alpha}^T \mathbf{p}^0)^{-2} \mathbf{p}^{0T} \mathbf{A} \mathbf{p}^0 \boldsymbol{\alpha}] u^0; \quad (6.84)$$

$$\mathbf{q}^1 = [\mathbf{b} + (\boldsymbol{\alpha}^T \mathbf{p}^1)^{-1} \mathbf{A} \mathbf{p}^1 - (1/2)(\boldsymbol{\alpha}^T \mathbf{p}^1)^{-2} \mathbf{p}^{1T} \mathbf{A} \mathbf{p}^1 \boldsymbol{\alpha}] u^1. \quad (6.85)$$

We now derive an exact index number formula that will enable us to calculate the utility ratio  $u^1/u^0$  using just the observable price and quantity data for the two situations,  $\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1$  and the parameter vector  $\boldsymbol{\alpha}$  (which is assumed to be known to us).

Premultiply both sides of (6.84) and (6.85) by the transpose of the price vector  $(\boldsymbol{\alpha}^T \mathbf{p}^0) \mathbf{p}^1 + (\boldsymbol{\alpha}^T \mathbf{p}^1) \mathbf{p}^0$ . After some simplification, we obtain the following formulae:

$$[(\boldsymbol{\alpha}^T \mathbf{p}^0) \mathbf{p}^1 + (\boldsymbol{\alpha}^T \mathbf{p}^1) \mathbf{p}^0]^T \mathbf{q}^0 = \{[(\boldsymbol{\alpha}^T \mathbf{p}^0) \mathbf{p}^1 + (\boldsymbol{\alpha}^T \mathbf{p}^1) \mathbf{p}^0]^T \mathbf{b} + \mathbf{p}^{1T} \mathbf{A} \mathbf{p}^0\} u^0; \quad (6.86)$$

$$[(\boldsymbol{\alpha}^T \mathbf{p}^0) \mathbf{p}^1 + (\boldsymbol{\alpha}^T \mathbf{p}^1) \mathbf{p}^0]^T \mathbf{q}^1 = \{[(\boldsymbol{\alpha}^T \mathbf{p}^0) \mathbf{p}^1 + (\boldsymbol{\alpha}^T \mathbf{p}^1) \mathbf{p}^0]^T \mathbf{b} + \mathbf{p}^{0T} \mathbf{A} \mathbf{p}^1\} u^1. \quad (6.87)$$

Since  $\mathbf{A}$  is symmetric,  $\mathbf{p}^{1T} \mathbf{A} \mathbf{p}^0 = [\mathbf{p}^{1T} \mathbf{A} \mathbf{p}^0]^T = \mathbf{p}^{0T} \mathbf{A}^T \mathbf{p}^1 = \mathbf{p}^{0T} \mathbf{A} \mathbf{p}^1$ , and hence, we have:<sup>43</sup>

$$u^1/u^0 = [(\boldsymbol{\alpha}^T \mathbf{p}^0) \mathbf{p}^1 + (\boldsymbol{\alpha}^T \mathbf{p}^1) \mathbf{p}^0]^T \mathbf{q}^1 / [(\boldsymbol{\alpha}^T \mathbf{p}^0) \mathbf{p}^1 + (\boldsymbol{\alpha}^T \mathbf{p}^1) \mathbf{p}^0]^T \mathbf{q}^0 \equiv Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha}) \quad (6.88)$$

where  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha})$  is the *normalized quadratic quantity index*.<sup>44</sup> Thus if we know  $\boldsymbol{\alpha}$ ,  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha})$  can be calculated using only observable price and quantity data pertaining to the two situations being considered and (6.88) tells us that this quantity index is equal to the utility

<sup>43</sup> This result was obtained by Diewert (1992b; 576)[103].

<sup>44</sup> Diewert (1992b; 576)[103] introduced this index to the economics literature.

ratio  $u^1/u^0$ , which is equal to  $f(\mathbf{q}^1)/f(\mathbf{q}^0)$  where  $f$  is the linearly homogeneous utility function that is dual to the expenditure function defined by (6.71)-(6.72). Thus  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha})$  is a *superlative index number formula* since  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha})$  is exactly equal to the utility ratio  $f(\mathbf{q}^1)/f(\mathbf{q}^0)$  where  $f$  is dual to a flexible functional form for a unit cost function.

It is possible to rewrite (6.88) in a more intuitive form. Define the period  $t$  real prices or normalized prices  $\boldsymbol{\rho}^t$  as the nominal period  $t$  prices  $\mathbf{p}^t$  divided by the period  $t$  fixed weight price index (with fixed quantity weights  $\boldsymbol{\alpha}$ ),  $\mathbf{p}^t \cdot \boldsymbol{\alpha}$ :

$$\boldsymbol{\rho}^t \equiv \mathbf{p}^t / \mathbf{p}^t \cdot \boldsymbol{\alpha}; \quad t = 0, 1. \quad (6.89)$$

Now divide the numerator and denominator in (6.88) by  $\boldsymbol{\alpha}^T \mathbf{p}^0 \boldsymbol{\alpha}^T \mathbf{p}^1$  and we obtain the following expressions for  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha})$ :

$$\begin{aligned} Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha}) &= [\boldsymbol{\rho}^0 + \boldsymbol{\rho}^1] \cdot \mathbf{q}^1 / [\boldsymbol{\rho}^0 + \boldsymbol{\rho}^1] \cdot \mathbf{q}^0 \\ &= [(1/2)\boldsymbol{\rho}^0 + (1/2)\boldsymbol{\rho}^1] \cdot \mathbf{q}^1 / [(1/2)\boldsymbol{\rho}^0 + (1/2)\boldsymbol{\rho}^1] \cdot \mathbf{q}^0. \end{aligned} \quad (6.90)$$

Thus utility in period  $t$ ,  $f(\mathbf{q}^t)$ , can be set equal to  $[(1/2)\boldsymbol{\rho}^0 + (1/2)\boldsymbol{\rho}^1] \cdot \mathbf{q}^t$ , the inner product of the arithmetic average of the real prices pertaining to the two periods,  $(1/2)\boldsymbol{\rho}^0 + (1/2)\boldsymbol{\rho}^1$ , and the period  $t$  quantity vector  $\mathbf{q}^t$ . Thus we have an additive superlative quantity index!<sup>\*45</sup>

The price index  $P_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha})$  that corresponds to the normalized quadratic quantity index defined by (6.88),  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha})$ , is defined using the product test as follows:

$$P_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha}) \equiv \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0 Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha}). \quad (6.91)$$

Since the vector  $\boldsymbol{\alpha}$  could be any nonnegative, nonzero vector, there is nothing to prevent us from setting  $\boldsymbol{\alpha}$  equal to  $\mathbf{q}^0$  or  $\mathbf{q}^1$ . We will consider these two special cases in turn.

*Case 1:  $\boldsymbol{\alpha} = \mathbf{q}^0$ :*

Replacing  $\boldsymbol{\alpha}$  by  $\mathbf{q}^0$  in (6.88) leads to the following special case for the normalized quadratic quantity index:

$$\begin{aligned} Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^0) &= [(\mathbf{q}^0 \cdot \mathbf{p}^0)\mathbf{p}^1 + (\mathbf{q}^0 \cdot \mathbf{p}^1)\mathbf{p}^0] \cdot \mathbf{q}^1 / [(\mathbf{q}^0 \cdot \mathbf{p}^0)\mathbf{p}^1 + (\mathbf{q}^0 \cdot \mathbf{p}^1)\mathbf{p}^0] \cdot \mathbf{q}^0 \\ &= [(\mathbf{q}^0 \cdot \mathbf{p}^0)\mathbf{p}^1 \cdot \mathbf{q}^1 + (\mathbf{q}^0 \cdot \mathbf{p}^1)\mathbf{p}^0 \cdot \mathbf{q}^1] / 2(\mathbf{p}^0 \cdot \mathbf{q}^0)(\mathbf{p}^1 \cdot \mathbf{q}^0) \\ &= (1/2)[\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^1 \cdot \mathbf{q}^0] + (1/2)[\mathbf{p}^0 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0] \\ &= (1/2)Q_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) + (1/2)Q_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \end{aligned} \quad (6.92)$$

where  $Q_L \equiv \mathbf{p}^0 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0$  and  $Q_P \equiv \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^1 \cdot \mathbf{q}^0$  are the Laspeyres and Paasche quantity indexes. Thus when the parameter vector  $\boldsymbol{\alpha}$  is equal to  $\mathbf{q}^0$ , the normalized quadratic quantity index reduces to the arithmetic average of the Paasche and Laspeyres quantity indexes and this index is superlative.

The price index  $P_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^0)$  which corresponds to the normalized quadratic quantity index defined by (6.92),  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^0)$ , can be defined as follows using (6.91):

$$\begin{aligned} P_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^0) &\equiv \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0 Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^0) \\ &= \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0 \{ (1/2)[\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^1 \cdot \mathbf{q}^0] + (1/2)[\mathbf{p}^0 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0] \} \\ &= \{ (1/2)[\mathbf{p}^0 \cdot \mathbf{q}^0 / \mathbf{p}^1 \cdot \mathbf{q}^0] + (1/2)[\mathbf{p}^0 \cdot \mathbf{q}^1 / \mathbf{p}^1 \cdot \mathbf{q}^1] \}^{-1} \\ &= \{ (1/2)[\mathbf{p}^1 \cdot \mathbf{q}^0 / \mathbf{p}^0 \cdot \mathbf{q}^0]^{-1} + (1/2)[\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^1]^{-1} \}^{-1} \\ &= \{ (1/2)[P_L]^{-1} + (1/2)[P_P]^{-1} \}^{-1}. \end{aligned} \quad (6.93)$$

<sup>\*45</sup> The Walsh quantity index,  $Q_W(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \equiv [\sum_{n=1}^N (p_n^0 p_n^1)^{1/2} q_n^1] / [\sum_{n=1}^N (p_n^0 p_n^1)^{1/2} q_n^0] = Q^{1*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ , also has this additivity property.

Thus the superlative price index  $P_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^0)$  which matches up with the normalized quadratic quantity index  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha})$  when we choose  $\boldsymbol{\alpha}$  equal to  $\mathbf{q}^0$  is the *harmonic mean of the Paasche and Laspeyres price indexes*, which were defined in (6.3) and (6.4) above.

Case 2:  $\boldsymbol{\alpha} = \mathbf{q}^1$ :

Replacing  $\boldsymbol{\alpha}$  by  $\mathbf{q}^1$  in (6.88) leads to the following special case for the normalized quadratic quantity index:

$$\begin{aligned} Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^1) &= [(\mathbf{q}^1 \cdot \mathbf{p}^0)\mathbf{p}^1 + (\mathbf{q}^1 \cdot \mathbf{p}^1)\mathbf{p}^0] \cdot \mathbf{q}^1 / [(\mathbf{q}^1 \cdot \mathbf{p}^0)\mathbf{p}^1 + (\mathbf{q}^1 \cdot \mathbf{p}^1)\mathbf{p}^0] \cdot \mathbf{q}^0 \\ &= 2\mathbf{p}^1 \cdot \mathbf{q}^1 \mathbf{p}^0 \cdot \mathbf{q}^1 / \{\mathbf{p}^0 \cdot \mathbf{q}^1 \mathbf{p}^1 \cdot \mathbf{q}^0 + \mathbf{p}^1 \cdot \mathbf{q}^1 \mathbf{p}^0 \cdot \mathbf{q}^0\} \\ &= \{(1/2)[\mathbf{p}^1 \cdot \mathbf{q}^0 / \mathbf{p}^1 \cdot \mathbf{q}^1] + (1/2)[\mathbf{p}^0 \cdot \mathbf{q}^0 / \mathbf{p}^0 \cdot \mathbf{q}^1]\}^{-1} \\ &= \{(1/2)[\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^1 \cdot \mathbf{q}^0]^{-1} + (1/2)[\mathbf{p}^0 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0]^{-1}\}^{-1} \\ &= \{(1/2)Q_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)^{-1} + (1/2)Q_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)^{-1}\}^{-1} \end{aligned} \quad (6.94)$$

where  $Q_L$  and  $Q_P$  are the Laspeyres and Paasche quantity indexes. Thus when the parameter vector  $\boldsymbol{\alpha}$  is equal to  $\mathbf{q}^1$ , the normalized quadratic quantity index reduces to the harmonic average of the Paasche and Laspeyres quantity indexes, which is a superlative index.

The price index  $P_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^1)$  which corresponds to the normalized quadratic quantity index defined by (6.94),  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^1)$ , can be defined as follows using (6.91):

$$\begin{aligned} P_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^1) &\equiv \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0 Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^1) \\ &= \{\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0\} \{ (1/2)[\mathbf{p}^1 \cdot \mathbf{q}^0 / \mathbf{p}^1 \cdot \mathbf{q}^1] + (1/2)[\mathbf{p}^0 \cdot \mathbf{q}^0 / \mathbf{p}^0 \cdot \mathbf{q}^1] \} \\ &= (1/2)[\mathbf{p}^1 \cdot \mathbf{q}^0 / \mathbf{p}^0 \cdot \mathbf{q}^0] + (1/2)[\mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^1] \\ &= (1/2)P_L + (1/2)P_P \\ &= P_{SB}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \end{aligned} \quad (6.95)$$

where  $P_{SB}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is the Sidgwick Bowley price index defined by (6.12). Thus the price index  $P_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \mathbf{q}^1)$  which matches up with the normalized quadratic quantity index  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha})$  when we choose  $\boldsymbol{\alpha}$  equal to  $\mathbf{q}^1$  is the *arithmetic mean of the Paasche and Laspeyres price indexes*.

As in the previous section, we can ask how different are the various normalized quadratic quantity indexes,  $Q_{NQ}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1; \boldsymbol{\alpha})$ , as the predetermined vector  $\boldsymbol{\alpha} > \mathbf{0}_N$  changes. Again, a straightforward differentiation exercise shows that all of these approximate each other to the second order around an equal price (i.e.,  $\mathbf{p}^0 = \mathbf{p}^1$ ) and equal quantity (i.e.,  $\mathbf{q}^0 = \mathbf{q}^1$ ) point. They also approximate all of the mean of order  $r$  quantity indexes,  $Q^r(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  and  $Q^{r*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ , to the second order around an equal price and equal quantity point.\*<sup>46</sup> Thus for “normal” data sets that do not fluctuate too violently, all of these superlative indexes will approximate each other reasonably closely.

The theory of superlative indexes presented in sections 6.5-6.7 provide reasonable methods for aggregation over commodities when the task at hand is to form subindexes. However, these techniques are not suitable for forming overall cost of living indexes or overall quantity indexes when we deal with broad consumer aggregates, because the assumption of homothetic preferences is not likely to be satisfied. Thus in the following sections, we look for methods of aggregation that do not depend on the homotheticity assumption.

\*<sup>46</sup> Diewert (1992b; 578)[103] noted this result.

## 6.8 Nonhomothetic Preferences and Cost of Living Indexes

Before we derive our main results, we require some preliminary results. Suppose the function of  $N$  variables,  $f(z_1, \dots, z_N) \equiv f(\mathbf{z})$ , is quadratic; i.e.,

$$f(\mathbf{z}) \equiv a_0 + \mathbf{a}^T \mathbf{z} + (1/2)\mathbf{z}^T \mathbf{A} \mathbf{z}; \quad \mathbf{A} = \mathbf{A}^T \quad (6.96)$$

where  $\mathbf{a}$  is a vector of parameters and  $\mathbf{A}$  is a symmetric matrix of parameters. It is well known that the second order Taylor series approximation to a quadratic function is *exact*; i.e., if  $f$  is defined by (6.96) above, then for any two points,  $\mathbf{z}^0$  and  $\mathbf{z}^1$ , we have

$$f(\mathbf{z}^1) - f(\mathbf{z}^0) = \nabla f(\mathbf{z}^0)^T (\mathbf{z}^1 - \mathbf{z}^0) + (1/2)(\mathbf{z}^1 - \mathbf{z}^0)^T \nabla^2 f(\mathbf{z}^0) (\mathbf{z}^1 - \mathbf{z}^0). \quad (6.97)$$

It is less well known that *an average of two first order Taylor series approximations* to a quadratic function is also *exact*; i.e., if  $f$  is defined by (6.96) above, then for any two points,  $\mathbf{z}^0$  and  $\mathbf{z}^1$ , we have<sup>\*47</sup>

$$f(\mathbf{z}^1) - f(\mathbf{z}^0) = (1/2)[\nabla f(\mathbf{z}^0) + \nabla f(\mathbf{z}^1)]^T [\mathbf{z}^1 - \mathbf{z}^0]. \quad (6.98)$$

Diewert (1976; 118)[82] and Lau (1979)[287] showed that equation (6.98) characterized a quadratic function and called the equation the *quadratic approximation lemma*. We will refer to (6.98) as the *quadratic identity*.

We now suppose that the consumer's cost function,  $C(u, \mathbf{p})$ , has the following *translog functional form*:<sup>\*48</sup>

$$\begin{aligned} \ln C(u, \mathbf{p}) \equiv & a_0 + \sum_{i=1}^N a_i \ln p_i + (1/2) \sum_{i=1}^N \sum_{k=1}^N a_{ik} \ln p_i \ln p_k \\ & + b_0 \ln u + \sum_{i=1}^N b_i \ln p_i \ln u + (1/2)b_{00}[\ln u]^2 \end{aligned} \quad (6.99)$$

where  $\ln$  is the natural logarithm function and the parameters  $a_i$ ,  $a_{ik}$ , and  $b_i$  satisfy the following restrictions:

$$a_{ik} = a_{ki}; \quad i, k = 1, \dots, N; \quad (6.100)$$

$$\sum_{i=1}^N a_i = 1; \quad (6.101)$$

$$\sum_{i=1}^N b_i = 0; \quad (6.102)$$

$$\sum_{k=1}^N a_{ik} = 0; \quad i = 1, \dots, N. \quad (6.103)$$

The parameter restrictions (6.100)-(6.103) ensure that  $C(u, \mathbf{p})$  defined by (6.99) is linearly homogeneous in  $\mathbf{p}$ . It can be shown that the translog cost function defined by (6.100)-(6.103) can provide a second order Taylor series approximation to an arbitrary cost function.<sup>\*49</sup>

<sup>\*47</sup> To prove that (6.97) and (6.98) are true, substitute definition (6.96) and its derivatives into (6.97) and (6.98). Recall Problem 10 in Chapter 5: Part I.

<sup>\*48</sup> Christensen, Jorgenson and Lau (1975)[55] introduced this function into the economics literature.

<sup>\*49</sup> It can also be shown that if  $b_0 = 1$  and all of the  $b_i = 0$  for  $i = 1, \dots, N$  and  $b_{00} = 0$ , then  $C(u, \mathbf{p}) = uC(1, \mathbf{p}) \equiv uc(\mathbf{p})$ ; i.e., with these additional restrictions on the parameters of the general translog cost function, we have homothetic preferences. Note that we also assume that utility  $u$  is scaled so that  $u$  is always positive. Finally, we assume that for each of our translog results, the regularity region contains the observed price and quantity data.

We assume that the consumer engages in cost minimizing behavior during periods 0 and 1 so that equations (6.1) hold. Applying Shephard's Lemma to the translog cost function leads to the following equations:

$$s_i^t = a_i + \sum_{k=1}^N a_{ik} \ln p_k^t + b_i \ln u^t; \quad i = 1, \dots, N; t = 0, 1 \quad (6.104)$$

where as usual,  $s_i^t$  is the period  $t$  expenditure share on commodity  $i$ . Define the geometric average of the period 0 and 1 utility levels as  $u^*$ ; i.e., define

$$u^* \equiv [u^0 u^1]^{1/2}. \quad (6.105)$$

Now observe that the right hand side of the equation that defines the natural logarithm of the translog cost function, equation (6.99), is a quadratic function of the variables  $z_i \equiv \ln p_i$  if we hold utility constant at the level  $u^*$ . Hence we can apply the quadratic identity, (6.98), and get the following equation:

$$\begin{aligned} & \ln C(u^*, \mathbf{p}^1) - \ln C(u^*, \mathbf{p}^0) \\ &= (1/2) \sum_{i=1}^N [\partial \ln C(u^*, \mathbf{p}^0) / \partial \ln p_i + \partial \ln C(u^*, \mathbf{p}^1) / \partial \ln p_i] [\ln p_i^1 - \ln p_i^0] \\ &= (1/2) \sum_{i=1}^N [a_i + \sum_{k=1}^N a_{ik} \ln p_k^0 + b_i \ln u^* + a_i + \sum_{k=1}^N a_{ik} \ln p_k^1 + b_i \ln u^*] [\ln p_i^1 - \ln p_i^0] \\ & \quad \text{differentiating (6.99) at the points } (u^*, \mathbf{p}^0) \text{ and } (u^*, \mathbf{p}^1) \\ &= (1/2) \sum_{i=1}^N [a_i + \sum_{k=1}^N a_{ik} \ln p_k^0 + b_i \ln [u^0 u^1]^{1/2} + a_i + \sum_{k=1}^N a_{ik} \ln p_k^1 + b_i \ln [u^0 u^1]^{1/2}] \\ & \quad \cdot [\ln p_i^1 - \ln p_i^0] \quad \text{using definition (6.105) for } u^* \\ &= (1/2) \sum_{i=1}^N [a_i + \sum_{k=1}^N a_{ik} \ln p_k^0 + b_i \ln u^0 + a_i + \sum_{k=1}^N a_{ik} \ln p_k^1 + b_i \ln u^1] [\ln p_i^1 - \ln p_i^0] \\ &= (1/2) \sum_{i=1}^N [\partial \ln C(u^0, \mathbf{p}^0) / \partial \ln p_i + \partial \ln C(u^1, \mathbf{p}^1) / \partial \ln p_i] [\ln p_i^1 - \ln p_i^0] \\ & \quad \text{differentiating (6.99) at the points } (u^0, \mathbf{p}^0) \text{ and } (u^1, \mathbf{p}^1) \\ &= (1/2) \sum_{i=1}^N [s_i^0 + s_i^1] [\ln p_i^1 - \ln p_i^0] \quad \text{using (6.104)} \\ &\equiv \ln P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1). \end{aligned} \quad (6.106)$$

The last equation in (6.106) defines the logarithm of an observable index number formula,  $P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ , which is known as the Törnqvist (1936)[373] (1937)[374] Theil (1967)[371] price index. Hence exponentiating both sides of (6.106) yields the following equality between the true cost of living between periods 0 and 1, evaluated at the intermediate utility level  $u^*$  and the observable price index  $P_T$ :<sup>\*50</sup>

$$C(u^*, \mathbf{p}^1) / C(u^*, \mathbf{p}^0) = P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1). \quad (6.107)$$

Since the translog cost function is a flexible functional form, the Törnqvist-Theil price index  $P_T$  is also a *superlative index*.<sup>\*51</sup> The importance of (6.107) as compared to our earlier exact index number results is that we no longer have to assume that preferences are homothetic. However, we do have to choose a particular utility level on the left hand side of (6.107) in order to obtain our new exact result, the geometric mean of  $u^0$  and  $u^1$ .

It is somewhat mysterious how a ratio of *unobservable* cost functions of the form appearing on the left hand side of the above equation can be *exactly* estimated by an *observable* index number formula

<sup>\*50</sup> This result is due to Diewert (1976; 122)[82].

<sup>\*51</sup> Diewert (1978; 888)[85] showed that  $P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  approximates the other superlative indexes  $P^r$  and  $P^{r*}$  to the second order around an equal price and quantity point.

but the key to this mystery is the assumption of cost minimizing behavior and the quadratic identity (6.98) along with the fact that derivatives of cost functions are equal to quantities, as specified by Shephard's Lemma. In fact, all of the exact index number results derived in sections 6.5 and 6.6 can be derived using transformations of the quadratic identity along with Shephard's Lemma (or Wold's identity).<sup>\*52</sup>

It is possible to generalize the above results using some results in Caves, Christensen and Diewert (1982; 1409-1411)[49]. We will conclude this section by explaining those results.

We now assume that in period  $t$ , the consumer has the utility function  $f^t(\mathbf{q}, \mathbf{z}^t)$  for  $t = 0, 1$ , where  $\mathbf{z}^t$  is a period  $t$  vector of *environmental or demographic variables* that affect the consumer's choices over market goods and services,  $\mathbf{q}$ . Note that we are also allowing for taste changes as we move from period 0 to 1. We assume that  $f^t(\mathbf{q}, \mathbf{z}^t)$  is nonnegative, increasing, continuous and quasiconcave in  $\mathbf{q}$  for  $\mathbf{q} \geq \mathbf{0}_N$ .

For  $\mathbf{p} \gg \mathbf{0}_N$ , and  $u$  in the range of  $f^t(\mathbf{q}, \mathbf{z}^t)$ , we define the consumer's period  $t$  cost function  $C^t$  as follows:

$$C^t(u, \mathbf{p}^t, \mathbf{z}^t) \equiv \min_{\mathbf{q}} \{\mathbf{p}^t \cdot \mathbf{q} : f^t(\mathbf{q}, \mathbf{z}^t) = u\}; \quad t = 0, 1. \quad (6.108)$$

Let  $\mathbf{q}^t$  be the consumer's observed market consumption vector for period  $t$  and define the period  $t$  utility level as:

$$u^t \equiv f^t(\mathbf{q}^t, \mathbf{z}^t); \quad t = 0, 1. \quad (6.109)$$

Suppose the consumer faces the market price vector  $\mathbf{p}^t$  in period  $t$  for  $t = 0, 1$ . As usual, we assume that the observed period  $t$  consumption vector  $\mathbf{q}^t$  solves the following *period  $t$  cost minimization problem*:

$$C^t(u^t, \mathbf{p}^t, \mathbf{z}^t) \equiv \min_{\mathbf{q}} \{\mathbf{p}^t \cdot \mathbf{q} : f^t(\mathbf{q}, \mathbf{z}^t) = u^t\} = \mathbf{p}^t \cdot \mathbf{q}^t; \quad t = 0, 1. \quad (6.110)$$

Define a *family of generalized Konüs true cost of living indexes* between periods 0 and 1 as follows:

$$P_{CCD}(\mathbf{p}^0, \mathbf{p}^1, u, \mathbf{z}, t) \equiv C^t(u, \mathbf{p}^1, \mathbf{z}) / C^t(u, \mathbf{p}^0, \mathbf{z}). \quad (6.111)$$

Note that all variables are exactly the same in the numerator and denominator on the right hand side of (6.111), except that the period 1 price vector  $\mathbf{p}^1$  appears in the numerator and the period 0 price vector  $\mathbf{p}^0$  appears in the denominator. Thus the resulting index is a valid measure of pure price change.

Caves, Christensen and Diewert (1982; 1409-1410)[49] singled out the two natural special cases of (6.111), where the common variables in the numerator and denominator on the right hand side of (6.111) are chosen to be the period 0 variables or the period 1 variables:

$$P_{CCD}(\mathbf{p}^0, \mathbf{p}^1, u^0, \mathbf{z}^0, 0) \equiv C^0(u^0, \mathbf{p}^1, \mathbf{z}^0) / C^0(u^0, \mathbf{p}^0, \mathbf{z}^0); \quad (6.112)$$

$$P_{CCD}(\mathbf{p}^0, \mathbf{p}^1, u^1, \mathbf{z}^1, 1) \equiv C^1(u^1, \mathbf{p}^1, \mathbf{z}^1) / C^1(u^1, \mathbf{p}^0, \mathbf{z}^1). \quad (6.113)$$

It turns out that we will not be able to provide empirical approximations to the individual price indexes defined by (6.112) and (6.113) but we will be able to provide an exact index number formula for the geometric mean of these two indexes. In order to accomplish this task, we will require the following generalization of the quadratic identity, (6.98):

---

<sup>\*52</sup> See Diewert (2002)[122].

**Proposition 4** Let  $\mathbf{x}$  and  $\mathbf{y}$  be  $N$  and  $M$  dimensional vectors respectively and let  $f^1$  and  $f^2$  be two general quadratic functions defined as follows:

$$f^1(\mathbf{x}, \mathbf{y}) \equiv a_0^1 + \mathbf{a}^{1T} \mathbf{x} + \mathbf{b}^{1T} \mathbf{y} + (1/2) \mathbf{x}^T \mathbf{A}^1 \mathbf{x} + (1/2) \mathbf{y}^T \mathbf{B}^1 \mathbf{y} + \mathbf{x}^T \mathbf{C}^1 \mathbf{y}; \quad \mathbf{A}^{1T} = \mathbf{A}^1; \quad \mathbf{B}^{1T} = \mathbf{B}^1; \quad (6.114)$$

$$f^2(\mathbf{x}, \mathbf{y}) \equiv a_0^2 + \mathbf{a}^{2T} \mathbf{x} + \mathbf{b}^{2T} \mathbf{y} + (1/2) \mathbf{x}^T \mathbf{A}^2 \mathbf{x} + (1/2) \mathbf{y}^T \mathbf{B}^2 \mathbf{y} + \mathbf{x}^T \mathbf{C}^2 \mathbf{y}; \quad \mathbf{A}^{2T} = \mathbf{A}^2; \quad \mathbf{B}^{2T} = \mathbf{B}^2 \quad (6.115)$$

where the  $a_0^i$  are scalar parameters, the  $\mathbf{a}^i$  and  $\mathbf{b}^i$  are parameter vectors and the  $\mathbf{A}^i, \mathbf{B}^i$  and  $\mathbf{C}^i$  are parameter matrices for  $i = 1, 2$ . Note that the  $\mathbf{A}^i$  and  $\mathbf{B}^i$  are symmetric matrices. If  $\mathbf{A}^1 = \mathbf{A}^2$ , then the following equation holds for all  $\mathbf{x}^1, \mathbf{x}^2, \mathbf{y}^1$  and  $\mathbf{y}^2$ :<sup>\*53</sup>

$$f^1(\mathbf{x}^2, \mathbf{y}^1) - f^1(\mathbf{x}^1, \mathbf{y}^1) + f^2(\mathbf{x}^2, \mathbf{y}^2) - f^2(\mathbf{x}^1, \mathbf{y}^2) = [\nabla_x f^1(\mathbf{x}^1, \mathbf{y}^1) + \nabla_x f^2(\mathbf{x}^2, \mathbf{y}^2)]^T [\mathbf{x}^2 - \mathbf{x}^1]. \quad (6.116)$$

**Proof.** Straightforward differentiation and substitution establishes (6.116). ■

We now suppose that the consumer's period  $t$  cost function,  $C^t(u, \mathbf{p}, \mathbf{z})$ , has the following functional form:<sup>\*54</sup>

$$\begin{aligned} \ln C^t(u, \mathbf{p}, \mathbf{z}) &\equiv a_0^t + \sum_{n=1}^N a_n^t \ln p_n + b_0^t \ln u + \sum_{m=1}^M b_{0m}^t z_m \ln u + \sum_{n=1}^N b_n^t \ln p_n \ln u \\ &+ (1/2) b_{00}^t [\ln u]^2 + (1/2) \sum_{i=1}^N \sum_{n=1}^N a_{in}^t \ln p_i \ln p_n \\ &+ (1/2) \sum_{i=1}^M \sum_{m=1}^M b_{im}^t z_i z_m + \sum_{n=1}^N \sum_{m=1}^M c_{nm}^t z_m \ln p_n \end{aligned} \quad (6.117)$$

where the parameters satisfy the following restrictions, which impose linear homogeneity in prices  $\mathbf{p}$  on  $C^t(u, \mathbf{p}, \mathbf{z})$ :

$$a_{in}^t = a_{ni}^t; \quad i, n = 1, \dots, N; \quad (6.118)$$

$$b_{im}^t = b_{mi}^t; \quad i, m = 1, \dots, M; \quad (6.119)$$

$$\sum_{n=1}^N a_n^t = 1; \quad (6.120)$$

$$\sum_{n=1}^N b_n^t = 0; \quad (6.121)$$

$$\sum_{i=1}^N a_{in}^t = 0; \quad n = 1, \dots, N; \quad (6.122)$$

$$\sum_{n=1}^N c_{nm}^t = 0; \quad m = 1, \dots, M. \quad (6.123)$$

It can be shown that the  $C^t(u, \mathbf{p}, \mathbf{z})$  defined by (6.117) can provide a second order approximation in the variables  $u, \mathbf{p}$  and  $\mathbf{z}$  to an arbitrary twice continuously differentiable cost function,  $C^*(u, \mathbf{p}, \mathbf{z})$ , and hence  $C^t$  is a flexible functional form.

If the consumer in period  $t$  has preferences that are dual to the  $C^t$  defined by (6.117)-(6.123), then Shephard's Lemma implies that the period  $t$  market expenditure shares,  $s_n^t$ , will satisfy the following equations:

$$s_n^t = \partial \ln C^t(u^t, \mathbf{p}^t, \mathbf{z}^t) / \partial \ln p_n = a_n^t + b_n^t \ln u^t + \sum_{i=1}^N a_{ni}^t \ln p_i + \sum_{m=1}^M c_{nm}^t z_m; \quad n = 1, \dots, N; t = 0, 1. \quad (6.124)$$

With the above preliminaries, we can now prove the following Proposition:

<sup>\*53</sup> Balk (1998; 225-226)[20] established this result using the Translog Lemma in Caves, Christensen and Diewert (1982; 1412)[49]. The CCD Translog Lemma is simply a logarithmic version of (6.116).

<sup>\*54</sup> Caves, Christensen and Diewert (1982; 1397)[49] assumed that  $C^t$  was a general translog functional form whereas we are assuming a "mixed" translog functional form, which allows the components of the  $\mathbf{z}$  vector to be 0 if this is required.

**Proposition 5** Suppose the consumer has preferences in period  $t$  that are dual to the cost function  $C^t$  defined by (6.117)-(6.123) for  $t = 0, 1$  and the consumer engages in cost minimizing behavior in each period so that equations (6.110) and (6.124) are satisfied. Finally, suppose that the quadratic coefficients on prices are the same for the two periods under consideration so that:

$$a_{in}^0 = a_{in}^1; \quad i, n = 1, \dots, N. \quad (6.125)$$

Then the geometric mean of the two CCD true cost of living indexes defined by (6.112) and (6.113) is exactly equal to the observable Törnqvist Theil price index  $P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  defined in (6.106) above; i.e., we have:

$$[P_{CCD}(\mathbf{p}^0, \mathbf{p}^1, u^0, \mathbf{z}^0, 0)P_{CCD}(\mathbf{p}^0, \mathbf{p}^1, u^1, \mathbf{z}^1, 1)]^{1/2} = P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1). \quad (6.126)$$

**Proof.** Take twice the logarithm of the left hand side of (6.126). Using definitions (6.112) and (6.113) and using the quadratic nature of  $\ln C^t$  in  $\ln \mathbf{p}$  and  $\mathbf{z}$  (see (6.117)), we obtain the following equation:

$$\begin{aligned} & \ln C^0(u^0, \mathbf{p}^1, \mathbf{z}^0) - \ln C^0(u^0, \mathbf{p}^0, \mathbf{z}^0) + \ln C^1(u^1, \mathbf{p}^1, \mathbf{z}^1) - \ln C^1(u^1, \mathbf{p}^0, \mathbf{z}^1) \\ &= \sum_{n=1}^N [\partial \ln C^0(u^0, \mathbf{p}^0, \mathbf{z}^0) / \partial \ln p_n + \partial \ln C^1(u^1, \mathbf{p}^1, \mathbf{z}^1) / \partial \ln p_n] [\ln p_n^1 - \ln p_n^0] \\ & \hspace{15em} \text{using assumption (6.125) and Proposition 4} \\ &= \sum_{n=1}^N [s_n^0 + s_n^1] [\ln p_n^1 - \ln p_n^0] \quad \text{using (6.124)} \\ &= 2 \ln P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \quad \text{using the definition of } P_T \text{ in (6.106)}. \end{aligned} \quad (6.127)$$

Equation (6.127) is equivalent to (6.126). ■

The above result is essentially equivalent to Theorem 5 in Caves, Christensen and Diewert (1982; 1410)[49].\*<sup>55</sup> The result in Proposition 5 provides a reasonably powerful justification for the use of the Törnqvist Theil price index as a measure of a consumer's change in his or her cost of living index even if preferences are nonhomothetic.\*<sup>56</sup>

Up to this point, we have not studied quantity indexes for the case of nonhomothetic preferences. In the case of a linearly homogeneous aggregator function,  $f(\mathbf{q})$  say, we have noted that the companion quantity index to the Konüs price index  $c(\mathbf{p}^1)/c(\mathbf{p}^0)$  (the unit cost ratio) was the ratio of the quantity aggregates  $f(\mathbf{q}^1)/f(\mathbf{q}^0)$ . In the following section, we will show how to find quantity indexes when preferences are nonhomothetic.

**Problem 5** Prove (6.98).

**Problem 6** Prove (6.116).

## 6.9 Allen Quantity Indexes

Suppose that we make the same assumptions on preferences that we made at the beginning of section 6.2. Let  $C(f(\mathbf{q}), \mathbf{p})$  be the consumer's cost function that is dual to the aggregator function  $f(\mathbf{q})$ . We again assume cost minimizing behavior in periods 0 and 1 so that equations (6.1) are satisfied.

\*<sup>55</sup> CCD assumed that their translog cost functions were quadratic in the logs of prices and the logs of the demographic variables. Balk (1989)[18] also obtained a special case of Proposition 5 where there were no demographic variables but there was taste change. However, Balk's case is also a special case of Theorem 5 in CCD.

\*<sup>56</sup> Note that we have provided two separate interpretations for Törnqvist Theil price index in the context of nonhomothetic preferences.

The *Allen* (1949)[6] *family of true quantity indexes*,  $Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p})$ , is defined for an arbitrary positive reference price vector  $\mathbf{p}$  as follows:

$$Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}) \equiv C(f(\mathbf{q}^1), \mathbf{p})/C(f(\mathbf{q}^0), \mathbf{p}) \quad (6.128)$$

The basic idea of the Allen quantity index dates back to Hicks (1941-42)[220] who observed that if the price vector  $\mathbf{p}$  were held fixed and the quantity vector  $\mathbf{q}$  is free to vary, then  $C(f(\mathbf{q}), \mathbf{p})$  is a perfectly valid cardinal measure of utility.<sup>\*57</sup>

As was the case with the true cost of living, the Allen definition simplifies considerably if the utility function happens to be linearly homogeneous. In this case, (6.128) simplifies to:<sup>\*58</sup>

$$Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}) = f(\mathbf{q}^1)C(1, \mathbf{p})/f(\mathbf{q}^0)C(1, \mathbf{p}) = f(\mathbf{q}^1)/f(\mathbf{q}^0). \quad (6.129)$$

However, in the general case where the consumer has nonhomothetic preferences, we do not obtain the nice simplification given by (6.129).

As usual, it is useful to specialize the general definition of the Allen quantity index and let the reference price vector equal either the period 0 price vector  $\mathbf{p}^0$  or the period 1 price vector  $\mathbf{p}^1$ :

$$Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^0) \equiv C(f(\mathbf{q}^1), \mathbf{p}^0)/C(f(\mathbf{q}^0), \mathbf{p}^0); \quad (6.130)$$

$$Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^1) \equiv C(f(\mathbf{q}^1), \mathbf{p}^1)/C(f(\mathbf{q}^0), \mathbf{p}^1). \quad (6.131)$$

Index number formula that are exact for either of the theoretical indexes defined by (6.130) and (6.131) do not seem to exist, at least for the case of nonhomothetic preferences that can be represented by a flexible functional form. However, we can find an index number formula that is exactly equal to the geometric mean of the Allen indexes defined by (6.130) and (6.131) where the underlying preferences are represented by a flexible functional form. Thus assume that the consumer's preferences can be represented by the general translog cost function,  $C(u, \mathbf{p})$  defined by (6.99), with the restrictions (6.100)-(6.103). This functional form is a special case of the functional form which appears in Proposition 5, with the demographic variables omitted and with no taste changes between periods 0 and 1. Hence we can apply Proposition 5 in the present context, and conclude that the following simplified version of equation (6.126) is satisfied for our plain vanilla translog consumer (but with general nonhomothetic preferences):

$$[\{C(f(\mathbf{q}^0), \mathbf{p}^1)/C(f(\mathbf{q}^0), \mathbf{p}^0)\}\{C(f(\mathbf{q}^1), \mathbf{p}^1)/C(f(\mathbf{q}^1), \mathbf{p}^0)\}]^{1/2} = P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1). \quad (6.132)$$

The implicit quantity index,  $Q_{T^*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  that corresponds to the Törnqvist Theil price index  $P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is defined as the value ratio,  $\mathbf{p}^1 \cdot \mathbf{q}^1/\mathbf{p}^0 \cdot \mathbf{q}^0$ , divided by  $P_T$ . Thus we have:

$$\begin{aligned} Q_{T^*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) &\equiv [\mathbf{p}^1 \cdot \mathbf{q}^1/\mathbf{p}^0 \cdot \mathbf{q}^0]/P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \\ &= [C(f(\mathbf{q}^1), \mathbf{p}^1)/C(f(\mathbf{q}^0), \mathbf{p}^0)]/P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1) \quad \text{using (6.1)} \\ &= [C(f(\mathbf{q}^1), \mathbf{p}^1)/C(f(\mathbf{q}^0), \mathbf{p}^0)]/[\{C(f(\mathbf{q}^0), \mathbf{p}^1)/C(f(\mathbf{q}^0), \mathbf{p}^0)\}\{C(f(\mathbf{q}^1), \mathbf{p}^1)/C(f(\mathbf{q}^1), \mathbf{p}^0)\}]^{1/2} \\ &\quad \text{using (6.132)} \\ &= [\{C(f(\mathbf{q}^1), \mathbf{p}^0)/C(f(\mathbf{q}^0), \mathbf{p}^0)\}\{C(f(\mathbf{q}^1), \mathbf{p}^1)/C(f(\mathbf{q}^0), \mathbf{p}^1)\}]^{1/2} \\ &= [Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^0)Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^1)]^{1/2} \end{aligned} \quad (6.133)$$

where the last equality follows using definitions (6.130) and (6.131). Thus the observable implicit Törnqvist Theil quantity index,  $Q_{T^*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$ , is exactly equal to the geometric mean of the two Allen quantity indexes defined by (6.130) and (6.131).

<sup>\*57</sup> Samuelson (1974)[347] called this a money metric measure of utility.

<sup>\*58</sup> See Diewert (1981)[91] for references to the literature.

Note that in general, the geometric mean of the two “natural” Allen quantity indexes defined by (6.130) and (6.131) matches up with the geometric mean of the two “natural” Konüs price indexes defined by (6.3) and (6.4); i.e., using these definitions, we have:

$$\begin{aligned} [P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0)P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^1)]^{1/2}[Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^0)Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^1)]^{1/2} &= C(f(\mathbf{q}^1), \mathbf{p}^1)/C(f(\mathbf{q}^0), \mathbf{p}^0) \\ &= \mathbf{p}^1 \cdot \mathbf{q}^1 / \mathbf{p}^0 \cdot \mathbf{q}^0. \end{aligned} \quad (6.134)$$

Thus in general, these two “natural” geometric mean price and quantity indexes satisfy the product test. Under our translog assumptions, we have a special case of (6.134) where  $Q_{T^*}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is equal to  $[Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^0)Q_A(\mathbf{q}^0, \mathbf{q}^1, \mathbf{p}^1)]^{1/2}$  and  $P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1)$  is equal to  $[P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0)P_K(\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^1)]^{1/2}$ .

There is an alternative concept for a theoretical quantity index in the case of nonhomothetic preferences that appears frequently in the literature and that is the *Malmquist* (1953)[301] *quantity index*. Results that are similar to the results that we have already derived can be obtained for this concept but we will leave these results to the interested reader.\*<sup>59</sup>

The economic approach to index number theory is based on either the consumer or producer solving an constrained maximization or minimization problem. In this Chapter, we have focused on the consumer’s cost or expenditure minimization problem. In Chapter 9, we will focus on producer optimization problems and look at the economic approach to index number theory from the producer perspective.

## 6.10 Conclusion

It can be seen that it is not necessary to use econometric methods in order to form estimates for price and quantity aggregates; instead, exact index numbers can be used. In particular, empirical index number formula can be used to closely approximate a consumer’s cost of living index or his or her welfare change, even in the case of nonhomothetic preferences.

## 6.11 References

- Allen, R.G.D. (1949), “The Economic Theory of Index Numbers”, *Economica* 16, 197-203.
- Balk, B.M. (1989), “Changing Consumer Preferences and the cost of Living Index: Theory and Nonparametric Expressions”, *Journal of Economics* 50, 157-169.
- Balk, B.M. (1998), *Industrial Price, Quantity and Productivity Indices*, Boston: Kluwer Academic Publishers.
- Blackorby, C. and W.E. Diewert (1979), “Expenditure Functions, Local Duality and Second Order Approximations”, *Econometrica* 47, 579-601.
- Bowley, A.L. (1901), *Elements of Statistics*, Westminster: P.S. King and Son.
- Bowley, A.L. (1919), “The Measurement of Changes in the Cost of Living”, *Journal of the Royal Statistical Society* 82, 343-372.
- Caves, D.W., L.R. Christensen and W.E. Diewert (1982), “The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity”, *Econometrica* 50, 1393-1414.
- Christensen, L.R., D.W. Jorgenson and L.J. Lau (1975), “Transcendental Logarithmic Utility Functions”, *American Economic Review* 65, 367-383.

\*<sup>59</sup> See Diewert (1981)[91] and Caves, Christensen and Diewert (1982)[49] for additional material on this index concept. Diewert (1976; 123-124)[82] provides a nonhomothetic translog result for this index number concept that is an exact analogue to the result in equation (6.106) for a nonhomothetic cost function.

- Denny, M. (1974), "The Relationship Between Functional Forms for the Production System", *Canadian Journal of Economics* 7, 21-31.
- Debreu, G. (1959), *Theory of Value*, New York: John Wiley and Sons.
- Diewert, W.E., 1974. "Applications of Duality Theory," pp. 106-171 in M.D. Intriligator and D.A. Kendrick (ed.), *Frontiers of Quantitative Economics*, Vol. II, Amsterdam: North-Holland.
- Diewert, W.E. (1976), "Exact and Superlative Index Numbers", *Journal of Econometrics* 4, 114-145.
- Diewert, W.E. (1978), "Superlative Index Numbers and Consistency in Aggregation", *Econometrica* 46, 883-900.
- Diewert, W.E. (1981), "The Economic Theory of Index Numbers: A Survey", pp. 163-208 in *Essays in the Theory and Measurement of Consumer Behavior in Honour of sir Richard Stone*. A. Deaton (ed.), London: Cambridge University Press.
- Diewert, W.E. (1983a), "The Theory of the Cost of Living Index and the Measurement of Welfare Change", pp. 163-233 in *Price Level Measurement*, W.E. Diewert and C. Montmarquette (eds.), Ottawa: Statistics Canada, reprinted as pp. 79-147 in *Price Level Measurement*, W.E. Diewert (ed.), Amsterdam: North-Holland, 1990.
- Diewert, W.E. (1983b), "The Theory of the Output Price Index and the Measurement of Real Output Change", pp. 1049-1113 in *Price Level Measurement*, W.E. Diewert and C. Montmarquette (eds.), Ottawa: Statistics Canada.
- Diewert, W.E. (1992a), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *Journal of Productivity Analysis* 3, 211-248.
- Diewert, W.E. (1992b), "Exact and Superlative Welfare Change Indicators", *Economic Inquiry* 30, 565-582.
- Diewert, W.E. (1993a), "The Early History of Price Index Research", pp. 33-65 in *Essays in Index Number Theory*, Volume 1, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North-Holland.
- Diewert, W.E. (1993b), "Duality Approaches To Microeconomic Theory", in *Essays in Index Number Theory*, pp. 105-175 in Volume I, Contributions to Economic Analysis 217, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North Holland.
- Diewert, W.E. (1993c), "Symmetric Means and Choice under Uncertainty", pp. 355-433 in *Essays in Index Number Theory*, Volume 1, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North-Holland.
- Diewert, W.E. (1997), "Commentary on Mathew D. Shapiro and David W. Wilcox: Alternative Strategies for Aggregating Prices in the CPI", *The Federal Reserve Bank of St. Louis Review*, Vol. 79:3, 127-137.
- Diewert, W.E. (2001), "The Consumer Price Index and Index Number Purpose", *Journal of Economic and Social Measurement* 27, 167-248.
- Diewert, W.E. (2002), "The Quadratic Approximation Lemma and Decompositions of Superlative Indexes", *Journal of Economic and Social Measurement* 28, 63-88.
- Diewert, W.E. and R.J. Hill (2009), "Comment on Different Approaches to Index Number Theory", Discussion Paper 09-05, Department of Economics, University of British Columbia, Vancouver, Canada, V6T 1Z1.
- Diewert, W.E. and D. Lawrence (2002), "The Deadweight Costs of Capital Taxation in Australia", pp. 103-167 in *Efficiency in the Public Sector*, Kevin J. Fox (ed.), Boston: Kluwer Academic Publishers.
- Diewert, W.E. and T.J. Wales (1987), "Flexible Functional Forms and Global Curvature Conditions", *Econometrica* 55, 43-68.
- Diewert, W.E. and T.J. Wales (1988a), "Normalized Quadratic Systems of Consumer Demand Functions", *Journal of Business and Economic Statistics* 6, 303-12.

- Diewert, W.E. and T.J. Wales (1988b), "A Normalized Quadratic Semiflexible Functional Form", *Journal of Econometrics* 37, 327-42.
- Diewert, W.E. and T.J. Wales (1992), "Quadratic Spline Models For Producer's Supply and Demand Functions", *International Economic Review* 33, 705-722.
- Diewert, W.E. and T.J. Wales (1993), "Linear and Quadratic Spline Models for Consumer Demand Functions", *Canadian Journal of Economics* 26, 77-106.
- Eichhorn, W. and J. Voeller (1976), *Theory of the Price Index*, Lecture Notes in Economics and Mathematical Systems, Vol. 140, Berlin: Springer-Verlag.
- Fisher, I. (1911), *The Purchasing Power of Money*, London: Macmillan.
- Fisher, I. (1922), *The Making of Index Numbers*, Houghton-Mifflin, Boston.
- Frisch, R. (1930), "Necessary and Sufficient Conditions Regarding the Form of an Index Number Which Shall Meet Certain of Fisher's Tests", *American Statistical Association Journal* 25, 397-406.
- Frisch, R. (1936), "Annual Survey of General Economic Theory: The Problem of Index Numbers", *Econometrica* 4, 1-39.
- Hicks, J.R. (1941-42), "Consumers' Surplus and Index Numbers", *The Review of Economic Studies* 9, 126-137.
- Hill, R.J. (2006), "Superlative Indexes: Not All of Them are Super", *Journal of Econometrics* 130, 25-43.
- Konüs, A.A. (1939), "The Problem of the True Index of the Cost of Living", *Econometrica* 7, 10-29. [Originally published in 1924]
- Konüs, A.A. and S.S. Byushgens (1926), "K probleme pokupatelnoi cili deneg", *Voprosi Konyunkturi* 2, 151-172.
- Lau, L.J. (1979), "On Exact Index Numbers", *Review of Economics and Statistics* 61, 73-82.
- Malmquist, S. (1953) "Index Numbers and Indifference Surfaces", *Trabajos de Estadística* 4, 209-242.
- Pollak, R.A. (1983), "The Theory of the Cost-of-Living Index", pp. 87-161 in *Price Level Measurement*, W.E. Diewert and C. Montmarquette (eds.), Ottawa: Statistics Canada; reprinted as pp. 3-52 in R.A. Pollak, *The Theory of the Cost-of-Living Index*, Oxford: Oxford University Press, 1989.
- Samuelson, P.A. (1953), "Prices of Factors and Goods in General Equilibrium", *Review of Economic Studies* 21, 1-20.
- Samuelson, P.A. (1974), "Complementarity—An Essay on the 40th Anniversary of the Hicks-Allen Revolution in Demand Theory", *Journal of Economic Literature* 12, 1255-1289.
- Samuelson, P.A. and S. Swamy (1974), "Invariant Economic Index Numbers and Canonical Duality: Survey and Synthesis", *American Economic Review* 64, 566-593.
- Shephard, R.W. (1953), *Cost and Production Functions*, Princeton: Princeton University Press.
- Shephard, R.W. (1970), *Theory of Cost and Production Functions*, Princeton: Princeton University Press.
- Sidgwick, H. (1883), *The Principles of Political Economy*, London: Macmillan.
- Theil, H. (1967), *Economics and Information Theory*, Amsterdam: North-Holland Publishing.
- Törnqvist, L. (1936), "The Bank of Finland's Consumption Price Index," *Bank of Finland Monthly Bulletin* 10: 1-8.
- Törnqvist, L. and E. Törnqvist (1937), "Vilket är förhållandet mellan finska markens och svenska kronans köpkraft?", *Ekonomiska Samfundets Tidskrift* 39, 1-39 reprinted as pp. 121-160 in *Collected Scientific Papers of Leo Törnqvist*, Helsinki: The Research Institute of the Finnish Economy, 1981.
- Walsh, C.M. (1901), *The Measurement of General Exchange Value*, New York: Macmillan and Co.
- Walsh, C.M. (1921), *The Problem of Estimation*, London: P.S. King & Son.

- 
- Wold, H. (1944), "A Synthesis of Pure Demand Analysis, Part 3", *Skandinavisk Aktuarietidskrift* 27, 69-120.
- Wold, H. (1953), *Demand Analysis*, New York: John Wiley.



## Chapter 7

# The Measurement of Productivity

### 7.1 Introduction

The productivity of a production unit<sup>\*1</sup> is defined as the output produced by the unit divided by the input used over the same time period. If the input measure is comprehensive, then the productivity concept is called *Total Factor Productivity* (TFP) or *Multifactor Productivity*. If the input measure is labour hours, then the productivity concept is called *Labour Productivity* (LP).

In this chapter, we will focus on the determinants of TFP and how to measure it rather than Labour Productivity. The Labour Productivity concept has its uses but the problem with this concept is that LP could be very high in one country compared to another country but the difference could be entirely due to a larger amount of nonlabour input in the first country. On the other hand, if TFP is much higher in country A compared to country B, then country A will be genuinely more efficient than country B and it will be useful to study the organization of production in country A in order to see if the techniques used there could be exported to less efficient countries.

A problem with the Total Factor Productivity concept is that it depends on the units of measurement for outputs and inputs. Hence TFP can only be compared across production units if the production units are basically in the same line of business so that they are producing the same (or closely similar) outputs and using the same inputs. However, in the time series context, Total Factor Productivity growth rates can be compared over dissimilar production units and hence, we will focus most of our attention on measuring Total Factor Productivity Growth (TFPG).

In section 7.2, we provide an introduction to the measurement issues involved in measuring TFP growth by considering the special case where the production unit produces only a single output and uses only a single input. It turns out in this case, that there are four equivalent ways for measuring TFP growth. In section 7.9, we will deal with the multiple output and multiple input case.

Sections 7.3 to 7.8 discuss the role of TFP growth in explaining economic growth in nontechnical terms (there are no equations in these sections).

Section 7.3 takes a general look at some of the factors that might help to explain economic growth, including *technical change*. Technical change is an outward shift in the production unit's production possibility set, which is due to new process and product innovation and the diffusion of new methods of organizing production. Technical change or a shift in the production function is part of TFP growth.

Section 7.4 singles out productivity growth as one of the most important factors in explaining per capita growth. In addition to technical change, another part of productivity growth is *increasing returns to scale*; i.e., as the scale of the production unit increases, there are a priori reasons for

---

<sup>\*1</sup> A production unit could be an establishment, a firm, an industry or an entire economy.

expecting more output to be produced per unit of input used. Section 7.5 reviews these theories that imply increasing returns to scale. Note that increasing returns to scale are part of productivity growth as we have defined it.

Section 7.6 looks at a variety of other factors that might help to explain economic growth and section 7.7 summarizes the various theories presented in previous sections.

Section 7.8 looks at the role of government in facilitating growth.

Section 7.9 returns to more technical material. The one output, one input measure of productivity growth developed in section 7.2 is generalized to the case of many outputs and many inputs. Index number theory is used to form output and input aggregates and then the analysis proceeds in much the same way as in the one output, one input case.

Section 7.10 uses exact index number theory to set up a simple bivariate regression model that will enable us to estimate returns to scale for a production unit. Unfortunately, running regressions of output growth on input growth usually produces much lower estimates of returns to scale than the ones obtained by running regressions of input growth on output growth. Since both regressions probably do not satisfy the required classical conditions for running a regression (i.e., a truly exogenous independent variable measured without error), it seems that a more appropriate estimation technique is required in order to obtain accurate estimates of returns to scale. Thus section 7.11 examines whether the use of instrumental variable techniques would solve this econometric problem. The discussion in section 7.11 indicates that instrumental variable estimation is unlikely to be completely satisfactory. Thus there is a need for more research on this very fundamental econometric problem.

Section 7.11 concludes with a theoretical model that shows that even if production units are technically efficient in isolation, when we aggregate over production units, the resulting aggregate technology may not be technically efficient. The reason is that different production units may face different prices for the same output or input due to commodity taxes that are industry specific or due to monopolistic pricing power on the part of some production units.

## 7.2 Productivity Measurement in the Case of One Input and One Output

We consider in this section the problem of measuring the total factor productivity (TFP) (and the growth of total factor productivity, TFPG) of a one output, one input firm.\*<sup>2</sup> To do this, we require data on the amounts of output produced,  $y^0$  and  $y^1$ , during two time periods, 0 and 1, and on the amounts of input utilised,  $x^0$  and  $x^1$ , during those same two time periods. It is also convenient to define the firm's revenues  $R^t$  and total costs  $C^t$  for period  $t$  where  $t = 0, 1$ . The average selling price of a unit of output in period  $t$  is assumed to be  $p^t$  and the average cost of a unit of input in period  $t$  is  $w^t$  for  $t = 0, 1$ . Thus we have:

$$R^t = p^t y^t \text{ for } t = 0, 1 \text{ and} \quad (7.1)$$

$$C^t = w^t x^t \text{ for } t = 0, 1. \quad (7.2)$$

Our first definition of the *total factor productivity growth* of the firm going from period 0 to period 1 (or more briefly, of the productivity of the firm) is:

$$\text{TFPG}(1) \equiv [y^1/y^0]/[x^1/x^0]. \quad (7.3)$$

Note that  $y^1/y^0$  is (one plus) the firm's output growth rate\*<sup>3</sup> going from period 0 to period 1 while  $x^1/x^0$  is the corresponding input growth rate going from period 0 to period 1. If  $\text{TFPG}(1) > 1$ ,

\*<sup>2</sup> The material in this section is largely taken from Diewert (1992)[101] and Diewert and Nakamura (2003)[149].

\*<sup>3</sup> In what follows, we will somewhat incorrectly refer to  $y^1/y^0$  as the output growth rate and  $x^1/x^0$  as the input growth rate.

then the output growth rate was greater than the input growth rate and we say that the firm has experienced a *productivity improvement* going from period 0 to period 1. If  $\text{TFPG}(1) < 1$ , then we say that the firm has experienced a *productivity decline*.

The output growth rate,  $y^1/y^0$ , can also be interpreted as a *quantity index of outputs*. In the following section, we will replace  $y^1/y^0$  by a quantity index for outputs. However in the following section, if there is only one output, it can be verified that the output quantity indexes defined there all reduce to the output growth rate,  $y^1/y^0$  defined here for the one output case. Similarly, the input growth rate,  $x^1/x^0$ , can be interpreted as a quantity index of inputs. Hence, our first definition of productivity growth,  $\text{TFPG}(1)$  defined by (7.3), can be interpreted as an output quantity index divided by an input quantity index.

An alternative method for measuring productivity in a one output, one input firm is the *change in technical coefficients* method. Define the input-output coefficient of the firm in period  $t$  as:

$$a^t \equiv y^t/x^t, \quad t = 0, 1. \quad (7.4)$$

Thus,  $a^t$  is the total amount of output  $y^t$  produced by the firm in period  $t$  divided by the total amount of input utilised by the firm in period  $t$ ,  $x^t$ . It can be interpreted as a coefficient which summarises the engineering and economic characteristics of the firm's technology in period  $t$ :  $a^t$  describes the rate at which inputs are transformed into outputs during period  $t$ .

Our second definition of total factor productivity can be expressed in terms of the output-input coefficients,  $a^0$  and  $a^1$ , as follows:

$$\text{TFPG}(2) \equiv a^1/a^0. \quad (7.5)$$

Thus, if  $a^1$  is greater than  $a^0$ , so that the firm is producing more output per unit input in period 1 compared to period 0, then  $\text{TFPG}(2)$  and the firm has experienced an increase in productivity going from period 0 to period 1.

It should be noted that the two productivity growth concepts that we have defined thus far,  $\text{TFPG}(1)$  and  $\text{TFPG}(2)$ , are both relative concepts. This is a general feature of economic definitions of productivity: the performance of the firm in a current period 1 is always compared to its performance in a base period 0. In contrast, an engineering concept of productivity or efficiency is usually an absolute one, concerned with obtaining the maximum amount of output in period one,  $y^1$ , given an available amount of input in period one,  $x^1$ , consistent with the laws of physics.\*<sup>4</sup>

Using (7.3), (7.4), and (7.5), it is easy to show that  $\text{TFPG}(2)$  coincides with an earlier  $\text{TFPG}(1)$  concept in this simple one output, one input model of production; i.e., we have:

$$\text{TFPG}(2) \equiv a^1/a^0 = [y^1/x^1]/[y^0/x^0] = [y^1/y^0]/[x^1/x^0] \equiv \text{TFPG}(1). \quad (7.6)$$

We turn now to a third possible method for defining productivity:

$$\text{TFPG}(3) \equiv [(R^1/R^0)/(p^1/p^0)]/[(C^1/C^0)/(w^1/w^0)]. \quad (7.7)$$

\*<sup>4</sup> Thus, the engineers Norman and Bahiri (1972, p.27)[322] define productivity as the quotient obtained by dividing output by one of the factors of production. Since our simple model has only one factor of production, this engineering definition of productivity reduces to  $a^1 = y^1/x^1$ . However, even engineers recognize that this definition of productivity is unsatisfactory, since it is not invariant to changes in the units of measurement. Thus, Norman and Bahiri (1972, p.28)[322] later define productivity as a relative concept as the following quotation indicates:

"Consequently, we define and measure relative productivity levels in comparison with a level achieved in the past or in comparison with another establishment in the same industry, or in comparison with the national average achieved by another nation."

Thus,  $a^1$  is compared to  $a^0$  where  $a^0 = y^0/x^0$  is a reference input-output coefficient. Note that  $a^1/a^0$  is invariant to changes in the units of measurement. It should be mentioned that sometimes economists (such as Jorgenson and Griliches (1967, p.252)[258]) define productivity as total output divided by total input,  $y^1/x^1 = a^1$ , and then define productivity change as the rate of change of  $a^1$ . However, it is only their productivity change concept that is regarded as being meaningful.

Thus, TFPG(3) is equal to the firm's revenue ratio  $R^1/R^0$  deflated by the output price index  $p^1/p^0$  divided by the cost ratio between the two periods  $C^1/C^0$  deflated by the input price index  $w^1/w^0$ .

Using (7.1), we have

$$(R^1/R^0)/(p^1/p^0) = (p^1 y^1/p^0 y^0)/(p^1/p^0) = y^1/y^0 \quad (7.8)$$

and using (7.2), we have

$$(C^1/C^0)/(w^1/w^0) = (w^1 x^1/w^0 x^0)/(w^1/w^0) = x^1/x^0. \quad (7.9)$$

Thus, in this simple one input, one output model, (7.8) says that the deflated revenue ratio is equal to the output growth rate and (7.9) says that the deflated cost ratio is equal to the input growth rate. Hence, (7.7) equals (7.3) and we have, using (7.6):

$$\text{TFPG}(1) = \text{TFPG}(2) = \text{TFPG}(3). \quad (7.10)$$

There is a fourth way for measuring productivity change that is a generalization of a method originally suggested by Jorgenson and Griliches (1967)[258]. In order to explain this fourth method, we need to introduce the concept of the firm's period  $t$  margin,  $m^t$ ; i.e., define

$$1 + m^t \equiv R^t/C^t; \quad t = 0, 1. \quad (7.11)$$

Thus,  $1 + m^t$  is the ratio of the firm's period  $t$  revenues  $R^t$  to its period  $t$  costs  $C^t$ . If  $m^t$  is zero, then the firm's revenues equal its costs in period  $t$  and the economic profit of the firm is zero. If  $m^t$  is positive, then the bigger  $m^t$  is, the bigger are the firm's profits.

We can now define our fourth way for measuring productivity change in a one output, one input firm:

$$\text{TFPG}(4) \equiv [(1 + m^1)/(1 + m^0)][w^1/w^0]/[p^1/p^0]. \quad (7.12)$$

Thus, TFPG(4) is equal to the margin growth rate  $(1 + m^1)/(1 + m^0)$  times the rate of increase in input prices  $w^1/w^0$  divided by the rate of increase in output prices  $p^1/p^0$ .

If we use equations (7.11) to eliminate  $(1 + m^1)/(1 + m^0)$  in (7.12), we find that

$$\text{TFPG}(4) = \text{TFPG}(3) \quad (7.13)$$

and thus, by (7.10),  $\text{TFPG}(1) = \text{TFPG}(2) = \text{TFPG}(3) = \text{TFPG}(4)$ . Thus, in a one output, one input firm, we have four conceptually distinct methods for measuring productivity change that turn out to be equivalent. (Unfortunately, this equivalence does not generally extend to the multiple output, multiple input case.)

Definition (7.12) of productivity can be used to show the importance of achieving a productivity gain: a productivity improvement is the source for increases in margins or increases in input prices or decreases in output prices. Equation (7.12) also indicates the relationship between total factor productivity and increased profitability. Rearranging (7.12), we have:

$$(1 + m^1)/(1 + m^0) = [\text{TFPG}(4)][p^1/p^0]/[w^1/w^0]. \quad (7.14)$$

Thus, the rate of growth in margins is equal to TFPG times the output price growth rate divided by the input price growth rate.

If there are constant returns to scale in production or margins  $m^t$  are zero for whatever reason in periods 0 and 1, then TFPG(4) reduces to  $[w^1/w^0]/[p^1/p^0]$ , which is the input price index divided by the output price index, a formula due to Jorgenson and Griliches (1967; 252)[258].

We conclude this section with a rather lengthy discussion of the problem of distinguishing TFPG from the concept of technical change or technical progress, TP. In order to distinguish TFPG from

TP, it is necessary to introduce the concept of the firm's period  $t$  production function  $f^t$ ; i.e., in period  $t$ ,  $y = f^t(x)$  denotes the maximum amount of output  $y$  that can be produced by  $x$  units of the input. We assume that in periods 0 and 1, the observed amounts of output,  $y^0$  and  $y^1$ , are produced by the observed amounts of input,  $x^0$  and  $x^1$ , according to the following production function relationships:

$$y^0 = f^0(x^0); \quad (7.15)$$

$$y^1 = f^1(x^1). \quad (7.16)$$

Note that we are now explicitly assuming that production is technically efficient during the two periods under consideration.\*<sup>5</sup>

We define technical progress TP as a measure of the shift in the production function going from period 0 to period 1. There are an infinite number of possible shift measures but it turns out that four measures of technical progress (involving the observed data  $y^0, y^1, x^0$  and  $x^1$  in some way) are the most useful. First, define:

$$y^{0*} \equiv f^1(x^0) \text{ and } y^{1*} \equiv f^0(x^1). \quad (7.17)$$

Thus  $y^{0*}$  is the output that could be produced by the period 0 input  $x^0$  if the period 1 production function  $f^1$  were available and  $y^{1*}$  is the output which could be produced by the period 1 input  $x^1$  but using the period 0 technology which is summarised by the period 0 production function  $f^0$ . Note that in order to define these hypothetical outputs  $y^{0*}$  and  $y^{1*}$ , a knowledge of the period 0 and 1 production functions  $f^0$  and  $f^1$  is required. This knowledge is not easy to acquire but it could be obtained by engineering studies or by econometric (statistical) techniques.

With  $y^{0*}$  and  $y^{1*}$  defined, we can define the following two *output based indexes of technical progress* TP(1) and TP(2):\*<sup>6</sup>

$$\text{TP}(1) \equiv y^{0*}/y^0 = f^1(x^0)/f^0(x^0); \quad (7.18)$$

$$\text{TP}(2) \equiv y^1/y^{1*} = f^1(x^1)/f^0(x^1). \quad (7.19)$$

Thus, TP(1) is one plus the percentage increase in output due to technical and managerial improvements (going from period 0 to period 1) evaluated at the period 0 input level  $x^0$  and TP(2) is one plus the percentage increase in output due to the new technology evaluated at the period 1 input level  $x^1$ .

It is also possible to define input based measures of technical progress TP(3) and TP(4). First, define  $x^{0*}$  and  $x^{1*}$  as follows:

$$y^0 = f^1(x^{0*}) \text{ and } y^1 = f^0(x^{1*}). \quad (7.20)$$

Thus,  $x^{0*}$  is the input required to produce the period 0 output  $y^0$  but by using the period 1 technology, and so  $x^{0*}$  will generally be less than  $x^0$  (which is the amount of input required to produce the period 0 output using the period 0 technology). Similarly,  $x^{1*}$  is the amount of input required to produce

\*<sup>5</sup> In benchmarking studies or in studies where we compare the relative efficiency of different production units producing the same outputs and using the same inputs, we do not assume that each production unit is globally efficient; i.e., the best practice production unit is regarded as being technically efficient but the other production units may not be technically efficient relative to the global best practice technology. However, in the time series context, it seems acceptable to assume that each production unit is technically efficient in each period *relative to its own knowledge of the technology available to it*. In other words, individual production units are efficient relative to their own knowledge base but of course they can be inefficient relative to the world wide best practice technology.

\*<sup>6</sup> TP(1) and TP(2) are the one input, one output special cases of Caves, Christensen, and Diewert's (1982; 1402)[49] output based 'productivity' indexes.

the period 1 output  $y^1$  but by using the period 0 technology, and  $x^{1*}$  will generally be larger than  $x^1$  (because the period 0 technology will generally be less efficient than the period 1 technology). Now define the following two *input based measures of technical progress*, TP(3) and TP(4):<sup>\*7</sup>

$$TP(3) \equiv x^0/x^{0*}; \tag{7.21}$$

$$TP(4) \equiv x^{1*}/x^1. \tag{7.22}$$

The above four measures of TP can be illustrated with the aid of Figure 7.1. The diagram shows that each of the TP measures can be different.

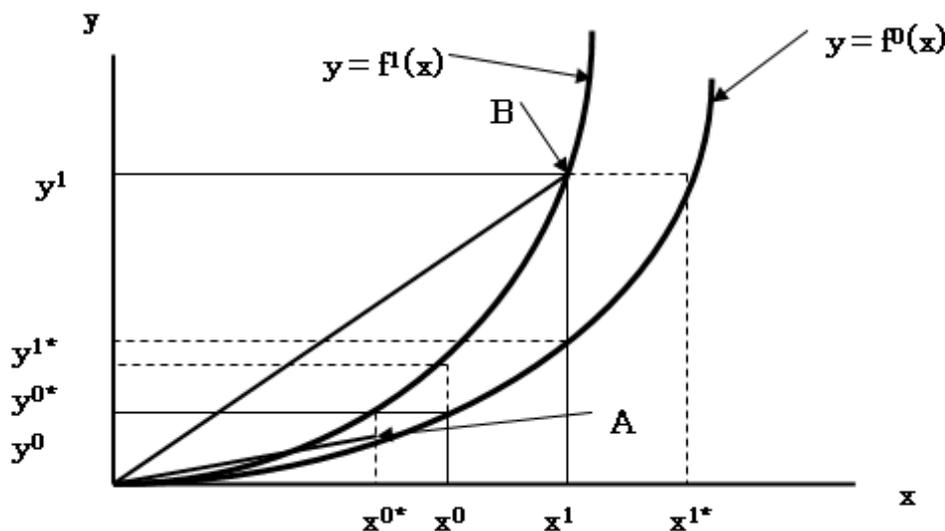


Fig. 7.1 Production Based Measures of Technical Progress

The lower curved line is the graph of the period 0 production function; i.e., it is the set of points  $(x, y)$  such that  $x \geq 0$  and  $y = f^0(x)$ . The higher curved line is the graph of the period 1 production function; i.e., it is the set of points  $(x, y)$  such that  $x \geq 0$  and  $y = f^1(x)$ . The observed data points are A, which has coordinates  $(x^0, y^0)$  and B, which has coordinates  $(x^1, y^1)$ . Note that the absolute amounts of production function shift in the direction of the  $y$  axis are  $y^{0*} - y^0$  (at point A) and  $y^1 - y^{1*}$  (at point B). The absolute amounts of production function shift in the direction of the  $x$  axis are  $x^0 - x^{0*}$  (at point A) and  $x^{1*} - x^1$  (at point B). We have chosen to measure TP in terms of the relative shifts,  $y^{0*}/y^0, y^1/y^{1*}, x^0/x^{0*}$  and  $x^{1*}/x^1$  rather than the absolute shifts,  $y^{0*} - y^0, y^1 - y^{1*}, x^0 - x^{0*}$  and  $x^{1*} - x^1$  in order to obtain measures of shift that are invariant to changes in the units of measurement. Note that  $TFPG = TFPG(2) = [y^1/x^1]/[y^0/x^0]$  is equal to the slope of the straight line OB divided by the slope of the straight line OA.

It turns out that there is a relationship between each of our technical progress measures, TP(1), TP(2), TP(3), TP(4), and total factor productivity growth, TFPG. We have:

$$TFPG = TP(i)RS(i); \quad i = 1, 2, 3, 4 \tag{7.23}$$

<sup>\*7</sup> TP(3) and TP(4) are the one input, one output special cases of Caves, Christensen, and Diewert's (1982; 1407)[49] input based 'productivity' indexes. However, in the present chapter, we regard these 'productivity' indexes as measures of the shift in the production functions and hence as measures of technical progress.

where the four returns to scale measures  $RS(i)$  are defined as follows:

$$RS(1) \equiv [y^1/x^1]/[y^{0*}/x^0]; \quad (7.24)$$

$$RS(2) \equiv [y^{1*}/x^1]/[y^0/x^0]; \quad (7.25)$$

$$RS(3) \equiv [y^1/x^1]/[y^0/x^{0*}]; \quad (7.26)$$

$$RS(4) \equiv [y^1/x^{1*}]/[y^0/x^0]. \quad (7.27)$$

The returns to scale measures  $RS(1)$  and  $RS(3)$  pertain to the period 1 production function  $f^1$  while the measures  $RS(2)$  and  $RS(4)$  pertain to the period 0 production function  $f^0$ . To interpret each of these returns to scale measures geometrically, see Figure 7.1. Each of these returns to scale measures is the ratio of two input-output coefficients, say  $[y^j/x^j]$  divided by  $[y^k/x^k]$ , where  $[y^j/x^j]$  and  $[y^k/x^k]$  are two points on the same production function and  $x^j > x^k$ . Thus, if the returns to scale measure  $[y^j/x^j]/[y^k/x^k]$  is greater than 1, then  $[y^j/x^j] > [y^k/x^k]$  and we say that the production function exhibits increasing returns to scale between the two points. If  $RS(i) = 1$ , then the production function exhibits constant returns to scale between the two points and finally if  $RS(i) < 1$ , then the production function exhibits decreasing returns to scale between the two points.

The total factor productivity growth decompositions given by equations (7.23) tell us that TFPG is equal to the product of a technical progress term  $TP(i)$  (this corresponds to a shift in the production function going from period 0 to period 1) and a returns to scale term  $RS(i)$  (this corresponds to a movement along one of the production functions). The reader can use Figure 7.1 and definitions (7.18)–(7.22) and definitions (7.24)–(7.27) to verify that each of the four decompositions of TFPG given by (7.23) corresponds to a different combination of shifts and movements along a production function that take us from point A to point B.

For firms in a regulated industry, returns to scale will generally be greater than one, since increasing returns to scale in production is often the reason for regulation in the first place. Thus, TFPG will exceed TP for growing firms in a regulated industry (provided that there are increasing returns to scale for that firm).

We note that the technical progress and returns to scale measures defined above cannot in general be calculated without a knowledge of the production functions that describe the technology for the two periods under consideration. However, in a one input, one output firm, the TFPG measures defined above can be calculated unambiguously provided that we know inputs used and outputs produced during the two periods.

In the final sections of this chapter, we shall generalise the above production function based definitions of productivity and technical progress to cover the case of many outputs and many inputs. However, in the following 5 sections, we will look at some of the institutional factors that might increase a country's TFP growth and possible roles for government to improve a country's TFP growth.

### 7.3 The Determinants of Economic Growth: Primary Input Growth and Other Factors

There are a variety of theories that attempt to explain why some countries grow faster than others. In a fairly recent review of the New Zealand economy, Bates (2001)[25] explains that the main determinants of output growth are *input growth* (the growth of capital and labour inputs) and the *growth of Total Factor Productivity* (TFPG). Bates goes on to note the importance of resource discoveries as another factor that can help explain growth. Resource discoveries and the exploitation of resources are somewhat important in the New Zealand context with agriculture, forests, oil and gas and perhaps fishing all playing a role. Other important factors in determining growth rates are:

- Changes in the terms of trade\*<sup>8</sup>;
- Immigration and population growth (obviously these factors influence the growth of labour input);
- Changes in domestic savings rates (this influences investment and the growth of capital input);
- Openness of the economy to foreign investment;
- Changes in the educational composition of the labour force;
- National entrepreneurial capacity\*<sup>9</sup> and
- The role of government in facilitating competition and the development of efficient markets.\*<sup>10</sup>

Bates (2001)[25] has an excellent discussion of the importance of institutions, property rights and government policies on growth rates; I can add little on his discussion of these factors.\*<sup>11</sup> However, the other factors listed above are also important.

In addition to the growth of primary inputs, the main factor which “explains” output growth is an increase in the *Total Factor Productivity* of the economy:

“The finding that TFP is largely responsible for differences in per capita incomes and growth rates does not take us very far. It merely implies that something other than capital accumulation may be important, without identifying what it is. Solow referred to this residual as ‘technical change’ and used this as a shorthand to cover education of the labour force and any kind of shift in the production function.” Winton Bates (2001; 11)[25].

---

\*<sup>8</sup> See Diewert and Morrison (1986)[146], Morrison and Diewert (1990)[314], Kohli (1990)[274] and Fox and Kohli (1998)[189] on how to measure this factor.

\*<sup>9</sup> Alfred Marshall (1898; 377)[305] described some of the characteristics of entrepreneurial ability as follows: “The ideal manufacturer, for instance, if he makes goods not to meet special orders but for the general market, must, in his first role as merchant and organizer of production, have a thorough knowledge of things in his own trade. He must have the power of forecasting the broad movements of production and consumption, of seeing where there is an opportunity for supplying a new commodity that will meet a real want or improving the plan of producing an old commodity. He must be able to judge cautiously and undertake risks boldly; and he must of course understand the materials and machinery used in his trade.”

\*<sup>10</sup> The development of efficient markets is a tricky business. Of course, there must be a legal system that will enforce contracts. But contracts are simply not able to deal with the complexities of real world transactions. Hence for markets to function efficiently, there must be *trust* among economic agents as they deal with one another, trust that they will not try to cheat each other and that each party to a transaction has faith that the other party will behave honorably as external conditions perhaps not originally envisioned change. As usual, Alfred Marshall (1898; 7)[305] had some pertinent observations on this point: “Again, the modern era has undoubtedly given new openings for dishonesty in trade. The advance of knowledge has discovered new ways of making things appear other than they are, and has rendered possible many new forms of adulteration. The producer is now far removed from the ultimate consumer; and his wrong-doings are not visited with the prompt and sharp punishment which falls on the head of a person who, being bound to live and die in his native village, plays a dishonest trick on one of his neighbours. The opportunities for knavery are certainly more numerous than they were; but there is no reason for thinking that people avail themselves of a larger proportion of such opportunities than they used to do. On the contrary, modern methods of trade imply habits of trustfulness on the one side and a power of resisting temptation to dishonesty on the other, which do not exist among a backward people.” I do not agree with Marshall that some people are necessarily “backward”. It seems that most market economies when in the early stages of their development have problems with determining what are appropriate rules of the game but as they evolve, appropriate norms also evolve.

\*<sup>11</sup> Hicks (1969)[224] had a nice discussion on how custom and command economies have historically evolved into market oriented economies. Alfred Marshall (1898; 24-25)[305] noted how governments can hinder the development of market economies: “Until a few years ago complete and direct self government by the people was impossible in a great nation: it could exist only in towns or very small territories. Government was necessarily in the hands of the few, who looked upon themselves as privileged upper classes, and who treated the workers as lower classes. Consequently, the workers, even when permitted to manage their own local affairs, were often wanting in the courage, the self-reliance, and the habits of mental activity, which are required as the basis of business enterprise. And as a matter of fact both the central Government and the local magnates did interfere directly with the freedom of industry; prohibiting migration, and levying taxes and tolls of the most burdensome and vexatious character. Even those of the lower classes who were nominally free, were plundered by arbitrary fines and dues levied under all manner of excuses, by the partial administration of justice, and often by direct violence and open pillage.”

TFP growth is output growth that cannot be directly explained by input growth.<sup>\*12</sup> Put another way, TFP growth is sometimes taken to be an upward shift in the private sector aggregate production function. Bates goes on to note that differences in TFP growth rates have been largely responsible for differences in growth rates and he provides a good review of the endogenous growth theories of Romer and others. However, there are other much older growth theories that might be relevant. In the following section, we provide a review of some of these older theories.

## 7.4 The Determinants of Economic Growth: Productivity Growth

How does the production function or production possibilities set of a country expand over time?<sup>\*13</sup> We think of the production possibilities set of a given firm as a given set of plans or operating procedures that are known to the management of the production unit. But where does this knowledge of the production possibilities set come from? And how does this knowledge expand over time; i.e., how does innovation and the expansion of society's feasible set of outputs occur?

Knowledge of the set of feasible input and output combinations that a business unit in a specific geographic location could use and produce during an accounting period comes from at least three sources: (i) operating manuals or other written (or computer accessible) materials that are available in the establishment; (ii) knowledge of production techniques that is embodied in employees and managers who work in the establishment and (iii) knowledge that is embedded in establishment machines. This provides a brief answer to the first question above.

Note that it may be difficult to separate a *shift* in the establishment production function (due to innovative activity) from a *movement along* a production function (due to a change in scale or to a change in input prices. This point was made by Hicks (1973; 120)[225] many years ago:

“I have so far been telling the story in the conventional terms, of shifts in technology and switches within the technology; but, at the point we have reached, do not the ‘technology’ and the ‘technological frontier’ themselves become suspect? They are essential tools of static analysis; but in dynamic analysis, such as this, do we need them? . . . The notion of a ‘technology’, as a collection of techniques, laid up in a library (or museum) to be taken down from their shelves as required, has been deservedly criticized; in itself it is a caricature of the inventive process . . . . Why should we not say that every change in technique is an invention, which may be large or small? It certainly partakes, to some degree, of the character of an invention; for it requires, for its application, some new knowledge, or some new expertise. There is no firm line, on the score of novelty, between shifts that change the technology and shifts that do not.”

We turn now to our second question; i.e., how does the production possibilities set of a country expand over time? Put another way, *how does knowledge of new techniques of production (process*

<sup>\*12</sup> Technically, TFP growth can be obtained by subtracting 1 from an output index divided by an input index; see Caves, Christensen and Diewert (1982)[49], Diewert (1992)[101] and Diewert and Nakamura (2003)[149] for further discussion. Lipsey and Carlaw (2000)[293] present a fairly devastating critique of the neoclassical production function approach to the measurement of TFP. However, I think that many of their criticisms of the TFP concept can be mitigated if we just define TFP as an output index divided by an input index. This focuses attention on the proper measurement of outputs and inputs and on the choice of functional forms for the indexes of output and input. Diewert (1997)[115] noted that four distinct approaches to the index number functional form problem led to either the Fisher (1922)[187] ideal index or the Törnqvist index as the preferred choice. We note that the index number approach to the measurement of TFP will summarize the *combined* effect of three separate effects: (i) technical progress (i.e., an outward expansion of the production possibilities set); (ii) increasing (or decreasing) returns to scale (a movement along the frontier of the production possibilities set) and (iii) increases in managerial or organizational efficiency (a movement towards the frontier of the production possibilities set).

<sup>\*13</sup> We follow Nordhaus (1969; 19)[321] in viewing an innovation as the introduction of a new process or vector of input-output coefficients into the economy.

*innovations) and of new products (product innovations) get created?* Traditional production theory as is embedded in general equilibrium theory is silent on this point (even though many economists have noted that knowledge creation cannot be regarded as exogenous<sup>\*14</sup> and critics<sup>\*15</sup> have noted this deficiency of traditional production theory).

Obviously, specialized schools, universities and publicly supported research labs are a primary source of the creation of new knowledge but a considerable amount of innovative activity is undertaken by individual inventors and the research departments of private firms.

Arrow<sup>\*16</sup> and others<sup>\*17</sup> have attributed increases in productivity (more output for the same amount of input) to experience or the incidental effect of new investments. Arrow (1962; 155-157)[11] explains his theory of innovation as follows:

“I would like to suggest here an endogenous theory of the changes in knowledge which underlie intertemporal and international shifts in production functions. The acquisition of knowledge is what is usually termed ‘learning’ and we might perhaps pick up some clues from the many psychologists who have studied this phenomenon . . . . I advance the hypothesis here that technical change in general can be ascribed to experience, that it is the very activity of production which gives rise to problems for which favorable responses are selected over time . . . . The first question is that of choosing the economic variable which represents ‘experience’ . . . . I therefore take instead cumulative gross investment (cumulative production of capital goods) as a index of experience”.

A somewhat similar theory of innovation was advanced by Allen (1983)[3] which he called *collective invention*.<sup>\*18</sup> Allen explained his theory as follows:

<sup>\*14</sup> “Analysis of production functions over the last twelve years has suggested strongly that (a) a major proportion of the increase in per capita income cannot be explained by increases in the capital-labor ratio, and (b) production functions differ strongly among nations and indeed among regions . . . . An economist could just leave the analysis at that, asserting that the causes which determine the amount of technological knowledge at any one time and place lie as much outside his province as the tastes which determine consumption patterns. But in fact, we know that significant quantities of resources are being expended by profit-making institutions on research and development . . . . Hence, it is suggested, we must regard the body of technological knowledge as the result as well as the cause of economic changes”. Kenneth J. Arrow (1969; 29)[12].

<sup>\*15</sup> “. . . the basic assumptions of economic theory are either of a kind that are unverifiable—such as that producers ‘maximise’ their profits or consumers ‘maximise’ their utility— or of a kind that are directly contradicted by observation—for example, perfect competition, perfect divisibility, linear-homogeneous and continuously differentiable production functions, wholly impersonal market relations, exclusive role of prices in information flows and perfect knowledge of all relevant prices by all agents and perfect foresight. There is also the requirement of a constant and unchanging set of products (goods) and of a constant and unchanging set of processes of production (or production functions) over time . . . . The latest theoretical models, which attempt to construct an equilibrium path through time with all prices for all periods fully determined at the start under the assumption that everyone foresees future prices correctly to eternity, require far more fundamental ‘relaxations’ for their applicability than was thought to be involved in the original Walrasian scheme”. Nicholas Kaldor (1972; 1238-1239)[265].

“Dynamic general equilibrium models with state contingent goods and convex production sets may be useful for some purposes, but the critics are right that there is something fundamental and important about the evolution of an economy that equilibrium models based on convex sets cannot capture”. Paul Romer (1994; 14)[336].

<sup>\*16</sup> “Knowledge arises from deliberate seeking, but it also arises from observations incidental or other activities. Haavelmo, Kaldor and I . . . have all stressed that the activities of production and investment may lead to increases in productivity without any identifiable allocation of resources to that end”. Kenneth J. Arrow (1969; 30)[12].

“The Horndal iron works in Sweden had no new investment (and therefore presumably no significant change in its methods of production) for a period of 15 years, yet productivity (output per manhour) rose on the average close to 2% per annum. We find again steadily increasing performance which can only be imputed to learning from experience”. Kenneth J. Arrow (1969; 156)[12].

<sup>\*17</sup> See Allen (1983)[3] and the references in Arrow (1962; 156)[11].

<sup>\*18</sup> “Who invents? Why do they invent? In attempting to answer these questions, economists have identified and studied three kinds of institutions—nonprofit institutions like universities and government agencies, firms

“Thus, if a firm constructed a new plant of novel design and that plant proved to have lower costs than other plants, these facts were made available to other firms in the industry and to potential entrants. The next firm constructing a new plant could build on the experience of the first by introducing and extending the design change that had proved profitable . . . . Collective invention was thus like modern research and development in that firms (and not individual inventors) generated the new technical knowledge. However, collective invention differs from R & D since the firms did not allocate resources to invention—the new technical knowledge was a by-product of normal business operation—and the technical information produced was exploited by agents other than the firms that discovered it”. Robert C. Allen (1983; 2)[3].

“As long as the rate of investment was high, the rate of experimentation and the discovery of new technical knowledge was also high. On the other hand, if the rate of investment fell for any reason, the rates of experimentation and invention fell with it”. Robert C. Allen (1983; 3)[3].

Allen illustrated his theory using data on changes in the height and operating temperatures of blast furnaces in England between 1850 and 1875 and he summarized his results as follows:

“Increasing furnace height and blast temperature led to lower fuel consumption and costs. The first firms to build tall furnaces might have treated this knowledge as a trade secret, but they did not. This information was made available to other parties through two channels: informal disclosure and publication in the engineering literature”. Robert C. Allen (1983; 6-7)[3].

Thus Allen modeled innovation as follows: as firms invested in new facilities, bolder firms undertook marginal changes in the design of their facilities or machines; successful design changes were then communicated to the industry as a whole through trade associations or formal publication in journals or magazines. It is interesting to note that Marshall advanced similar ideas many years ago.\*<sup>19</sup>

Arrow and Allen both saw *investment* as a key input into the innovation process. Many modern growth theorists have noted that improvements in total factor productivity are often associated with increased investment activity. This association has a structural basis for the following reasons:

- New scientific and engineering information is often embodied in new investment goods;
- If investment is stagnant or declining, then capital services into the economy are also declining and since the growth in labour and other primary inputs is generally small, the decline in capital input leads to an overall decline in inputs used by the economy and hence average fixed costs increase and there is a decline in the overall efficiency of the economy. We discuss the role of fixed costs and increasing returns to scale in more detail below.

The next batch of theories of innovation date back to the origins of economics. Adam Smith (1963; 8)[361] observed that many inventions or innovations are made by workers who simply figure out better ways of accomplishing a task that they are presently engaged in\*<sup>20</sup>:

“I shall only observe, therefore, that the invention of all those machines by which labour is so much facilitated and abridged, seems to have been originally owing to the division of labour. Men are much more likely to discover easier and readier methods of attaining any object when the whole attention of their minds is directed towards that single object, than

---

that undertake research and development, and individual inventors. In this paper, it is proposed that a fourth inventive institution be recognized. This institution is called collective invention”. Robert C. Allen (1983; 1)[3].

\*<sup>19</sup> “Again, it is to his interest also that the secrecy of business is on the whole diminishing, and that the most important improvements in method seldom remain secret for long after they have passed from the experimental stage. It is to his advantage that changes in manufacture depend less on mere rules of thumb and more on broad developments of scientific principle; and that many of these are made by students in the pursuit of knowledge for its own sake, and are promptly published in the general interest”. Alfred Marshall (1920; 285)[305].

\*<sup>20</sup> This is obviously related to Arrow’s learning by doing theory of productivity improvements.

when it is dissipated among a great variety of things. But in consequence of the division of labour, the whole of every man's attention comes naturally to be directed towards some one very simple object. It is naturally to be expected, therefore, that some one or other of those who are employed in each particular branch of labour should soon find out easier and readier methods of performing their own particular work, whenever the nature of it admits of such improvement".\*<sup>21</sup>

Smith also observed that many improvements in productivity result from the *specialization of labour*: a worker who is able to concentrate or specialize on one task will become more proficient at that single task due to: (i) improvements in dexterity or physical skill and (ii) the elimination of the fixed costs in going from one type of task to another:

"This great increase of the quantity of work, which, in consequence of the division of labour, the same number of people are capable of performing, is owing to three different circumstances; first, to the increase of dexterity in every particular workman; secondly, to the saving of the time which is commonly lost in passing from one species of work to another; and lastly, to the invention of a great number of machines which facilitate and abridge labour, and enable one man to do the work of many". Adam Smith (1963; 7)[361].

Note that Smith suggested a third productivity benefit due to the increased specialization of labour: *specialized routine operations by workers lend themselves to replacement by more efficient machines*. Marshall\*<sup>22</sup> and Young\*<sup>23</sup> made similar observations. These observations are still valid today; e.g., many clerical and lower level managerial jobs are being replaced by computers and other machines.\*<sup>24</sup>

---

\*<sup>21</sup> Smith (1963; 8-9)[361] illustrated this general statement by the following specific example:

"In the first fire-engines, a boy was constantly employed to open and shut alternately the communication between the boiler and the cylinder, according as the piston either ascended or descended. One of those boys, who loved to play with his companions, observed that, by tying a string from the handle of the valve which opened this communication to another part of the machine, the valve would open and shut without his assistance, and leave him at liberty to divert himself with his play-fellows. One of the greatest improvements that has been made upon this machine, since it was first invented, was in this manner the discovery of a boy who wanted to save his own labour".

\*<sup>22</sup> "We are thus led to a general rule, the action of which is more prominent in some branches of manufacture than others, but which applies to all. It is, that any manufacturing operation that can be reduced to uniformity, so that exactly the same thing has to be done over and over again in the same way, is sure to be taken over sooner or later by machinery . . . . Thus the two movements of the improvement of machinery and the growing subdivision of labour have gone together and are in some measure connected". Alfred Marshall (1920; 255)[305].

\*<sup>23</sup> "It is generally agreed that Adam Smith, when he suggested that the division of labour leads to inventions because workmen engaged in specialised routine operations come to see better ways of accomplishing the same results, missed the main point. The important thing, of course, is that with the division of labour a group of complex processes is transformed into a succession of simpler processes, some of which, at least, lend themselves to the use of machinery. In the use of machinery and the adoption of indirect processes there is a further division of labour, the economies of which are again limited by the extent of the market. It would be wasteful to make a hammer to drive a single nail; it would be better to use whatever awkward implement lies conveniently at hand. It would be wasteful to furnish a factory with an elaborate equipment of specially constructed jigs, gauges, lathes, drills, presses and conveyors to build a hundred automobiles; it would be better to rely mostly upon tools and machines of standard types, so as to make a relatively larger use of directly-applied and a relatively smaller use of indirectly-applied labour. Mr. Ford's methods would be absurdly uneconomical if his output were very small, and would be unprofitable even if his output were what many other manufactures of automobiles would call large". Allyn A. Young (1928; 530)[408].

\*<sup>24</sup> Nakamura and Lawrence (1994; 248)[320] have a nice analysis of some of the differences between machines and workers that might cause managers to substitute machines for workers: "The comparative advantages of using machine labour are readily apparent. Computers and computer controlled machines are consistent in their responses, time after time. Machines are not vulnerable to feelings of boredom, fears that technological change may render them obsolete, or inopportune promotion aspirations. They never get pregnant, ask for maternity leaves, file discrimination or harassment suits, object if they are not given training opportunities, demand to be paid time-and-a-half for overtime work, or strikes. When parts of machines wear out, they can be replaced (or the whole machine can be replaced) without concerns about Workers' Compensation or disability claims being filed. Machines may not always perform as desired, but this is never a consequence of hard-to-handle

Smith (1963; 14)[361] also pointed out *that the division of labour was limited by the extent of the market*; i.e., as the scale of the establishment grows due to the growth of markets for its outputs, the possibility of using specialized labour (and capital!) inputs also grows. As a corollary to his general principle, Smith pointed out that cities had larger markets than small towns and hence would support a higher degree of specialization in labour markets:

“There are some sorts of industry, even of the lowest kind, which can be carried on no where but in a great town. A porter, for example, can find employment and subsistence in no other place. A village is by much too narrow a sphere for him; even an ordinary market town is scarce large enough to afford him constant occupation”. Adam Smith (1963; 14)[361].

Hence smallness of the local market hinders specialization and the resulting increases in efficiency. *This point is extremely important for a small isolated economy like New Zealand.* Because of New Zealand’s smallness and geographic distance from major markets, it is difficult for New Zealand to provide specialized exports of goods and services to the world market and to develop a large variety of specialized domestic inputs. Consider the following quotation from *The Economist*, December 2, 2000:

“New Zealand’s small population and geographic isolation from large markets also limit its scope for exploiting economies of scale. As ‘the last bus stop on the planet’, New Zealand is at a disadvantage compared with other small economies such as Ireland or Finland. A circle with a radius of 2,200 kilometers centered on Wellington encompasses only 3.8 million people and a lot of seagulls. A circle of the same size centered on Helsinki would capture well over 300 million people. Even if New Zealand had the best economic policies in the world, its isolation would probably still constrain its growth rate.”

*The Economist* sums up its article on New Zealand’s economy as follows:

“New Zealand’s smallness and remoteness mattered less when it produced mainly for the British market and when people had less choice about where to work and invest. But in today’s more integrated world it is a serious handicap. As the OECD points out in its report, to offset its natural disadvantages, New Zealand needs to have better economic policies than other countries, if it is to be an attractive location for investment and for skilled workers to live. As other countries, notably in continental Europe, continue to liberalise their own economies, New Zealand’s policies are no longer so exceptional. By reversing its reforms now, New Zealand could snatch defeat from the jaws of victory.”

Alfred Marshall further refined Adam Smith’s idea that a larger market allows for increases in specialization and hence increased output for the same amount of aggregate input by introducing the ideas of *internal and external economies of scale*. In the following section, we shall examine his ideas and those of others on this topic in more depth.

## 7.5 Increasing Returns to Scale

We may divide the economies arising from an increase in the scale of production of any kind of goods, into two classes—firstly, those dependent on the general development of the industry; and, secondly, those dependent on the resources of the individual houses of business engaged in it, on their organization and the efficiency of their management. We may call the former external economies, and the latter internal economies”. Alfred Marshall (1920; 266)[305].

---

attitudes or substance abuse problems. Rather, straight-forward methods of scientific and engineering inquiry can usually be relied on to solve the performance difficulties of mechanical devices. And machines never have to be monitored to prevent them from intentionally shirking or stealing”.

*Internal economies of scale* occur if output expansion leads to a less than proportional increase in the use of inputs; i.e., internal economies are equivalent to increasing returns to scale in more modern language. The increasing returns to scale phenomenon could be regarded as meaning that the production possibilities set of an establishment has a particular shape and hence it might appear that the increasing returns to scale phenomenon can be accommodated by traditional production theory. This is true once a business unit has actually run an establishment at a higher scale and has demonstrated that the technology works at the higher output levels, but the first successful demonstration of operating a technology at a higher scale has much the same character as establishing the feasibility of an innovation.\*<sup>25</sup> In any case, the benefits due to a firm being able to increase its scale when its technology is subject to increasing returns to scale is entirely similar to a productivity improvement due to an innovation. Hence increasing returns to scale may help to “explain” where improvements in total factor productivity come from.

There appear to be *six main sources of internal economies of scale*:

- (1) *Simple Indivisibilities*; i.e., most labour and capital inputs cannot be purchased in fractional amounts and all capital inputs have upper and lower limits on their capacities.\*<sup>26</sup> Thus a tiny firm will generally have higher costs than larger firms because it cannot purchase its inputs in small enough amounts.
- (2) *Multiple Stages of Production Indivisibilities*. This source of increasing returns to scale is an extension of the first source to deal with the complexities of multistage production. It is due to Babbage (1835; 212)[15] and will be explained below.
- (3) *The Laws of Physics*; i.e., Kaldor\*<sup>27</sup> (and Marshall\*<sup>28</sup>) noted that the three dimensional nature of space leads to certain economies of scale.\*<sup>29</sup>
- (4) *The Laws of Geometry*. This source of increasing returns to scale was flagged by Lipsey (2000)[292] and it is closely related to the previous source. We discuss some of Lipsey’s examples below.
- (5) *The Existence of Fixed Costs*; i.e., these are the efficiencies which result from averaging or amortizing fixed costs (a kind of indivisibility) over higher output levels. Before a machine yields a benefit from its operation, it may require the services of an operator who may have to be transported from one location to another\*<sup>30</sup> and the machine may require a warming up

---

\*<sup>25</sup> Allen (1983; 10)[3] pointed out that increasing the height of blast furnaces eventually ran into diminishing returns: “These tall furnaces proved to be disasters”.

\*<sup>26</sup> For example, vehicles used to transport goods (trucks) cannot be constructed above and below certain capacities.

\*<sup>27</sup> “As was shown above, not all causes of increasing returns can be attributed to indivisibility of one kind or another and there is no reason to suppose that ‘economies of scale’ become inoperative above certain levels of production. There is first of all the steady and step-wise improvement in knowledge gained from experience—the so-called ‘dynamic economies of scale’ which have nothing to do with indivisibilities. But even in the field of ‘static’ or ‘reversible’ economies, there is the important group of cases which I described above as being due to the three dimensional nature of space—i.e., the fact that the capacity of, say, a pipeline can be quadrupled by doubling its diameter while the costs (in terms of labour and materials) are more nearly related to the diameter than to its capacity”. Nicholas Kaldor (1972; 1253)[265].

\*<sup>28</sup> “A ship’s carrying power varies as the cube of her dimensions, while the resistance offered by the water increases only a little faster than the square of her dimensions; so that a large ship requires less coal in proportion to its tonnage than a small one. It also requires less labour, especially that of navigation: while to passengers it offers greater safety and comfort, more choice of company and better professional attendance.” Alfred Marshall (1920; 290)[305].

\*<sup>29</sup> For a more recent discussion of this topic, see Lipsey (2000)[292].

\*<sup>30</sup> This example of a fixed cost is of course due to Adam Smith (1963; 7)[361]. A classic example of a returns to scale effect due to the existence of fixed costs is the square root inventory replenishment rule discovered by the industrial engineers Green (1915)[202] and Harris (1915; 48-52)[215], and the economists Allais (1947; 238-241)[2], Baumol (1952)[26], Tobin (1956)[372] and many others; see Whitin (1952; 503)[398] (1957; 32 and 230)[399] and Hadley and Whitin (1963; 3-4)[206] for additional references to the literature.

period before production can begin. These are examples of fixed costs whose effect becomes relatively smaller the greater the length of time that the machine is continuously operated.

- (6) *The Law of Large Numbers*; i.e., these are efficiencies that result from the laws of probability theory. For example, consider a power plant that uses a number of identical engines. If the probabilities of engine failure are independently distributed, then having one set of spare parts on hand will generally be sufficient whether the plant has one engine or ten engines. Similarly, a large bank will not require as high a proportion of cash reserves to meet random demands as a small bank.\*<sup>31</sup> In a similar vein, a large property insurance company whose risks are geographically diversified faces a smaller probability of bankruptcy than a small insurance company,\*<sup>32</sup> etc.

The fact that machines have lower limits on their size and upper limits on their capacities means that for any single manufacturing process, there will generally be an output level and a machine that will minimize the average costs of production.\*<sup>33</sup> Babbage (1835)[15] takes this observation one step further by considering how a factory or multistage manufacturing process could be arranged to produce the final output at minimum cost. To take a simple example, suppose a finally demanded product can be produced by two separate stages of production. Suppose that the average cost of production of the first stage can be minimized if 100 units are produced but the average cost of production of the second stage can only be minimized if 200 units are produced. Then obviously, the overall unit cost of production can only be minimized if we produce 200 units (or a multiple of 200 units using a replication argument). Thus the threshold level of output that is necessary to achieve overall economies of scale in producing a product that is manufactured in multiple stages will generally be higher than a simple average of the efficient threshold levels of output for each stage. Babbage\*<sup>34</sup> expressed this very subtle principle as follows:

---

\*<sup>31</sup> This application of probability theory to the determination of adequate bank reserves dates back to Edgeworth (1888; 122)[164]; for additional applications and references to the literature, see Whitin (1952; 506-511)[398] (1957; 234-236)[399] and Hadley and Whitin (1963; chapters 4-8)[206]. Edgeworth (1888; 124)[164] also applied his statistical reasoning to the inventory stocking problem faced by a restaurant or club and noted that optimal inventory stocks are proportional to the square root of anticipated demands: "Suppose now the number of members in the club to be doubled or trebled, while their habits are unaltered. At first sight it might appear that the reserve of provisions which the manager requires should increase proportionately. But the corrected theory is that the ratio of the new reserve to the old should not be two or three but the square root of two or three".

\*<sup>32</sup> Hicks gave great importance to this factor. "The evolution of the institutions of the Mercantile Economy is largely a matter of finding ways of diminishing risks." John Hicks (1969; 48)[224]. "Neither of these methods would in fact be as powerful as they have proved to be, if it were not for the possibility of spreading risks, the so-called 'Law of Large Numbers' which is the basis of Insurance. We know that the medieval Italians were acquainted with insurance contracts; maritime insurance, insurance against the loss of a cargo in transit, was already possible in the fourteenth century. To undertake a single insurance of this type— involving a small but significant chance of a large loss, with no more than a moderate gain in the other event to set against it— would be intolerably risky; but it must soon have been observed that by combining a number of such risks, if they were reasonably independent of each other, the risk could be greatly reduced. If this had not been perceived, insurance could not have developed, as we know it did. We cannot tell at what point it was observed that the same principle applied to banking." John Hicks (1969; 79)[224].

\*<sup>33</sup> If the demand for the output is large relative to the output level that minimizes average cost, then the optimal machine could in theory be replicated and the industry production function would exhibit approximate constant returns to scale for large industry outputs; see Samuelson (1967)[346] and Diewert (1981)[92] for arguments along these lines.

\*<sup>34</sup> Babbage (1835)[15] in his preface explains how he came to be the world's first industrial engineer (or management consultant): "The present volume may be considered as one of the consequences that have resulted from the Calculating-Engine, the construction of which I have been so long superintending. Having been induced, during the last ten years, to visit a considerable number of workshops and factories, both in England and on the Continent, for the purpose of endeavouring to make myself acquainted with the various resources of mechanical art, I was insensibly led to apply to them those principles of generalization to which my other pursuits had naturally given rise."

“When the number of processes into which it is most advantageous to divide it, are ascertained, then all factories which do not employ a direct multiple of this latter number, will produce the article at a greater cost. This principle ought always to be kept in view in great establishments, although it is quite impossible, even with the best division of the labour, to attend to it rigidly in practice. . . . But it is quite certain that no individual, nor in the case of pin-making could any five individuals, ever hope to compete with an extensive establishment. Hence arises one cause of the great size of manufacturing establishments, which have increased with the progress of civilization.” Charles Babbage (1835; 212-213)[15].

Babbage also noted that the growth of large factories facilitated the division of labour:

“Perhaps the most important principle on which the economy of a manufacture depends, is the *division of labour* amongst the persons who perform the work. The first application of this principle must have been made in a very early stage of society; for it must have soon been apparent, that a larger number of comforts and conveniences could be acquired by each individual, if one man restricted his occupation to the art of making bows, another to that of building houses, a third boats, and so on. This division of labour into trades was not, however, the result of an opinion that the general riches of the community would be increased by such an arrangement; but it must have arisen from the circumstance of each individual so employed discovering that he himself could thus make a greater profit of his labour than by pursuing more varied occupations. Society must have made considerable advances before this principle could be carried into the workshop; for it is only in countries which have attained a high degree of civilization, and in articles in which there is a great competition amongst the producers, that the most perfect system of the division of labour is to be observed.” Charles Babbage (1835; 169)[15].

Babbage then went on to give a list of principles which would lead to the most perfect system of the division of labour in factories:

1. *Of the time required for learning.* It will be readily admitted that the portion of time occupied in the acquisition of any art will depend on the difficulty of its execution; and that the greater number of distinct processes, the longer will be the time which the apprentice must employ in acquiring it. . . .
2. *Of waste of materials in learning.* A certain quantity of material will, in all cases, be consumed unprofitably, or spoiled by every person who learns an art; and as he applies himself to each new process, he will waste some of the raw material, or of the partly manufactured commodity. But if each man commit this waste in acquiring successively every process, the quantity of waste will be much greater than if each person confine his attention to one process; in this view of the subject, therefore, the division of labour will diminish the price of production.
3. Another advantage resulting from the division of labour is, *the saving of that portion of time which is always lost in changing from one occupation to another.* . . .
4. *Change of tools.* The employment of different tools in the successive processes is another cause of the loss of time in changing from one operation to another. If these tools are simple and the change of tools is not frequent, the loss of time is not considerable; but in many processes of the arts the tools are of great delicacy, requiring accurate adjustment every time they are used; and in many cases the time employed in adjusting bears a large proportion to that employed in using the tool. The sliding-rest, the dividing and the drilling-engine, are of this kind; . . .
5. *Skill acquired by frequent repetition of the same processes.* The constant repetition of the same process necessarily produces in the workman a degree of excellence and rapidity in his particular department, which is never possessed by a person who is obliged to execute many

different processes. . . .

6. *The division of labour suggests the contrivance of tools and machinery to execute its processes.* When each process, by which any article is produced, is the sole occupation of one individual, his whole attention being devoted to a very limited and simple operation, improvements in the form of his tools, or in the mode of using them, are much more likely to occur to his mind, than if it were distracted by a greater variety of circumstances. Such an improvement in the tool is generally the first step towards a machine.” Charles Babbage (1835; 170-174)[15].

The above observations on the effects of an increasing division of labour reducing unit costs owe much to Adam Smith but it can be seen that Babbage put his own spin on Smith’s observations.\*<sup>35</sup> Babbage concludes his discussion on the division of labour by deriving a *seventh new principle*:

*“That the master manufacturer, by dividing the work to be executed into different processes, each requiring different degrees of skill or force, can purchase exactly that precise quantity of both which is necessary for each process; whereas, if the whole work were executed by one workman, that person must possess sufficient skill to perform the most difficult, and sufficient strength to execute the most laborious, of the operations into which the art is divided.”* Charles Babbage (1835; 175-176)[15].

Babbage (1835; 176-186)[15] illustrated his new principle by describing in great detail the mechanics of making pins, which could be broken down into a number of distinct processes, each of which had its own labour requirements (of different skills). He found that the unit cost of a pin made by a single worker (who necessarily must be the most skilled) would exceed the unit cost of a pin made using his new principle by a considerable margin:

*“The pins would therefore cost, in making, three times and three quarters as much as they now do by the application of the division of labour.”* Charles Babbage (1835; 186)[15].

Finally, Babbage (1835; 212-213)[15] tied the above material on the mechanics of making pins into his multiple processes principle of optimum production, (2) above, in our list of six sources of returns to scale.

We note that industrial engineering, operations research and management science have developed mathematical techniques that enable the business unit to achieve internal economies of scale with respect to many of the six factors listed above.

We turn now to Lipsey’s (2000)[292] discussion of *geometry* as a source of increasing returns to scale. His first example is extremely simple and has to do with the mechanics of pasturing horses:

*“This example is chosen because its transparency allows the issues to be easily identified. It concerns a firm that is in the business of pasturing other people’s horses. One square unit of fenced space is required to accommodate one horse. The grass is free and the only production cost is the fence, which is continuously variable. When the firm wishes to pasture more horses, it increases the size of its one fenced field.”* Richard G. Lipsey (2000, 3)[292].

Thus if  $L$  is the length of fence used by the firm, its costs are proportional to  $L$  but its output is proportional to  $L^2$ . Thus the firm’s unit cost will be proportional to  $1/L$  and hence we have decreasing unit costs and increasing returns to scale. Lipsey stresses that the source of the increasing returns has nothing to do with indivisibilities:

*“There are no indivisibilities in this example. The physical nature of the capital good is unchanged and the area of the pasture is a continuous variable. The neoclassical production function, defined in terms of inputs of service flows, displays constant returns to scale. Yet*

---

\*<sup>35</sup> Babbage (1835; 175)[15] explicitly acknowledges the contributions of Smith.

there are scale economies. These are rooted in the geometry of our three-dimensional world. The fenced area increases with the square of the length of the fence, while the cost increases linearly with the length of the fence.” Richard G. Lipsey (2000, 4)[292].

Lipsey<sup>\*36</sup> gives several additional examples of scale effects that arise from geometrical relations:

“The geometrical relation governing any container typically makes the amount of material used, and hence its cost (given constant prices of the materials with which it is made), proportional to *one dimension less* than the service output, giving increasing returns to scale over the whole range of output (at least with respect to the inputs of materials). This holds for more than just storage. Blast furnaces, ships, and steam engines are a few examples of the myriad technologies that show such geometrical scale effects.

Costs of construction also often increase less than in proportion to the increase in the capacity of any container. Consider just one example. The capacity of a closed cubic container of sides  $s$  is  $s^3$ . The amount of welding required is proportional to the total length of the seams, which is  $12s$ . The amount of material required for construction is  $6s^2$ . So material required per unit of capacity is  $6/s$  while [per unit volume] welding cost is  $12/s^2$ . Not only are both of these rates falling as the capital good is reconfigured to increase its capacity, they fall at different rates.” Richard G. Lipsey (2000, 6)[292].

We turn now to a discussion of Marshall’s *external economies of scale*. Two examples are:

- reduced prices for inputs due to bulk purchasing<sup>\*37</sup> and
- the large scale of a business unit can translate into a large demand for inputs and this in turn can encourage specialized suppliers to come into existence.<sup>\*38</sup> Thus external economies of scale reflect favorable changes in the environment facing the expanding business unit (lower input prices and new intermediate input suppliers).

Another way of explaining the second example is that a large demander of intermediate or primary inputs may facilitate the specialization of suppliers, leading to lower unit input prices for the large demander.

In the following section, we list some related factors that help to explain TFP growth.

## 7.6 Other Factors that Might Explain Growth

What is the underlying cause of both internal and external economies? It seems that Adam Smith (1963; 14)[361] had the answer to this question: *growth of the market*. Some of the obvious *factors that facilitate growth of the market* are:

- transportation and infrastructure improvements<sup>\*39</sup>;

<sup>\*36</sup> Lipsey (2000)[292] also gives many examples of scale effects that arise from physical laws and from indivisibilities.

<sup>\*37</sup> Bulk purchasing means that the supplying firm may be able to achieve internal economies of scale and thus can offer lower selling prices.

<sup>\*38</sup> This observation is of course due to Adam Smith as we have seen. Krugman summarizes Marshall’s elaboration of Smith as follows: “It was Alfred Marshall who presented the basic classic economic analysis of the phenomenon. (Actually, it was the observation of industry localization that underlay Marshall’s concept of external economies, which makes the modern neglect of the subject even more surprising). Marshall (1920)[305] identified three distinct reasons for localization. First by concentrating a number of firms in an industry in the same place, an industrial center allows a pooled market for workers with specialized skills; this pooled market benefits both workers and firms . . . . Second, an industrial center allows provision of nontraded [i.e., non internationally traded] inputs specific to an industry in greater variety and at lower cost . . . . Finally, because information flows locally more easily than over great distances, an industrial center generates what we would now call technological spillovers . . . .” Paul Krugman (1991; 36-37)[283].

<sup>\*39</sup> Adam Smith (1963; 15)[361] was well aware of this factor: “As by means of water-carriage a more extensive market is opened to every sort of industry than what land-carriage alone can afford it, so it is upon the sea-coast,

- population growth<sup>\*40</sup>;
- reduction in trade barriers<sup>\*41</sup>;
- reduction of taxes on commodities, labour services and capital<sup>\*42</sup>;
- the provision of personal security and the security of property rights<sup>\*43</sup>;
- improvements in advertising and the transmission of information about products<sup>\*44</sup>;
- improvements in communications<sup>\*45</sup>; and
- growth of physical and human capital.

The role of population growth in facilitating the growth of the market should take into account the growth of *rural versus urban* population since it is the growth of population in the *cities* of a country that leads to the growth of specialized markets. Thus for the first 60 or 70 years of the past century, growth in the cities of most advanced economies was fueled not only by higher rates of natural population growth than prevail now but urban growth was also fueled by migration of workers from the farm to the city. These within country population shifts helped fuel the productivity boom in the previous century that fell off after the first OPEC price shock in 1973.<sup>\*46</sup>

---

and along the banks of navigable rivers, that industry of every kind naturally begins to subdivide and improve itself, and it is frequently not till a long time after that those improvements extend themselves to the inland parts of the country”.

<sup>\*40</sup> “... every increase in [population] is likely for the time to be accompanied by a more than proportionate increase in their power of obtaining material goods. For it enables them to secure the many various economies of specialized skill and specialized machinery, of localized industries and production on a large scale: it enables them to have increased facilities of communication of all kinds; while the very closeness of their neighbourhood diminishes the expense of time and effort involved in every sort of traffic between them, and gives them new opportunities of getting social enjoyments and the comforts and luxuries of culture in every form. No doubt deduction must be made for the growing difficulty of finding solitude and quiet and even fresh air: but there is in most cases some balance of good.” Alfred Marshall (1920; 320-321)[305]. Perhaps Marshall could be considered the first environmentalist!

<sup>\*41</sup> As tariffs were reduced in the years following World War II, trade between countries grew faster than GDP growth. The North American Free Trade agreement led to a 75% increase in trade between Canada and the U.S. in 5 years.

<sup>\*42</sup> Bates (2001)[25] has an instructive example showing how high rates of labour taxation can cause taxpayers to allocate their time to doing various domestic chores like mowing the lawn instead of contracting specialized service providers. Thus high taxes inhibit the formation of specialized markets. A Canadian economist, William Watson (1999)[395] explains the problem as follows: “I spent Labour Day, fittingly, at work. . . . I was scraping my front porch and filling the holes with wood filler, in preparation for painting it . . . Objectively speaking, the reason I found myself scraping and patching was taxes. My comparative advantage, as we economists say, is typing, not hand tools. I should really be paying someone else to paint the front porch. The reason I don’t is taxes. Taxes mean I have to pay roughly four times what the job is worth. First, because my marginal rate is 50 plus per cent, I have to earn twice as much in pre-tax income as a painter would charge me. And, depending on the painter’s income tax rate and GST status, he has to charge me close to twice what he wants in after-tax income. Two times two being four (even in Tax-land), to pay for the job, I end up having to earn four times what the folks I would hire think their time is worth.”

<sup>\*43</sup> Bates’ (2001; Chapter 2)[25] discussion on this topic is more than adequate. Obviously physical intimidation and corruption is not conducive to economic growth in a country. Corruption acts like an uncertain tax on investments and hence will deter investment.

<sup>\*44</sup> Advertising makes potential purchasers aware of new products and thus stimulates market growth. A particularly effective recent innovation in this area is use of targeted mailing and email lists.

<sup>\*45</sup> Particularly important today are the improvements in telecommunications technology (fax machines, the internet, etc.). The recent growth in business to business provision of services over the internet should make it possible for New Zealanders to compete in international services markets. Communications improvements were also important in Marshall’s time: “Meanwhile an increase in the aggregate scale of production of course increases those economies, which do not directly depend on the size of individual houses of business. The most important of these result from the growth of correlated branches of industry which mutually assist one another, perhaps being concentrated in the same localities, but anyhow availing themselves of the modern facilities for communication offered by steam transport, by the telegraph and by the printing-press”. Alfred Marshall (1920; 317)[305].

<sup>\*46</sup> See Diewert and Fox (1999; 255-257)[137] for the falloff in OECD country TFP growth after 1973.

The previous paragraph makes the point that an economy's productivity will be improved if workers are shifted from lower productivity jobs in agriculture to higher productivity jobs in manufacturing and services. However, the same point applies to shifts of workers from lower to higher productivity establishments *within an industry*. John Haltiwanger (2000; 16)[211] sums up recent research in this area as follows:

“In this study we have focused on the contribution of the reallocation of activity across individual producers in accounting for aggregate productivity growth. A growing body of empirical analysis reveals striking patterns in the behaviour of establishment-level reallocation and productivity. First, there is a large ongoing pace of reallocation of outputs and inputs across establishments in market economies. Second, the pace of reallocation varies secularly, cyclically and by industry. Third, there are large and persistent productivity differentials across establishments in the same industry even in well functioning market economies. Fourth, entering establishments tend to have higher productivity than exiting establishments. Large productivity differentials and substantial reallocation are the necessary ingredients for an important role for reallocation in aggregate productivity growth. The emerging evidence suggests that the process of economic growth at the micro level is incredibly noisy and complex there is a vast amount of churning as businesses and workers seek to find the best methods, products, locations and matches. This churning is an inevitable and vital component of economic growth. However, a number of conceptual and measurement issues remain. We don't have a clear understanding of the sources of within and between country variation in the nature and magnitude of this churning, we don't have a clear understanding of the sources of within industry heterogeneity in productivity levels and growth rates, and in turn we don't have a clear understanding of all of this variation for within and between country outcomes like economic growth. A key obstacle for current work is that the requisite data development is still in early stages.”

Harberger (1998)[212] makes many of the same points as Haltiwanger in his Presidential Address to the American Economic Association. The available evidence indicates that establishments in the same industry differ tremendously in their efficiency and it can take long periods of time before these inefficient establishments are driven out of business. This indicates a potentially large role for business consultants or governments to bring a knowledge of best practice techniques to the attention of the inefficient establishments.<sup>\*47</sup> There is another important role for government and that is to facilitate the reallocation of resources from inefficient establishments to efficient ones. Thus other things being equal, it is likely that a highly regulated economy will not do as well as one where it is easy to set up new businesses and to hire (and fire) workers.<sup>\*48</sup>

As Bates explains in his Chapter 3, it is well known that tariffs and taxes have excess burdens associated with them. As long as these taxes and tariffs are not *increased*, they should not affect growth rates (in theory); they should only affect the *level* of economic activity. However, Feldstein, hints at a possible dynamic effect of high taxes on labour supply:

“The relevant distortion to labour supply is not just the effect of tax rates on participation rates and hours but also their effect on education, occupational choice, effort, location, and all the other aspects of behavior that affect the short-run and long-run productivity and income

<sup>\*47</sup> Often this can be accomplished by benchmarking exercises where similar establishments are compared. See Zeitsch and Lawrence (1996)[409] and Diewert and Nakamura (1999)[148].

<sup>\*48</sup> However, on social grounds (and even on efficiency grounds) it would be desirable to have an effective unemployment insurance scheme that would offer laid off workers temporary income support as they searched for new jobs. What is not desirable is a scheme that discourages interregional labour mobility or a scheme that encourages seasonal workers to stay in seasonal jobs. See Nakamura and Diewert (2000)[319] on recent reforms to the Canadian Employment Insurance scheme. The recent great expansion in internet job market companies should also improve the efficiency of labour markets.

of the individual.” Martin Feldstein (1996; 22)[178].

Thus high or increasing marginal rates of taxation on labour income can discourage individuals from using their after tax income to invest in higher education or specialized training courses, given that any increases in earnings might be taxed at very high rates.<sup>\*49</sup> This has the effect of hindering the growth of specialized labour markets and will divert effort into untaxed leisure or inefficient home production.

We return to our analysis of factors that might explain Total Factor Productivity growth. One such factor is of course the creation of *new* scientific and engineering knowledge. The creation of new knowledge is fairly well understood and will not be discussed here. However, what is perhaps most relevant for New Zealand is *not* the *initial creation* of the new knowledge but its *diffusion* to the local establishment level. The fact that a new product or production process has been developed somewhere in the world is of little significance to a local establishment that could use the innovation if the original knowledge is not transmitted or diffused to the establishment. Some of the factors that facilitate the rapid diffusion of new (and old) knowledge into a local market area are:

- access to public libraries and university libraries<sup>\*50</sup>;
- access to newspapers, periodicals, journals, magazines, how to do it books, etc.<sup>\*51</sup>;
- memberships in trade associations, industry associations, professional societies, etc.<sup>\*52</sup>;
- access to international meetings and trade fairs where knowledge can be transmitted on a face to face basis<sup>\*53</sup> (adequate local transportation infrastructure will facilitate this access<sup>\*54</sup>);
- access to good schooling and specialized training programs<sup>\*55</sup>;
- access to specialized consulting services<sup>\*56</sup> and product information and

<sup>\*49</sup> If there is progressive income taxation and variable labour supply, then investments in education can push the individual into a higher tax bracket. Under these conditions, we can expect a fair amount of excess burden; see for example Driffill and Rosen (1983)[161] or Dupor, Lochner, Taber and Wittekind (1996)[163]. This literature is reviewed by Kesselman (1997; 47-49)[267].

<sup>\*50</sup> It is here where basic information on science and engineering can be obtained: “Let us then look at those elements of the wealth of a nation which are commonly ignored when estimating the wealth of the individuals composing it . . . . Scientific knowledge indeed, wherever discovered, soon becomes the property of the whole civilized world, and may be considered as cosmopolitan rather than as specially national wealth. The same is true of mechanical inventions and of many other improvements in the arts of production. . . .” Alfred Marshall (1920; 59)[305].

<sup>\*51</sup> “For External economies are constantly growing in importance relatively to Internal in all matters of Trade-knowledge: newspapers, and trade and technical publications of all kinds are perpetually scouting for him and bringing him much of the knowledge he wants—knowledge which a little while ago would have been beyond the reach of anyone who could not afford to have well-paid agents in many distant places . . . . Although therefore the small manufacturer can seldom be in the front of the race of progress, he need not be far from it, if he has the time and the ability for availing himself of the modern facilities for obtaining knowledge”. Alfred Marshall (1920; 284-285)[305].

<sup>\*52</sup> “But perhaps a greater though less conspicuous hindrance to the rise of the working man is the growing complexity of business. The head of a business has now to think of many things which he never used to trouble himself in earlier days; and these are just the kind of difficulties for which the training of the workshop affords the least preparation. Against this must be set the rapid improvement of the education of the working man not only at school, but what is more important, in after life by newspapers, and from the work of co-operative societies and trades-unions, and in the other ways”. Alfred Marshall (1920; 308-309)[305].

<sup>\*53</sup> “While mass media play a major role in alerting individuals to the possibility of an innovation, it seems to be personal contact that is most relevant in leading to its adoption. Thus, the diffusion of an innovation becomes a process formally akin to the spread of an infectious disease”. Kenneth J. Arrow (1969; 33)[12].

<sup>\*54</sup> Having an easily accessible local airport that has direct flights to many international destinations seems to be important in this respect.

<sup>\*55</sup> “Other things being equal, one person has more real wealth in its broadest sense than another, if the place in which the former lives has a better climate, better roads, better water, more wholesome drainage; and again better newspapers, books and places of amusement and instructions”. Alfred Marshall (1920; 58-59)[305].

<sup>\*56</sup> Several companies provide *benchmarking services*; i.e., the performance of a given production unit is compared to peer group units that face similar operating conditions. If inefficiency is revealed, then the given unit can attempt to duplicate the techniques used by the most efficient units; see for example Zeitsch and Lawrence

- access to telecommunications services<sup>\*57</sup> (i.e., having good local telecommunications infrastructure).

The point that we are trying to make here is that a small country does not *necessarily* have to devote a high percentage of its resources to primary research and development (i.e., to the creation of new products and processes): it need only have easy access to the sources of new knowledge.<sup>\*58</sup>

Jon Kesselman, in a private communication notes that the above paragraph may be a serious oversimplification as he indicates below:

“Even if a nation has *access* to all of these ways of getting information, it seems to me that the most critical factor is whether there are large numbers of well-educated, well-trained, and motivated, creative, and able individuals actually at work in industry. Having low-cost access to all of these informational resources will do an economy no good if there are not sufficient numbers of workers to seek, absorb, integrate, and apply the ideas. Moreover, the creation of new products, new industrial processes, and new business management and marketing techniques relies on having large numbers of these energetic, able, creative workers. If a nation’s workers are always imitating what is the well-documented ‘best technique’ in other countries, they will be well behind the curve of the best technology that is actually in development and in the earlier stages of application in the world’s leading companies. Don’t the big economic rents come from those who develop the new products and processes first? If they are able to patent, trademark, or otherwise keep some proprietary ownership of their discoveries, then they can either keep the gains to themselves or earn the royalties from others who wish to use their discoveries. Given today’s fast changing technology and new products, firms that start producing the goods or implementing the new processes one or two years after they were first developed are entering only when the ‘invention’ has become a ‘commodity’ with much lower economic returns.”

It should be noted that expending resources on the development of new products is not necessarily productive. Paul Romer noted that there are costs associated with expenditures on developing new products and processes:

“Every real economy is presented with an almost incomprehensible number of new goods that can be introduced. Some of these goods are like good Z in Figure 3. They would increase utility. Many others, perhaps the great majority of all possible new goods, would not be worth introducing. The fixed costs are too high and the benefits too low. Out of the enormous set of possible new goods, a very small number are somehow selected and introduced.” Paul Romer (1994; 14)[336].

We have reproduced the essence of a diagram due to Romer (1994; 13)[336] in Figure 7.2 below. There are two commodities in the economy: “old” commodities Y and a new commodity Z. The production possibilities set for the “old” economy before the introduction of the new commodity is the line segment OB. The fixed costs of developing the new commodity are equal to the line segment AB. Once these fixed costs have been paid, the production possibilities set for the new economy is the set enclosed by the production frontier DA, which is tangent to an indifference curve at the point E. Note that the introduction of the new commodity in this particular case has led to a higher utility level (OC in terms of old goods) than the utility level achieved by the economy

---

(1996)[409]. For references to the early history of benchmarking (which has its origins in the early industrial engineering literature), see Diewert and Nakamura (1999)[148]. This last study shows vast differences in the efficiency of diesel electric power plants around the world.

<sup>\*57</sup> It seems likely that internet services will eventually be substitutes for most of the knowledge transmission activities listed above.

<sup>\*58</sup> Japan might be an example of a country that focuses on using and commercializing basic research that has been done offshore.

before the introduction of the new commodity, OB. However, if the fixed costs of developing the new commodity, AB, were greater than the benefit AC, then it would not pay to introduce the new commodity.

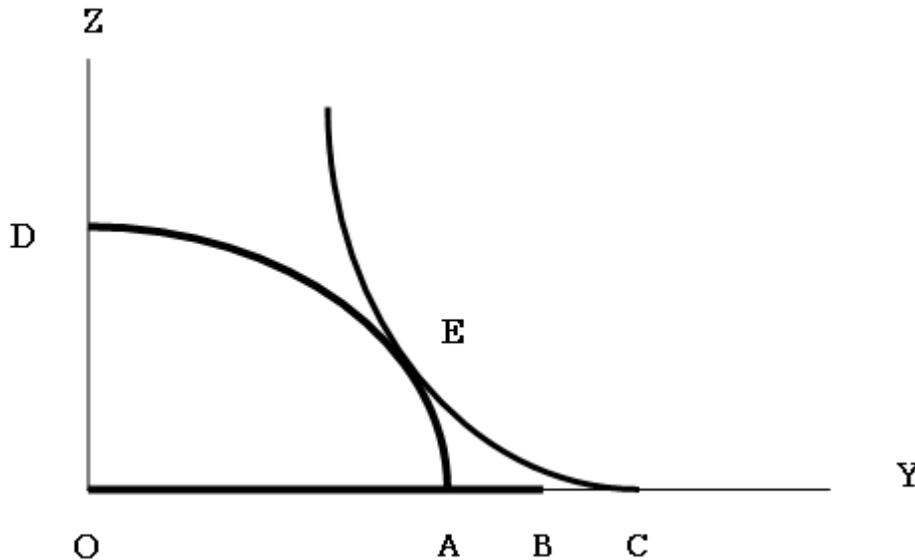


Fig. 7.2 The Costs and Benefits of New Products

The main point to note about Romer's observation is this: it is by no means certain that all expenditures on the development of new processes and products are beneficial.

We conclude our survey of causal factors that might help to explain total factor productivity growth by discussing the role of *macroeconomic stability*. The main factors here are stable fiscal policy, stable exchange rates, stable prices and stable interest rates. Macroeconomic stability by itself is not a main driver of productivity growth. However, a lack of stability often has strong *negative effects* on the rate of TFP growth. In particular, it appears that high or variable rates of inflation may have a such a negative effect. Diewert and Fox (1999)[137] identified *two mechanisms* that may cause high inflation to drive down TFP growth rates:

- Usually business income tax systems are not indexed to take into account the effects of inflation. Under these circumstances, even moderate inflation can lead to effective tax rates that diminish the real capital of businesses.
- Because many multiproduct firms did not know how to adjust depreciation allowances and user costs of capital for the effects of inflation, they were unable to properly price their products. As the capital intensity of advanced economies increased during the past century (and as firms greatly expanded the variety of products being produced), these effects of these pricing mistakes became relatively larger.<sup>\*59</sup>

Of course, under conditions of high and variable inflation, it becomes very difficult to determine the current real interest rate and to forecast future real interest rates. This increased uncertainty leads to incorrect investment decisions and a loss of efficiency for the economy. This is a *third mechanism* by which inflation translates into a loss of productive efficiency.

<sup>\*59</sup> Diewert and Fox (1999)[137] argue that the increase in inflation that started with the first oil shock in 1973 in most OECD economies and lasted until the early 1990's is the only major factor that changed abruptly that could perhaps explain the pronounced drop in TFP that hit virtually all OECD countries during that period.

In the following section, we attempt to summarize our rather diffuse discussion in this section.

## 7.7 A Summary of the Factors Explaining Productivity Growth

The previous sections gave an overview of factors that might explain variations in the growth of Total Factor Productivity, which in turn is the main driver of growth in per capita incomes. In summary, factors that will tend to *augment* TFP growth are:<sup>\*60</sup>

- Rapid *investment growth* (in reproducible or physical capital).<sup>\*61</sup>
- Rapid growth in *investments in education, training and human capital*.
- Rapid *growth in primary inputs* will tend to lead to an even more rapid growth of output due to *increasing returns to scale in production*. The main drivers of increasing returns to scale are: the existence of indivisibilities; the laws of geometry and physics; the existence of fixed costs and the laws of statistics.
- Increases in TFP are associated with *increased specialization*, which in turn is driven by growth in the size of the market. In brief: big tends to be better!
- *Improvements in the functioning of markets*, which could occur in a variety of ways, including: (i) improvements in personal security; (ii) improvements in property rights; (iii) reductions in trade barriers; (iv) improvements in telecommunications (in particular, the growth of internet driven markets) and (v) improvements in transportation and infrastructure.
- *Access to new knowledge* about the development of new commodities and processes. Recall our discussion about the importance of business consultants, trade associations and benchmarking to diffuse knowledge about best practices.

Factors which will tend to *reduce* the growth of Total Factor Productivity (in addition to the negative of the above factors) are:

- *High taxes*. In theory, this factor should just have one time level effects on economic efficiency but it is likely that high taxes have dynamic effects as well, tending to reduce investments in physical and human capital and retarding the formation of specialized markets.<sup>\*62</sup>
- *High inflation*. High or unpredictable rates of inflation tend to increase uncertainty about the real interest rate and future prices and hence lead to a misallocation of investment and a reduction in productive efficiency.

---

<sup>\*60</sup> Harris (2001; 5)[217] has the following list of factors that explain growth in what he calls the *modern macroeconomic growth perspective*: A. *Supply Side Growth Factors*: (1) Primary inputs (labour, resources); (2) Reproducible capital goods (physical and human capital); (3) Technology, management and the knowledge base; (4) Allocative efficiency of markets and external spillovers; (5) International comparative advantage; (6) Terms of trade; and (7) Public policy. B. *Demand Side Factors*: (1) External market access; (2) The global business cycle; and (3) Domestic macroeconomic policy. It can be seen that our list of explanatory factors is not all that different.

<sup>\*61</sup> Harris (2001)[217] characterizes productivity growth as being driven by three main factors: “The outcome of these studies, and of a host of other country-specific studies, has led to what I would call a consensus view on the three main correlates of national productivity growth — let’s call them the Big 3. They are, respectively, investment in machinery and equipment, human capital development, and openness to trade and investment. In the literally hundreds of studies that have been done these three variables show up as robustly and highly correlated with productivity growth or growth in per capita GDP.”

<sup>\*62</sup> In addition to the extensive literature cited by Bates (2001)[25] on the statistical relationships between taxation and growth rates, there are two more recent studies that could be cited. Kneller, Bleaney and Gemmill (1999)[270] find that capital taxation is more damaging to growth than taxes on consumption or labour as does Kesselman (2000; 47-57)[268]. Kesselman (2000; 55)[268] sums up his reading of the literature on this topic as follows: “What can be learned from the economic studies and comparative international experience is that taxing ‘smarter’ is more important than taxing less when promoting economic growth. Either shifting the total revenue mix toward greater reliance on indirect taxes on goods and services or on payroll-type taxes, or reforming the personal tax base to be more consumption oriented and less reliant on savings and capital incomes, would pay significant economic dividends.”

We turn now to a discussion of the role of government in optimizing growth.

## 7.8 The Role of Government in Facilitating Growth

Immediately above, we listed 6 factors that will tend to increase TFP growth and 2 factors that will tend to decrease it. Let us look at each of these factors in turn and see what possibilities there are for the government to optimize any of these factors.

(1) *Rapid investment growth (in reproducible or physical capital).*

Low rates of business income taxation are the key here. This leads to an argument for *smaller* government.

(2) *Rapid growth in investments in education, training and human capital.*

To optimize this factor, the government should aim for low rates of taxation on labour earnings in order to encourage individuals to invest in their human capital. Given the difficulties that individuals have in accessing capital markets in order to finance investments in education and training, there is also an argument for the government to subsidize these human capital investments. Hence the first argument implies a *smaller* government while the second argument implies a *larger* one.

(3) *Rapid growth in primary inputs.*

Low rates of taxation on business income will tend to encourage investment in physical capital while low rates of taxation on labour earnings and consumption will tend to encourage the growth of labour input. These are arguments for *smaller* government. There are other policies that the government can implement that can encourage primary input growth without much in the way of budgetary implications. Perhaps the most important of these policies might be to encourage immigration. Immigrants with large endowments of human and physical capital are particularly desirable as are immigrant groups who historically have had high labour force participation rates or high propensities to invest in physical and human capital. Another set of policies that might encourage primary input growth but are not necessarily very costly are associated with the exploitation of natural resources. In particular, the subsidization of tree planting comes to mind.\*<sup>63</sup>

(4) *Increased specialization and growth of the market.*

Lower taxes on business and labour income should facilitate increased specialization. This is another argument for *smaller* government. The government should also explore possible free trade agreements with its major trading partners. The budgetary implications of this policy are small.

(5) *Improvements in the functioning of markets.*

This factor includes: (i) improvements in personal security; (ii) improvements in property rights; (iii) reductions in trade barriers; (iv) improvements in telecommunications (in particular, the growth of internet driven markets) and (v) improvements in transportation and infrastructure. There is little that the New Zealand government could do in areas (i) to (iii) above. For many countries, the above factors will lead to arguments for a *larger* government.

(6) *Access to new knowledge about the development of new commodities and processes.*

Obviously, small countries will have difficulties in making their higher education sectors major players in the development of new knowledge. However, one could still make an argument for subsidizing Universities for two reasons:

---

\*<sup>63</sup> Poor resource management can of course negatively impact long run growth. The management of the cod fishery on the east coast of Canada has led to a complete shutdown of the fishery!

- Some Universities in relatively isolated locations have still managed to be incubators for high tech firms. The University of Waterloo in Canada comes to mind as do the University of British Columbia in Vancouver and the University of Alberta in Edmonton. Thus it may make sense to subsidize engineering, science, medicine and business departments in particular.\*<sup>64</sup>
- In order to transfer knowledge from abroad, it is necessary that the country have access to higher education in order to facilitate the diffusion and transfer processes.

The above considerations lead to an argument for a *larger* government. In addition to subsidizing higher education, there are some other things a government could do to facilitate knowledge transfers that do not have large budgetary implications. In particular, the government could publicize *benchmarking* to its business community so that the performance of local businesses could be compared to their peers offshore. Benchmarking of government enterprises and regulated enterprises should also be encouraged.

(7) *High taxes.*

We have discussed this factor already. Without taking into account to what purpose additional tax revenue would be used, it appears that growth is enhanced by having lower taxes in high tax and spending jurisdictions.

(8) *High inflation.*

As discussed above, economic growth will tend to be larger if the inflation rate is low and stable. Most countries have already achieved this so all that is needed is more of the same. This growth factor has no implications for the size of government.

Summing up, a detailed study of each growth factor would have to be undertaken to determine the optimal level of government expenditure in each of the various areas. However, a number of policies which would not require much government expenditure were mentioned above and should perhaps be considered.

There is one other role for government which should be mentioned explicitly at this point. An important role for government in encouraging growth is to create a transparent institutional environment which neither rewards nor tolerates rent seeking behaviour. In particular, business subsidies that are targeted to friends of the government or that are awarded on an almost random basis invites the diversion of resources from productive activity to wasteful rent seeking. In addition, the creation of an unlevel playing field will induce a loss of productive efficiency and slower productivity growth. Related to this point is the necessity for the government to create a transparent and effective system of business regulation. One need not look further than at the recent Californian energy crisis, which was created by a regulatory environment that did not encourage long term investment in power plants.

We conclude our review of the role of government by noting that it is also possible to do some fine tuning on the *tax collection* side of government as well as on the *expenditure* side. As Bates (2001; 52)[25] notes, the deadweight loss associated with a particular tax rises roughly as the square of the tax rate and is roughly proportional to the sum of the relevant magnitudes of the elasticities of supply and demand. Unfortunately, this means that for the government to set tax rates so as to minimize deadweight losses, a knowledge of elasticities of supply and demand is required. This knowledge is not easy to obtain. There are very few optimal tax studies that actually estimate empirically elasticities of supply and demand\*<sup>65</sup>. However, from the limited empirical evidence that

\*<sup>64</sup> Of course, science departments also require good mathematics departments, engineering departments require science departments and business schools require good economics departments. All faculties need their students to have some skills in English and of course, it is always good for science and engineering students to have some access to the arts and humanities.

\*<sup>65</sup> In addition to the work of Diewert and Lawrence (1994)[141] (2002)[143] noted by Bates, there is the work of Jorgenson and Yun (1986)[262] (1990)[263] (1991)[264].

is available, it appears that the marginal excess burden of taxing capital is higher than the marginal excess burdens of taxing labour or consumption. If this were true, then it would be desirable for a country to have a system of business income taxation that is at least as favorable as its major trading partners. For some recent papers on this topic, see Mintz (1999)[313] Harris (1999)[216]<sup>\*66</sup>, Kesselman (2000)[268] and Walsh (2000)[388].

Our overall conclusion is that the exact determinants of TFP growth are still not *precisely* known but, broadly speaking, the most important factors are probably known to us and are summarized in section 7.7 above. The present section looks at the possible role of governments in improving TFP growth. It is likely that at least some of the suggestions made in this section may be helpful in improving a country's TFP growth.

In the next section, we follow up on the technical material on measuring TFP growth that was introduced in section 7.2 but now we consider the case of production units that produce many outputs and use many inputs.

## 7.9 The Index Number Approach to the Measurement of Productivity

Recall our first definition of productivity growth in the one output, one input case (7.3),  $\text{TFPG}(1) \equiv [y^1/y^0]/[x^1/x^0]$ , which was the output ratio divided by the input ratio between periods 0 and 1. In order to find a counterpart to this definition in the multiple output, multiple input case, we need only replace the output ratio by an output quantity index,  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$ , and replace the input ratio by an input quantity index,  $Q^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)$ , where  $\mathbf{p}^t \equiv [p_1^t, \dots, p_M^t]$  and  $\mathbf{w}^t \equiv [w_1^t, \dots, w_N^t]$  are the period  $t$  output and input price vectors and  $\mathbf{y}^t \equiv [y_1^t, \dots, y_M^t]$  and  $\mathbf{x}^t \equiv [x_1^t, \dots, x_N^t]$  are the period  $t$  output and input quantity vectors for  $t = 0, 1$ . Thus an *output quantity index*,  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$ , is defined to be a function of the output prices and quantities for the two periods under consideration. Similarly, an *input quantity index*, between periods 0 and 1,  $Q^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)$ , is simply a function of  $4N$  variables, the input prices and quantities pertaining to the two periods under consideration.

Two of the most frequently used functional forms for quantity indexes are the Laspeyres (1871)[285] and Paasche (1874)[325] quantity indexes.<sup>\*67</sup> The *Laspeyres output quantity index* between periods 0 and 1 is defined as:

$$\begin{aligned} Q_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) &\equiv \sum_{m=1}^M p_m^0 y_m^1 / \sum_{m=1}^M p_m^0 y_m^0 \\ &= \sum_{m=1}^M (y_m^1/y_m^0) p_m^0 y_m^0 / \sum_{m=1}^M p_m^0 y_m^0 \\ &= \sum_{m=1}^M (y_m^1/y_m^0) s_m^0 \end{aligned} \quad (7.28)$$

<sup>\*66</sup> Harris (1999; 9)[216] raises an important point in the context of the discussion of whether government debt or taxes should be reduced in a surplus situation: "In balancing these concerns however one needs to factor in the impact of tax cuts on economic growth and output. Even ignoring dynamic effects, given a marginal excess burden of 30 cents on each dollar of revenue, at the margin a *permanent* tax reduction today and forever of \$1 will raise real output *permanently* by 30 cents. In the presence of a fiscal surplus, the choice to reduce taxes will be growth enhancing, while debt reduction will not immediately increase the size of the economic pie. Debt reduction pushes the growth benefits into the future. ... The point is that with a current tax system which is generating a large fiscal surplus and a substantial MEB from current levels of taxation, an output maximizing strategy would be to reduce taxes rather than to reduce the debt." The main counter argument that one could make against this Harris critique is that the debt reduction strategy might be intergenerationally "fairer" since the current generation ran up the debt, a point that Harris (1999; 10)[216] recognizes.

<sup>\*67</sup> Actually, Laspeyres and Paasche originally defined the price counterparts to the quantity indexes that we are defining here; see (7.36) and (7.37) below.

where the *period t revenue share for output m* is defined as

$$s_m^t \equiv p_m^t y_m^t / \sum_{k=1}^M p_k^t y_k^t; \quad m = 1, \dots, M; t = 0, 1. \quad (7.29)$$

Thus the Laspeyres output quantity index is a base period revenue share weighted sum of the  $M$  individual quantity ratios,  $y_m^1/y_m^0$ .

The *Paasche output quantity index* between periods 0 and 1 is defined as:

$$\begin{aligned} Q_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) &\equiv \sum_{m=1}^M p_m^1 y_m^1 / \sum_{m=1}^M p_m^1 y_m^0 \\ &= \left[ \sum_{m=1}^M p_m^1 y_m^0 / \sum_{m=1}^M p_m^1 y_m^1 \right]^{-1} \\ &= \left[ \sum_{m=1}^M (y_m^1/y_m^0)^{-1} p_m^1 y_m^1 / \sum_{m=1}^M p_m^1 y_m^1 \right]^{-1} \\ &= \left[ \sum_{m=1}^M (y_m^1/y_m^0)^{-1} s_m^1 \right]^{-1}. \end{aligned} \quad (7.30)$$

Thus the Paasche output quantity index is a current period revenue share weighted harmonic mean of the  $M$  individual quantity ratios,  $y_m^1/y_m^0$ .

In what follows, we shall concentrate on the problems involved in choosing a functional form for the output index  $Q$ ; an analogous discussion applies to the choice of a functional form for the input index  $Q^*$ .

Another commonly used functional form for a quantity index is the Fisher (1922; 234)[187] ideal quantity index  $Q_F$  which is equal to the square root of the product of the Laspeyres and Paasche quantity index defined by (7.28) and (7.30); i.e.:

$$Q_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) \equiv [Q_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) Q_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)]^{1/2}. \quad (7.31)$$

Another commonly used functional form for a quantity index is the Törnqvist (1936)[373] quantity index  $Q_T$ . The natural logarithm of  $Q_T$  is defined to be the right hand side of (7.32) below:

$$\ln Q_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) \equiv (1/2) \sum_{m=1}^M (s_m^0 + s_m^1) \ln(y_m^1/y_m^0) \quad (7.32)$$

where the revenue shares  $s_m^t$  are defined by (7.29) above. Note that the quantities  $y_m^t$  must all be positive in order for  $Q_T$  to be well defined.

The quantity index  $Q_T$  is also known as the *translog quantity index* (e.g.. see Jorgenson and Nishimizu (1978)[261] who introduced this terminology) because Diewert (1976; 120)[82] related  $Q_T$  to a translog production function. The index is also known as the Divisia index since Jorgenson and Griliches (1967)[258] (1972)[259] used  $Q_T$  to provide a discrete time approximation to the continuous time Divisia index.\*<sup>68</sup>

The four quantity indexes  $Q_L, Q_P, Q_F$  and  $Q_T$ , defined by (7.28), (7.30), (7.31), and (7.32) respectively, all have a common property: if the number of outputs  $M$  equals one, then each of these quantity indexes reduces to the output ratio,  $y_1^1/y_1^0$ . Thus, it can be seen that the use of quantity indexes for outputs and inputs can be used to generalize our one output, one input measure of productivity change, TFPG(1) defined by (7.3), discussed in section 7.2 above. More formally, let us

\*<sup>68</sup> Unfortunately, there are many discrete time approximations to the Divisia index including the Paasche and Laspeyres quantity indexes; see Frisch (1936)[192].

define the direct quantity index measure of productivity growth TFPG(5) in the general multiple output, multiple input case as follows:

$$\text{TFPG}(5) \equiv Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) / Q^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1) \quad (7.33)$$

where  $Q$  is the output quantity index and  $Q^*$  is the input quantity index. If the number of outputs equals one and the number of inputs equals one, if  $Q$  equals one of  $Q_L, Q_P, Q_F$  or  $Q_T$ , and if  $Q^*$  equals one of  $Q_L^*, Q_P^*, Q_F^*$  or  $Q_T^*$ , then  $\text{TFPG}(5) = \text{TFPG}(1)$ . Thus, the approach to productivity measurement outlined in this section reduces to the approach outlined in section 7.2 if there is only one input and only one output.

In the general multiple output, multiple input case, we still have to address a problem: which functional forms for the output index  $Q$  and the input index  $Q^*$  should we choose? We shall return to this functional form problem shortly.

We turn now to an index number measure of productivity that generalizes the deflated revenues divided by deflated costs productivity measure TFPG(3) that was defined earlier by (7.7).

Denote period  $t$  revenue by  $R^t$  and period  $t$  cost by  $C^t$ . We have:

$$R^t \equiv \sum_{m=1}^M p_m^t y_m^t; \quad C^t \equiv \sum_{n=1}^N w_n^t x_n^t; \quad t = 0, 1. \quad (7.34)$$

The multiple output analogue to the output price ratio which occurred in formula (7.7) above is the *output price index*,  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$ , which is a function of  $4M$  variables, the output prices and quantities that pertain to the two periods under consideration. The multiple input analogue to the input price ratio which occurred in (7.7) above is the *input price index*,  $P^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)$ , which is a function of  $4N$  variables, the input prices and quantities that pertain to the two periods under consideration.

Using the output price index  $P$  as a deflator for the revenue ratio  $R^1/R^0$  between periods 0 and 1 and using the input price index  $P^*$  as a deflator for the cost ratio  $C^1/C^0$  between the two periods leads to the following definition of the productivity growth of the production unit going from period 0 to 1:

$$\text{TFPG}(6) \equiv [(R^1/R^0)/P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)] / [(C^1/C^0)/P^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)]. \quad (7.35)$$

Note that (7.35) is a generalization to multiple inputs and outputs of our earlier productivity change measure TFPG(3) defined by (7.7).

Suppose that the output quantity index  $Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$  which appeared in definition (7.33) matches up with the output price index  $P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$  which appears in definition (7.35) in the sense that the product of the price and quantity index equals the revenue ratio for the two periods under consideration so that we have:<sup>\*69</sup>

$$R^1/R^0 = P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1). \quad (7.36)$$

Suppose further that the input quantity index  $Q^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)$  which appeared in definition (7.33) matches up with the input price index  $P^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)$  which appears in definition (7.35) in the sense that the product of the price and quantity index equals the cost ratio for the two periods under consideration so that we have:

$$C^1/C^0 = P^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1) Q^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1). \quad (7.37)$$

Now substitute (7.36) and (7.37) into (7.35) and we find that:

$$\text{TFPG}(5) = \text{TFPG}(6). \quad (7.38)$$

<sup>\*69</sup> This is the product test; see (7.47) below.

Thus if the two pairs of price and quantity indexes satisfy the relations (7.36) and (7.37), we find that both of the productivity measures introduced in this section, TFPG(5) defined by (7.33) and TFPG(6) defined by (7.35), are equal to each other.

Recall that in section 7.2, we defined the period  $t$  markup,  $m^t$ , for the production unit by  $1 + m^t = R^t/C^t$  for  $t = 0, 1$ . Using these definitions of the markup in each period again, it can be seen that we can rewrite TFPG(6) as follows:

$$\begin{aligned} \text{TFPG}(6) &\equiv [(R^1/R^0)/P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)]/[(C^1/C^0)/P^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)] \\ &= [(R^1/R^0)/(C^1/C^0)][P^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)/P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)] \\ &= [(1 + m^1)/(1 + m^0)][P^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)/P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)] \\ &\equiv \text{TFPG}(7). \end{aligned} \tag{7.39}$$

The above definition says that TFPG(7) is equal to the margin growth rate times the input price index divided by the output price index. Note that TFPG(7) is an exact analogue to our earlier one output, one input TFP growth measure TFPG(4) defined by (7.12) in section 7.2. Equations (7.39) show that this “new” measure of TFP growth is equal to the previous measure TFPG(5), which was the ratio of the output quantity index to the input quantity index, and to TFPG(6), which was equal to the revenue growth rate deflated by the output price index divided by the cost growth rate deflated by the input price index.<sup>\*70</sup> Thus we have obtained multiple output, multiple input counterparts to the equalities:

$$\text{TFPG}(1) = \text{TFPG}(3) = \text{TFPG}(4) \tag{7.40}$$

which were obtained in section 7.2 above.

There remains the problem of choosing a functional form for the output price index  $P$  and the input price index  $P^*$ . The same four index number formulae that were used for quantity indexes, (7.28), (7.30), (7.31), and (7.32), can also be used for price indexes, except that the role of prices and quantities are interchanged. Thus, define the Laspeyres price index  $P_L$ , the Paasche price index  $P_P$ , the Fisher price index  $P_F$ , and the translog price index  $P_T$  by (7.41), (7.42), (7.43), and (7.44), respectively:

$$P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) \equiv Q_L(\mathbf{y}^0, \mathbf{y}^1, \mathbf{p}^0, \mathbf{p}^1); \tag{7.41}$$

$$P_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) \equiv Q_P(\mathbf{y}^0, \mathbf{y}^1, \mathbf{p}^0, \mathbf{p}^1); \tag{7.42}$$

$$P_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) \equiv Q_F(\mathbf{y}^0, \mathbf{y}^1, \mathbf{p}^0, \mathbf{p}^1); \tag{7.43}$$

$$P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) \equiv Q_T(\mathbf{y}^0, \mathbf{y}^1, \mathbf{p}^0, \mathbf{p}^1) \tag{7.44}$$

Thus, the price indexes are equal to the corresponding quantity indexes with the role of prices and quantities interchanged in the quantity indexes. The Laspeyres, Paasche, Fisher and Translog input price indexes,  $P_L^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)$ ,  $P_P^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)$ ,  $P_F^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)$  and  $P_T^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)$  respectively, may be defined in an analogous manner.

If  $M = 1$ , so that there is only one output, then it can be verified that the output price indexes defined by (7.41)–(7.44) all collapse down to the output price ratio,  $p_1^1/p_1^0$ . Similarly, if  $N = 1$ , so that there is only one input, then  $P_L^*$ ,  $P_P^*$ ,  $P_F^*$  and  $P_T^*$  all collapse down to the input price ratio,  $w_1^1/w_1^0$ . Thus, the use of the Laspeyres, Paasche, Fisher or translog price indexes in (7.35) or (7.39) leads to the following equalities in the  $M = 1, N = 1$ :

$$\text{TFPG}(6) = \text{TFPG}(7) = \text{TFPG}(1). \tag{7.45}$$

<sup>\*70</sup> We require that (7.36) and (7.37) hold in order to obtain these equalities.

Thus, our new definitions of productivity change defined by (7.33), (7.35) or (7.39) are generalizations to the case of many outputs and inputs of our earlier one output, one input measure of productivity change defined by (7.3).

Returning to the general case of many outputs and many inputs, it can be seen that different choices of the output price index  $P$  and the input price index  $P^*$  will generate different productivity change measures TFPG(6) defined by (7.35). Similarly, different choices of the output quantity index  $Q$  and the input quantity index  $Q^*$  will generate different productivity change measures TFPG(5) defined by (7.33).

However, the degree of arbitrariness in the formulae (7.33) and (7.35) is not quite as large as it might seem at first glance. It turns out that the two families of productivity measures are related, because the deflated revenue ratio which occurs in the numerator of the right-hand side of (7.35),  $(R^1/R^0)/P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$ , can be interpreted as an implicit quantity index of outputs, and the denominator in (7.35),  $(C^1/C^0)/P^*(\mathbf{w}^0, \mathbf{w}^1, \mathbf{x}^0, \mathbf{x}^1)$ , can be interpreted as an implicit quantity index of inputs.

To see the above point more clearly, let us determine what  $(R^1/R^0)/P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$  equals when we let  $P$  equal the four specific price indexes defined by (7.41)–(7.44).

**Problem 1** Calculate  $(R^1/R^0)/P_L(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$ ,  $(R^1/R^0)/P_P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$  and  $(R^1/R^0)/P_F(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$  and show that the resulting quantity indexes are equal to either  $Q_L, Q_P, Q_F$  or  $Q_T$ .

*Hint:* You will need to use equations (7.34).

It can be shown that  $(R^1/R^0)/P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)$  is *not* equal to the translog quantity index,  $Q_T$ . Hence we simply define the *implicit Törnqvist Theil or Translog quantity index*,  $Q_{IT}$ , as follows:

$$Q_{IT}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) \equiv (R^1/R^0)/P_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1). \quad (7.46)$$

The five quantity indexes,  $Q_L, Q_P, Q_F, Q_T$  and  $Q_{IT}$ , defined by (7.28), (7.30), (7.31), (7.32) and (7.46) are the five functional forms for quantity indexes that are used most frequently in applied economics. The question now arises: which of these five formulae should we use in the multiple output, multiple input definition of TFP growth, TFPG(5) defined above by (7.33)?

In chapter 6, we showed that from the perspective of the economic approach to index number theory,  $Q_F, Q_T$  and  $Q_{IT}$  were to be clearly preferred to the Paasche and Laspeyres quantity indexes,  $Q_P$  and  $Q_L$ . If we wanted to use TFPG(6) or TFPG(7) as our multiple output, multiple input productivity growth concept, then again using the results in chapter 6, we showed that from the perspective of the economic approach to index number theory,  $P_F, P_T$  and  $P_{IT}$  were to be clearly preferred to the Paasche and Laspeyres price indexes,  $P_P$  and  $P_L$ . The economic approach was equally valid for  $Q_F, Q_T$  and  $Q_{IT}$  or for  $P_F, P_T$  and  $P_{IT}$ . Hence, any of these indexes would be equally good from the economic perspective.\*<sup>71</sup>

Another major approach to index number theory is the *test or axiomatic approach* to index number theory. This approach to the determination of the functional form for  $P$  and  $Q$  works as follows: researchers suggest various mathematical properties that  $P$  or  $Q$  should satisfy based on a priori reasoning — these properties are called “tests” or “axioms” — and then mathematical reasoning is applied to determine: (i) whether the a priori tests are mutually consistent and (ii) whether the a priori tests uniquely determine the functional form for  $P$  or  $Q$ . The main contributors to the test or axiomatic approach were Walsh (1901)[389] (1921a)[391] (1921b)[392], Irving Fisher (1911)[185]

\*<sup>71</sup> Moreover, we showed in chapter 6, that for normal time series data, all of these indexes would give much the same answer.

(1922)[187], Frisch (1936)[192], Eichhorn (1978)[170], Eichhorn and Voeller (1976)[171] and Funke and Voeller (1978)[194] (1979)[195].<sup>\*72</sup>

We will not cover the test approach in great detail in this chapter but we will present some material on this important approach to index number theory.

One fundamental test that the price and quantity index should jointly satisfy is the following property:

$$P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1)Q(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) = R^1/R^0; \quad (7.47)$$

i.e., the product of the output price and quantity indexes between periods 0 and 1 should equal the revenue or value ratio between the two periods,  $R^1/R^0 = \sum_{m=1}^M p_m^1 y_m^1 / \sum_{m=1}^M p_m^0 y_m^0$ . This test was called the *product test* by Frisch (1930; 399)[191], but it was first formulated by Irving Fisher (1911; 388)[185].

If we accept the validity of the product test (and virtually all researchers do accept its validity), then  $P$  and  $Q$  cannot be determined independently. For example, if the functional form for the price index  $P$  is given, then (7.47) determines the functional form for the quantity index  $Q$ .

Thus, in what follows, we focus in on the determination of the functional form for the price index  $P$ . Once  $P$  has been determined,  $Q$  will be determined residually by (7.47).

We list a few examples of tests that have been proposed for price indexes.

The *Identity* or *Constant Prices Test*, originally proposed by Laspeyres (1871; 308)[285] and also by Walsh (1901; 308)[389], and Eichhorn and Voeller (1976; 24)[171] is the following test:

$$P(\mathbf{p}, \mathbf{p}, \mathbf{y}^0, \mathbf{y}^1) = 1; \quad (7.48)$$

i.e., if  $\mathbf{p}^0 = \mathbf{p}^1 \equiv \mathbf{p}$ , so that for each commodity, prices are equal in the two periods being compared, then the price index is equal to 1 no matter what the quantities are in period 0 and 1,  $\mathbf{y}^0$  and  $\mathbf{y}^1$  respectively.

The *Constant Basket Test* or the *Constant Quantities Test*, proposed by many researchers including Walsh (1901; 540)[389] is the following test:

$$P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}, \mathbf{y}) = \sum_{m=1}^M p_m^1 y_m / \sum_{m=1}^M p_m^0 y_m; \quad (7.49)$$

i.e., if quantities are constant over the two periods 0 and 1 so that  $\mathbf{y}^0 = \mathbf{y}^1 \equiv \mathbf{y}$ , then the level of prices in period 1 compared to period 0 is the value of the constant basket of quantities evaluated at the period 1 prices,  $\sum_{m=1}^M p_m^1 y_m$ , divided by the value of the basket evaluated at the period 0 prices,  $\sum_{m=1}^M p_m^0 y_m$ .

The *Proportionality in Period  $t$  Prices Test*, proposed by Walsh (1901; 385)[389] and Eichhorn and Voeller (1976; 24)[171], is the following test:

$$P(\mathbf{p}^0, \lambda \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) = \lambda P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) \quad \text{for all } \lambda > 0; \quad (7.50)$$

i.e., if each price in period 1 is multiplied by the positive constant  $\lambda$ , then the level of prices in period 1 relative to the level of prices in period 0 increases by the same positive constant  $\lambda$ .

Our final example of a price index test is the *Time Reversal Test*, which was first informally proposed by Pierson (1896; 128)[326] and more formally by Walsh (1901; 368)[389] (1921b; 541)[392] and Fisher (1922; 64)[187]:

$$P(\mathbf{p}^1, \mathbf{p}^0, \mathbf{y}^1, \mathbf{y}^0) = 1/P(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1); \quad (7.51)$$

<sup>\*72</sup> For more recent contributions and surveys, see Diewert (1992b)[105] (1993)[109] (2004)[124] and Balk (1995)[19].

i.e., if the prices and quantities for periods 0 and 1 are interchanged, then the resulting price index is the reciprocal of the original price index.

The five tests (7.47)–(7.51) will suffice to give the reader the flavour of the test approach to index number theory. For a much more extensive list of twenty or so tests, see Diewert (1992b)[105].

There are five leading functional forms for the output price index  $P$  that are most frequently used in empirical work: (i) the Laspeyres price index  $P_L$  defined by (7.41) above, (ii) the Paasche price index  $P_P$  defined by (7.42), (iii) the Fisher price index  $P_F$  defined by (7.43), (iv) the translog price index  $P_T$  defined by (7.44), and (v) the implicit translog price index  $P_{IT}$  defined by:

$$P_{IT}(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) \equiv \left[ \sum_{m=1}^M p_m^1 y_m^1 / \sum_{m=1}^M p_m^0 y_m^0 \right] / Q_T(\mathbf{p}^0, \mathbf{p}^1, \mathbf{y}^0, \mathbf{y}^1) \quad (7.52)$$

where the translog quantity index  $Q_T$  is defined by (7.32). Do these five functional forms for  $P$  satisfy the four tests (7.48) to (7.51)?

The answer is yes in the case of the Fisher ideal price index  $P_F$  and no for the other four price indexes:  $P_L$  fails (7.51),  $P_P$  fails (7.51),  $P_T$  fails (7.49), and  $P_{IT}$  fails (7.48).

**Problem 2** Show that the Fisher ideal price index  $P_F$  satisfies the tests (7.48)–(7.51).

**Problem 3** Show that  $P_L$  fails (7.51),  $P_P$  fails (7.51),  $P_T$  fails (7.49), and  $P_{IT}$  fails (7.48).

**Problem 4** Show that  $P_F$  and  $Q_F$  satisfy the Product Test (7.47).

When more extensive lists of tests are compiled, the Fisher ideal price index  $P_F$  continues to satisfy more tests than other leading candidates; see Diewert (1976; 131)[82] (1992b)[105]. In fact, the Fisher price index satisfies all 20 tests utilised by Diewert (1992b)[105].<sup>\*73</sup> Moreover, satisfactory axiomatic characterizations of  $P_F$  have been obtained recently; see Funke and Voeller (1978; 180)[194] (1979)[195] and Diewert (1992b)[105]. Thus, from the viewpoint of the test approach to index number theory, the Fisher price index  $P_F$  defined by (7.43) and the corresponding Fisher quantity index  $Q_F$  defined by (7.31) seem to be the best choices. It should also be noted that  $P_F$  and  $Q_F$  satisfy the Product Test (7.47). Hence, if the Fisher indexes are used in the productivity measures defined by (7.33) and (7.35), then both of these productivity measures will coincide; i.e., if we use Fisher price and quantity indexes for  $P$  and  $Q$  and  $P^*$  and  $Q^*$  wherever they occur in (7.33), (7.35) and (7.39), we obtain the following equality:

$$\text{TFPG}_F(5) = \text{TFPG}_F(6) = \text{TFPG}_F(7) \quad (7.53)$$

where we have added a subscript  $F$  to the three productivity measures to indicate that Fisher indexes are being used. *Thus, an added benefit of using Fisher price and quantity indexes is that three conceptually distinct (but equally attractive) productivity change measures become identical.*

From section 7.2, it is evident that the total factor productivity growth measures that were defined there measure the combined effects of technological (and managerial) progress and increasing (or decreasing) returns to scale. The TFP growth measures defined in the present section also measure the combined effects of these two factors. In the following section, we attempt to devise a framework that will allow us to identify the separate effects of technical change and returns to scale in the many output and many input case under some conditions.

<sup>\*73</sup> However recently, Diewert (2004)[124] has obtained axiomatic justifications for the Translog price and quantity indexes,  $P_T$  and  $Q_T$ , that are comparable to the axiomatic justifications that have been obtained for the Fisher ideal indexes,  $P_F$  and  $Q_F$ .

## 7.10 The Estimation of Technical Progress and Returns to Scale

As in the previous section, consider the case of a single firm or production unit that produces  $N$  outputs and uses  $M$  inputs for periods  $0, 1, \dots, T$ . Let  $\mathbf{y} \equiv [y_1, \dots, y_N]$  denote the vector of positive outputs that is produced by the positive vector of inputs,  $\mathbf{x} \equiv [x_1, \dots, x_M]$ . Assume that in period  $t$ , the firm has a feasible set of inputs and outputs,  $S^t$ , and that it faces a positive vector of input prices,  $\mathbf{w} \equiv [w_1, \dots, w_M]$ . Assuming that the firm takes these input prices as fixed and beyond its control, the firm's *period  $t$  joint cost function*,  $C(\mathbf{w}, \mathbf{y}, t)$ , conditional on target set of outputs  $\mathbf{y}$  that must be produced, is defined as follows:

$$C(\mathbf{w}, \mathbf{y}, t) \equiv \min_{\mathbf{x}} \{ \mathbf{w} \cdot \mathbf{x} : (\mathbf{y}, \mathbf{x}) \text{ belongs to } S^t \} \quad (7.54)$$

where  $\mathbf{w} \cdot \mathbf{x} \equiv \sum_{m=1}^M w_m x_m$  denotes the inner product between the vectors  $\mathbf{w}$  and  $\mathbf{x}$ . The joint cost function provides a characterization of the firm's technology.

A measure of the (reciprocal) *local returns to scale* of a multiple output, multiple input firm can be defined as the percentage change in cost due to a one percent increase in all outputs. The technical definition is<sup>\*74</sup>:

$$\begin{aligned} \rho(\mathbf{w}, \mathbf{y}) &\equiv [C(\mathbf{w}, \mathbf{y}, t)]^{-1} dC(\mathbf{w}, \lambda \mathbf{y}, t) / d\lambda |_{\lambda=1} \\ &= \sum_{n=1}^N C_n(\mathbf{w}, \mathbf{y}, t) y_n / C(\mathbf{w}, \mathbf{y}, t) \\ &= \sum_{n=1}^N \partial \ln C(\mathbf{w}, \mathbf{y}, t) / \partial \ln y_n. \end{aligned} \quad (7.55)$$

Thus this measure of (inverse) returns to scale is equal to the sum of the cost elasticities with respect to the  $N$  outputs.

Now assume that the logarithm of the firm's period  $t$  cost function is the following *non constant returns to scale translog joint cost function*:<sup>\*75</sup>

$$\begin{aligned} \ln C(\mathbf{w}, \mathbf{y}, t) &\equiv -\tau t + \alpha_0 + \sum_{m=1}^M \alpha_m \ln w_m + \sum_{n=1}^N \beta_n \ln y_n + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln y_i \ln y_j \\ &\quad + (1/2) \sum_{k=1}^M \sum_{m=1}^M \delta_{km} \ln w_k \ln w_m + \sum_{m=1}^M \sum_{n=1}^N \phi_{mn} \ln w_m \ln y_n \end{aligned} \quad (7.56)$$

<sup>\*74</sup> This is the reciprocal of the usual returns to scale measure. Hence there are local decreasing costs (and increasing returns to scale) if  $\rho(\mathbf{w}, \mathbf{y}) < 1$  and constant costs (and constant returns to scale) if  $\rho(\mathbf{w}, \mathbf{y}) = 1$ .

<sup>\*75</sup> The basic translog functional form was introduced by Christensen, Jorgenson and Lau (1971)[53]. This particular functional form was introduced by Diewert (1974; 139)[76] as a joint revenue function, but the parameter  $k$  on the right hand side of (7.57) was set equal to 1 and the technical progress term,  $-\tau t$  was missing. The translog joint cost function was first introduced by Burgess (1974)[43].

where the parameters on the right hand side of (7.56) satisfy the following restrictions:

$$\sum_{n=1}^N \beta_n \equiv k > 0; \quad (7.57)$$

$$\sum_{j=1}^N \gamma_{ij} = 0 \text{ for } i = 1, \dots, N; \quad (7.58)$$

$$\gamma_{ij} = \gamma_{ji} \text{ for all } 1 \leq i < j \leq N; \quad (7.59)$$

$$\sum_{m=1}^M \alpha_m = 1; \quad (7.60)$$

$$\sum_{m=1}^M \delta_{km} = 0 \text{ for } k = 1, \dots, M; \quad (7.61)$$

$$\delta_{km} = \delta_{mk} \text{ for all } 1 \leq k < m \leq M; \quad (7.62)$$

$$\sum_{m=1}^M \phi_{mn} = 0 \text{ for } n = 1, \dots, N; \quad (7.63)$$

$$\sum_{n=1}^N \phi_{mn} = 0 \text{ for } m = 1, \dots, M. \quad (7.64)$$

The parameter  $\tau$  which occurs in the right hand side of (7.56) is a measure of *technical progress*, which in this case is expressed as exogenous cost reduction. Usually,  $\tau \geq 0$ ; if  $\tau < 0$ , then the technology exhibits *technological regress*.

**Problem 5** Show that the degree of reciprocal local returns to scale,  $\rho(\mathbf{w}, \mathbf{y})$ , using the  $C(\mathbf{w}, \mathbf{y}, t)$  defined by (7.56)-(7.64) is equal to:

$$\rho(\mathbf{p}, \mathbf{x}) = \sum_{n=1}^N \beta_n \equiv k. \quad (7.65)$$

*Hint:* Use (7.57), (7.58), (7.59) and (7.64).

If there are increasing returns to scale or decreasing costs so that the parameter  $k$  is less than one, then it is well known that competitive profit maximization breaks down in this case. Hence, since we do not want to restrict  $k$  to be equal or greater than one, it is necessary to allow for a monopolistic profit maximization problem in each period. Thus for period  $t$ , we assume that the firm or production unit faces the inverse demand function  $P_n^t(y_n)$  which gives the market clearing price for output  $n$  as a function of the amount of output  $y_n$  that the firm places on the market, for  $n = 1, \dots, N$ . Assuming that the firm faces the positive input price vector  $\mathbf{w}^t \equiv [w_1^t, \dots, w_M^t]$ , the *firm's period  $t$  monopolistic profit maximization problem* is the following unconstrained maximization problem involving the vector of period  $t$  output supplies  $\mathbf{y} \equiv [y_1, \dots, y_N]$ :

$$\max_{\mathbf{y}} \left\{ \sum_{n=1}^N P_n^t(y_n) y_n - C(\mathbf{w}^t, \mathbf{y}, t) \right\}. \quad (7.66)$$

The observed period  $t$  price for output  $n$  will be:

$$p_n^t \equiv P_n^t(y_n^t); \quad n = 1, \dots, N; t = 0, 1, \dots, T. \quad (7.67)$$

Assuming that the demand derivatives  $dP_n^t(y_n^t)/dy_n$  are nonpositive, the nonnegative *ad valorem monopolistic markup*  $m_n^t$  for the  $n$ th output in period  $t$  can be defined as follows:

$$m_n^t \equiv -[dP_n^t(y_n^t)/dy_n][y_n^t/p_n^t] \geq 0; \quad n = 1, \dots, N; t = 0, 1, \dots, T. \quad (7.68)$$

**Problem 6** Using definitions (7.67) and (7.68), *show* that the first order conditions for maximizing (7.66) can be written as follows:

$$p_n^t[1 - m_n^t] = \partial C(\mathbf{w}^t, \mathbf{y}^t, t)/\partial y_n; \quad n = 1, \dots, N; t = 0, 1, \dots, T. \quad (7.69)$$

In what follows, it will simplify the notation somewhat if we define one minus the markup for commodity  $n$  as the *markup factor* for output  $n$  in period  $t$ ,  $M_n^t$ :<sup>\*76</sup>

$$0 < M_n^t \equiv 1 - m_n^t \leq 1; \quad n = 1, \dots, N; t = 0, 1, \dots, T. \quad (7.70)$$

Using definitions (7.70), conditions (7.69) become:

$$p_n^t M_n^t = \partial C(\mathbf{w}^t, \mathbf{y}^t, t) / \partial y_n; \quad n = 1, \dots, N; t = 0, 1, \dots, T. \quad (7.71)$$

Assuming differentiability of the period  $t$  cost function with respect to the input prices, using Shephard's (1953; 11)[355] Lemma, the cost minimizing vector of input demands for the firm in period  $t$ ,  $\mathbf{x}^t \equiv [x_1^t, \dots, x_M^t]$ , will be equal to the vector of first order partial derivatives of the cost function with respect to the components of the input price vector:

$$\mathbf{x}^t \equiv \nabla_{\mathbf{w}} C(\mathbf{w}^t, \mathbf{y}^t, t); \quad t = 0, 1, \dots, T \quad (7.72)$$

and the period  $t$  observed total cost,  $C(\mathbf{w}^t, \mathbf{y}^t, t)$ , will be equal to the inner product of the period  $t$  input price and quantity vectors,  $\mathbf{w}^t$  and  $\mathbf{x}^t$  respectively:

$$C(\mathbf{w}^t, \mathbf{y}^t, t) = \mathbf{w}^t \cdot \mathbf{x}^t; \quad t = 0, 1, \dots, T. \quad (7.73)$$

**Problem 7** (a) *Show* that the following equations hold, using (7.71) to (7.73) and the specific translog functional form defined by (7.56): for  $n = 1, \dots, N; t = 0, 1, \dots, T$ :

$$[p_n^t y_n^t M_n^t] / \mathbf{w}^t \cdot \mathbf{x}^t = \beta_n + \sum_{j=1}^N \gamma_{nj} \ln y_j + \sum_{m=1}^M \phi_{mn} \ln w_m = \partial \ln C(\mathbf{w}^t, \mathbf{y}^t, t) / \partial \ln y_n. \quad (7.74)$$

(b) *Show* that equations (7.74), definition (7.65) and some of the restrictions (7.57)-(7.64) imply the following equations:

$$\begin{aligned} \sum_{n=1}^N [p_n^t y_n^t M_n^t] / \mathbf{w}^t \cdot \mathbf{x}^t &= \sum_{n=1}^N \left[ \beta_n + \sum_{j=1}^N \gamma_{nj} \ln y_j + \sum_{m=1}^M \phi_{mn} \ln w_m \right]; \quad t = 0, 1, \dots, T \\ &= k. \end{aligned} \quad (7.75)$$

Note that equations (7.75) can be rearranged to yield the following expressions for period  $t$  costs:

$$\mathbf{w}^t \cdot \mathbf{x}^t = k^{-1} \sum_{n=1}^N p_n^t y_n^t M_n^t; \quad t = 0, 1, \dots, T. \quad (7.76)$$

Thus for each period  $t$ , an estimate of the firm's (reciprocal) returns to scale  $k$  can be obtained as the ratio of period  $t$  markup adjusted revenues,  $\sum_{n=1}^N p_n^t y_n^t M_n^t$ , divided by period  $t$  total cost,  $\mathbf{w}^t \cdot \mathbf{x}^t = \sum_{m=1}^M w_m^t x_m^t$ .<sup>\*77</sup>

<sup>\*76</sup> If there are constant or increasing costs so that the parameter  $k \geq 1$ , then this situation is consistent with the competitive pricing of outputs. To model this case in what follows, simply set each  $M_n^t = 1$  and estimate the parameters  $k$  and  $\tau$ . In the production function literature on returns to scale and markups where there is only a single output, the markup factor is defined as price over marginal cost, which is the reciprocal of the markup factor  $M_n^t$  which appears in (7.18); see Hall (1988)[209] (1990)[210] and Basu and Fernald (1997; 253)[23] (2002; 975)[24] for these single output production function approaches.

<sup>\*77</sup> If there is only one output so that  $N = 1$ , then (7.76) can be rewritten as  $k^{-1} = [M_1^t]^{-1} [\mathbf{w}^t \cdot \mathbf{x}^t / p_1^t y_1^t]$ , which is a standard result in the one output production function literature on this topic: see Basu and Fernald (1997; 253)[23] (2002; 976)[24]. The term  $\mathbf{w}^t \cdot \mathbf{x}^t / p_1^t y_1^t$  is observed cost over observed revenue, which in turn is one minus the revenue share of pure profits.

Rearranging the second equality in (7.74) leads to the following system of equations:

$$\partial \ln C(\mathbf{w}^t, \mathbf{y}^t, t) / \partial \ln y_n = p_n^t y_n^t M_n^t / \mathbf{w}^t \cdot \mathbf{x}^t; \quad n = 1, \dots, N; t = 0, 1, \dots, T \quad (7.77)$$

$$= k p_n^t y_n^t M_n^t / \sum_{j=1}^N p_j^t y_j^t M_j^t \quad \text{using (7.76)}. \quad (7.78)$$

Now assume that the markup factors within each period are constant across commodities; i.e., assume:

$$M_n^t = M^t; \quad n = 1, \dots, N; t = 0, 1, \dots, T. \quad (7.79)$$

**Problem 8** Use assumptions (7.79) and the previous material to *show* that the following equations hold:

$$\begin{aligned} \partial \ln C(\mathbf{w}^t, \mathbf{y}^t, t) / \partial \ln y_n &= k p_n^t y_n^t / \sum_{j=1}^N p_j^t y_j^t & n = 1, \dots, N; t = 0, 1, \dots, T \\ &= k s_n^t \end{aligned} \quad (7.80)$$

where  $s_n^t \equiv p_n^t y_n^t / \mathbf{p}^t \cdot \mathbf{y}^t$  is the *observed revenue share* of output  $n$  in period  $t$ .

**Problem 9** Using (7.72) and (7.73), *show* that the logarithmic derivatives of the period  $t$  cost function with respect to input prices are equal to:

$$\begin{aligned} \partial \ln C(\mathbf{w}^t, \mathbf{y}^t, t) / \partial \ln w_m &= w_m^t x_m^t / \mathbf{w}^t \cdot \mathbf{x}^t & m = 1, \dots, M; t = 0, 1, \dots, T \\ &= S_m^t \end{aligned} \quad (7.81)$$

where  $S_m^t \equiv w_m^t x_m^t / \mathbf{w}^t \cdot \mathbf{x}^t$  is the *observed cost share* of input  $m$  in period  $t$ .

**Problem 10** Since the right hand side of (7.56) is a quadratic function in the logarithms of output quantities, the logarithms of input prices and time, *show* that Diewert's (1976; 118)[82] Quadratic Identity and the material in problems 5-9 leads to the following equations, relating the difference in the costs in periods  $t-1$  and  $t$ ,  $\mathbf{w}^{t-1} \cdot \mathbf{x}^{t-1} = C(\mathbf{w}^{t-1}, \mathbf{y}^{t-1}, t-1)$  and  $\mathbf{w}^t \cdot \mathbf{x}^t = C(\mathbf{w}^t, \mathbf{y}^t, t)$ :

$$\begin{aligned} &\ln C(\mathbf{w}^t, \mathbf{y}^t, t) - \ln C(\mathbf{w}^{t-1}, \mathbf{y}^{t-1}, t-1) & t = 1, 2, \dots, T \\ &= (1/2) \{ [\partial \ln C(\mathbf{w}^{t-1}, \mathbf{y}^{t-1}, t-1) / \partial t] + [\partial \ln C(\mathbf{w}^t, \mathbf{y}^t, t) / \partial t] \} [(t) - (t-1)] \\ &+ (1/2) \sum_{n=1}^N \{ [\partial \ln C(\mathbf{w}^{t-1}, \mathbf{y}^{t-1}, t-1) / \partial \ln y_n] + [\partial \ln C(\mathbf{w}^t, \mathbf{y}^t, t) / \partial \ln y_n] \} [\ln y_n^t - \ln y_n^{t-1}] \\ &+ (1/2) \sum_{m=1}^M \{ [\partial \ln C(\mathbf{w}^{t-1}, \mathbf{y}^{t-1}, t-1) / \partial \ln w_m] + [\partial \ln C(\mathbf{w}^t, \mathbf{y}^t, t) / \partial \ln w_m] \} [\ln w_m^t - \ln w_m^{t-1}] \\ &= -\tau + k \ln Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t) + \ln P_T(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t) \end{aligned} \quad (7.82)$$

where  $Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t)$  is the *Törnqvist* (1936)[373] (1937)[374] *quantity index* for output growth between periods  $t-1$  and  $t$  and  $P_T(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t)$  is the *Törnqvist input price index* for input price growth between periods  $t-1$  and  $t$ . As we know from the previous section, the logarithms of these two indexes are defined as follows:

$$\ln Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t) \equiv (1/2) \sum_{n=1}^N [s_n^{t-1} + s_n^t] [\ln y_n^t - \ln y_n^{t-1}]; \quad (7.83)$$

$$\ln P_T(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t) \equiv (1/2) \sum_{m=1}^M [S_m^{t-1} + S_m^t] [\ln w_m^t - \ln w_m^{t-1}]. \quad (7.84)$$

**Problem 11** The Törnqvist input price index between periods  $t-1$  and  $t$ ,  $P_T(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t)$ , can be used in order to define the *implicit Törnqvist input quantity index* between periods  $t-1$  and  $t$  as follows:

$$Q_T^*(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t) \equiv \mathbf{w}^t \cdot \mathbf{x}^t / \{ \mathbf{w}^{t-1} \cdot \mathbf{x}^{t-1} P_T(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t) \}. \quad (7.85)$$

Use the above definitions to *show that* equations (7.82) can be rewritten as follows:

$$\ln Q_T^*(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t) = -\tau + k \ln Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t); \quad t = 1, 2, \dots, T. \quad (7.86)$$

Thus if  $T \geq 2$ , then the technical change parameter  $\tau$  and the returns to scale parameter  $k$  can be estimated by running a linear regression using equations (7.86) after appending error terms.\*<sup>78</sup> If there is positive technical progress, then  $\tau > 0$  while if there are increasing returns to scale, then  $k < 1$ . Hence, a combination of technical progress and increasing returns to scale will cause input growth to be less than output growth. *Equations (7.86) enable us to assess the contribution of returns to scale versus technical progress (which is a shift in the production or cost function) in a very simple regression model that has eliminated all of the nuisance parameters that are in the translog cost function that was defined earlier by (7.56).* This is a rather remarkable result which is valid even if  $M$  and  $N$  are extremely large so that traditional econometric methods for estimating  $\tau$  and  $k$  fail.\*<sup>79</sup>

Now suppose that returns to scale are 1 so that the parameter  $k$  in (7.86) equals 1. Then recalling the results in the previous section, a traditional index number measure of Total Factor Productivity Growth can be defined as follows:

$$\gamma \equiv Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t) / Q_T^*(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t). \quad (7.87)$$

Now if  $k = 1$ , we can rewrite (7.86) as follows:

$$\ln[Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t) / Q_T^*(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t)] = \tau; \quad t = 1, 2, \dots, T. \quad (7.88)$$

Exponentiating both sides of (7.88) gives us the following relationships:

$$\gamma \equiv Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t) / Q_T^*(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t) = e^\tau; \quad t = 1, 2, \dots, T. \quad (7.89)$$

Hence, if there are constant returns to scale so that  $k = 1$ , then the productivity growth measure  $\gamma$  that was defined in the previous section is equal to the technical progress measure  $e^\tau$ .

In the general case where  $k$  is not necessarily equal to 1, we can rearrange (7.86) in order to obtain the following relationship between the translog productivity growth measure  $\gamma$  defined as  $Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t) / Q_T^*(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t)$  and the parameters  $k$  and  $\tau$ :

$$\gamma \equiv Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t) / Q_T^*(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t) = e^\tau Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t)^{1-k}; \quad t = 1, 2, \dots, T. \quad (7.90)$$

If there is positive output growth so that  $Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t) > 1$  and if there are increasing returns to scale so that  $k < 1$ , then it can be seen that  $Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t)^{1-k} > 1$  and the productivity growth measure  $\gamma$  is equal to the product of a technical progress term  $e^\tau$  (which is greater than 1 if  $\tau$  is greater than 0) times the term  $Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t)^{1-k}$ , which reflects the degree of returns to scale. Thus the decomposition (7.90) provides an economic justification for our earlier assertion that the measures of TFP growth defined in the previous section reflect both the effects of technical progress and returns to scale.

We now consider some of the more subtle problems involved in estimating the two parameters  $k$  and  $\tau$  in the linear regression equation (7.86) above. We start off by reviewing some material on simple linear regression models.

\*<sup>78</sup> Recall that we required the constant across commodities markup assumption (7.79) in order to derive this result. Of course, we also require the rate of cost reducing technical progress parameter  $\tau$  to be constant over the sample period in order to apply the linear regression.

\*<sup>79</sup> This general technique was introduced to the economics literature by Nakajima, Nakamura and Yoshioka (1998)[318] and Nakajima, Nakamura and Nakamura (2002)[317]. The specific results derived in this problem section are due to Diewert and Fox (2004)[138].

Let  $\mathbf{Y}$  and  $\mathbf{X}$  be  $N$  dimensional vectors and consider the linear regression model:

$$\mathbf{Y} = \mathbf{1}_N \alpha + \mathbf{X} \beta + \boldsymbol{\varepsilon} \quad (7.91)$$

where  $\mathbf{1}_N$  is a vector of ones of dimension  $N$ ,  $\alpha$  and  $\beta$  are scalar parameters and  $\boldsymbol{\varepsilon}$  is an  $N$  dimensional vector of error terms. It is well known that the vector of least squares estimators for  $\alpha$  and  $\beta$  is given by:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \{[\mathbf{1}_N, \mathbf{X}]^T [\mathbf{1}_N, \mathbf{X}]\}^{-1} [\mathbf{1}_N, \mathbf{X}]^T \mathbf{Y}. \quad (7.92)$$

Define the vectors of deviations from the mean for  $\mathbf{X}$  and  $\mathbf{Y}$  as follows:

$$\mathbf{x} \equiv \mathbf{X} - \mathbf{1}_N \mathbf{X}^*; \quad (7.93)$$

$$\mathbf{y} \equiv \mathbf{Y} - \mathbf{1}_N \mathbf{Y}^* \quad (7.94)$$

where  $\mathbf{X}^* \equiv \mathbf{1}_N^T \mathbf{X} / N$  and  $\mathbf{Y}^* \equiv \mathbf{1}_N^T \mathbf{Y} / N$  are the arithmetic means of the components of  $\mathbf{X}$  and  $\mathbf{Y}$  respectively.

**Problem 12** Show that  $b$ , the least squares estimator for  $\beta$ , can be written as follows:

$$b = (\mathbf{x}^T \mathbf{y}) / (\mathbf{x}^T \mathbf{x}). \quad (7.95)$$

**Problem 13** Recall problem 11 above where we suggested running a linear regression of the form:

$$X^t = \alpha + \beta Y^t + \varepsilon^t; \quad t = 1, \dots, T \quad (7.96)$$

in order to determine the reciprocal returns to scale parameter  $k = \beta$ , where  $X^t \equiv \ln Q_T^*(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t)$  is the log of input growth and  $Y^t \equiv \ln Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t)$  is the log of output growth. Hence in (7.96), output growth is regarded as the exogenous variable while input growth is regarded as the endogenous variable. In the production literature, input growth is usually regarded as the exogenous variable and output growth as being endogenous. If we take this traditional view, then (7.96) should be rewritten as follows:

$$Y^t = \gamma + \delta X^t + \eta^t; \quad t = 1, \dots, T \quad (7.97)$$

where the new parameters  $\gamma$  and  $\delta$  are related to the old  $\alpha$  and  $\beta$  as follows:

$$\gamma \equiv -\alpha / \beta; \quad (7.98)$$

$$\delta \equiv 1 / \beta. \quad (7.99)$$

Let  $\mathbf{X}$  and  $\mathbf{Y}$  be  $T$  dimensional vectors of the  $X^t$  and the  $Y^t$ . Assume that the variance of  $\mathbf{X}$  and  $\mathbf{Y}$  are positive (so that  $\mathbf{x}^T \mathbf{x} > 0$  and  $\mathbf{y}^T \mathbf{y} > 0$  using the notation in problem 12) and that the covariance between  $\mathbf{X}$  and  $\mathbf{Y}$  is also positive (so that  $\mathbf{x}^T \mathbf{y} > 0$ ). This last assumption is justified in our context since output growth and input growth will be positively correlated. Let  $d$  be the least squares estimator for  $\delta$  in equation (7.97) so that this is a direct measure of returns to scale (rather than being a reciprocal measure) and let  $b$  be the least squares estimator for  $\beta$  in equation (7.96). Hence  $1/b$  is also a direct measure of returns to scale.

(a) Show that under our assumptions

$$d \leq 1/b. \quad (7.100)$$

Hint: This is actually a straightforward application of problem 12 and the Cauchy Schwarz inequality.<sup>\*80</sup> This problem is of some practical importance, since it tells us if we run our initial regression

<sup>\*80</sup> This result was obtained recently by Bartelsman (1995)[22] but it is implicit in Cramér (1946; 273-275)[60].

(7.96), we will generally obtain a higher estimate of returns to scale than if we run the alternative regression (7.97)!

(b) What do you think that we should do in practice? Should we combine the two estimates of returns to scale or should we pick one or the other of our two possible estimates? Or is there some other estimation technique that we could use that might be more symmetric?

The result (7.100) obtained in the previous problem has some rather disturbing implications for other areas of applied economics. For example, let  $\mathbf{Y}$  be a quantity vector and let  $\mathbf{X}$  be the corresponding vector of prices and let  $\mathbf{y}$  and  $\mathbf{x}$  be the corresponding centered vectors. Then the regression (7.97) is a direct regression of quantity on price and generates a direct estimate,  $d$ , of the effects on quantity supplied or demanded of a change in price. The regression (7.96) generates an indirect estimate of the same effect,  $1/b$ . In the case where we are estimating an input demand function or a consumer demand function, it will usually be the case that  $\mathbf{x}^T \mathbf{y} < 0$  and in this case, the least squares estimators for  $\delta$  and  $\beta$ ,  $d$  and  $b$ , will be negative and the following inequalities will hold:

$$0 > d \equiv (\mathbf{x}^T \mathbf{y})/(\mathbf{x}^T \mathbf{x}) \geq (\mathbf{y}^T \mathbf{y})/(\mathbf{x}^T \mathbf{y}) \equiv 1/b \quad (7.101)$$

or since  $d$  and  $b$  are negative:

$$|d| \leq 1/|b|. \quad (7.102)$$

Thus elasticities of demand estimated by regressing quantity on price using (7.97) will tend to be smaller in magnitude than the corresponding elasticities of demand estimated by regressing price on quantity using (7.96).

In the case where we are estimating an output supply function,  $\mathbf{x}^T \mathbf{y}$  will tend to be positive and hence  $d$  and  $1/b$  will be positive, and in this case, the inequality (7.100) will hold; i.e., elasticities of supply estimated by regressing quantity on price using (7.97) will tend to be smaller in magnitude than the corresponding elasticities of supply estimated by regressing price on quantity using (7.96). Thus own price elasticities estimated *directly* by regressing quantities on prices will generally be *less in magnitude* than when estimated *indirectly* by regressing prices on quantities. This is a very troublesome result since the direct and indirect estimates can be very different.

A possible way out of the above difficulties might be to develop a symmetric regression model.<sup>\*81</sup> Thus let the  $N$  dimensional vectors  $\mathbf{x}$  and  $\mathbf{y}$  be given. We can interpret them as zero mean vectors like those defined by equations (7.93) and (7.94) above. We are interested in fitting linear regressions through the origin for these variables of the type  $\mathbf{y} = a\mathbf{x} + \mathbf{e}$  or  $\mathbf{x} = b\mathbf{y} + \mathbf{e}$  but we would like a procedure that would have the property that our estimator for  $a$  is equal to the reciprocal of the estimator for  $b$  (so that it would not matter which way we ran the regression). Consider the following method for fitting a regression of the type  $\mathbf{y} = a\mathbf{x} + \mathbf{e}$ :

$$\begin{aligned} & \min_{\mathbf{x}^*, \mathbf{y}^*, a} \{(\mathbf{y} - \mathbf{y}^*)^T (\mathbf{y} - \mathbf{y}^*) + (\mathbf{x} - \mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*) : \mathbf{y}^* = a\mathbf{x}^*\} \\ & = \min_{\mathbf{x}^*, a} \{(\mathbf{y} - a\mathbf{x}^*)^T (\mathbf{y} - a\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*)\} \equiv f(\mathbf{x}^*, a). \end{aligned} \quad (7.103)$$

Solving the problem (7.103) minimizes the sum of the squared distances of each  $(y_n, x_n)$  observation from the line through the origin that has the equation  $\mathbf{y} = a\mathbf{x}$ . Let us first minimize  $f(\mathbf{x}^*, a)$  with respect to the components of the  $\mathbf{x}^*$  vector conditional on a given scalar parameter,  $a$ .

**Problem 14** (a) *Show* that solving

$$\nabla_{\mathbf{x}^*} f(\mathbf{x}^*, a) = \mathbf{0}_N \quad (7.104)$$

<sup>\*81</sup> The model that we are about to describe dates back to Adcock (1878)[1] but it has been rediscovered many times since; see Madanski (1959; 202)[297] for references to the literature. Golub and Van Loan (1980)[200] renamed the method as “total least squares”.

leads to the following  $\mathbf{x}^*$  solution:

$$\mathbf{x}^{**} = [1 + a^2]^{-1}[\mathbf{x} + a\mathbf{y}]. \quad (7.105)$$

(b) Check that the solution given by (7.105) satisfies the second order conditions for minimizing  $f(\mathbf{x}^*, a)$  with respect to  $\mathbf{x}^*$ .

(c) Substitute the solution (7.105) into  $f(\mathbf{x}^*, a)$  defined in (7.103) and *show* that the resulting expression simplifies to

$$g(a) \equiv [1 + a^2]^{-1}[\mathbf{y} - a\mathbf{x}]^T[\mathbf{y} - a\mathbf{x}]. \quad (7.106)$$

(d) *Show* that the first order necessary condition for minimizing  $g(a)$  with respect to  $a$  is equivalent to finding a root of the following quadratic equation (we rule out infinite solutions to the first order conditions):

$$\mathbf{x}^T \mathbf{y} a^2 + [\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y}] a - \mathbf{x}^T \mathbf{y} = 0. \quad (7.107)$$

Assume that

$$\mathbf{x}^T \mathbf{y} > 0 \text{ so that the } \mathbf{x} \text{ and } \mathbf{y} \text{ vectors are positively correlated.} \quad (7.108)$$

(e) Under these conditions, what can you say about the signs of the two “ $a$ ” roots for (7.107)?

(f) *Show* that the largest root of (7.107) is given by

$$a^* = \{[\mathbf{y}^T \mathbf{y} - \mathbf{x}^T \mathbf{x}] + [(\mathbf{y}^T \mathbf{y} - \mathbf{x}^T \mathbf{x})^2 + 4(\mathbf{x}^T \mathbf{y})^2]^{1/2}\} / 2\mathbf{x}^T \mathbf{y}. \quad (7.109)$$

This root is the desired estimator for the parameter  $a$  in the regression line  $\mathbf{y} = a\mathbf{x}$ .

Now consider the following method for fitting a regression of the type  $\mathbf{x} = b\mathbf{y} + \mathbf{e}$ :

$$\min_{\mathbf{x}^*, \mathbf{y}^*, b} \{(\mathbf{y} - \mathbf{y}^*)^T(\mathbf{y} - \mathbf{y}^*) + (\mathbf{x} - \mathbf{x}^*)^T(\mathbf{x} - \mathbf{x}^*) : \mathbf{x}^* = b\mathbf{y}^*\}. \quad (7.110)$$

It can be seen that (7.110) is the same as (7.103) except the roles of  $\mathbf{x}$  and  $\mathbf{y}$  have been reversed. Hence we can simply note that the  $b$  solution to (7.110) is given by (7.109) except the roles of  $\mathbf{x}$  and  $\mathbf{y}$  must be reversed so that

$$b^* = \{[\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y}] + [(\mathbf{x}^T \mathbf{x} - \mathbf{y}^T \mathbf{y})^2 + 4(\mathbf{y}^T \mathbf{x})^2]^{1/2}\} / 2\mathbf{y}^T \mathbf{x}. \quad (7.111)$$

(g) *Show* that (again assuming that  $\mathbf{x}^T \mathbf{y} > 0$ ):

$$a^* b^* = 1. \quad (7.112)$$

Hence the regression methods defined by solving (7.103) or (7.110) are indeed symmetric.

**Problem 15** Let  $a^*$  be the “ $a$ ” solution to (7.103). Now suppose we change the units of measurement for the  $\mathbf{x}$  variable by multiplying  $\mathbf{x}$  and  $\mathbf{x}^*$  by the positive scalar  $\lambda$ . The new symmetric regression problem can be written as follows:

$$\begin{aligned} & \min_{\mathbf{x}^*, \mathbf{y}^*, a} \{(\mathbf{y} - \mathbf{y}^*)^T(\mathbf{y} - \mathbf{y}^*) + (\lambda\mathbf{x} - \mathbf{x}^*)^T(\lambda\mathbf{x} - \mathbf{x}^*) : \mathbf{y}^* = a\mathbf{x}^*\} \\ & = \min_{\mathbf{x}^*, a} \{(\mathbf{y} - a\mathbf{x}^*)^T(\mathbf{y} - a\mathbf{x}^*) + (\lambda\mathbf{x} - \mathbf{x}^*)^T(\lambda\mathbf{x} - \mathbf{x}^*)\} \equiv f(\mathbf{x}^*, a). \end{aligned} \quad (7.113)$$

Let  $a^{**}$  solve (7.113). *Show that* in general  $a^{**} \neq a^*/\lambda$ . Thus the estimator for “ $a$ ” that the symmetric regression generates is in general *not invariant to changes in the units of measurement of the variables in the  $\mathbf{x}$  and  $\mathbf{y}$  vectors.*<sup>\*82</sup>

<sup>\*82</sup> Allen (1939; 199)[5] pointed out this problem with the “orthogonal regression line” or the “line of best fit” and recommended that it should not be used because of this problem. We agree with his recommendation although in the present context, changing the units of measurement for outputs and inputs will not change our  $\mathbf{Y}$  and  $\mathbf{X}$  variables since they are rates of change.

The previous problem shows that the symmetric regression is not the answer to our problems in determining a symmetric approach to the bivariate regression problem. However, we leave this problem for now and consider a generalization of our basic theoretical regression model (7.86) to the case where we have some dummy variables involving time.

Define the function of one variable  $f(t)$  as follows:

$$f(t) \equiv \begin{cases} -\tau_0 t & \text{for } t \leq t^* \\ -\tau_0 t^* - \tau_1 [t - t^*] & \text{for } t \geq t^* \end{cases} \quad (7.114)$$

where  $\tau_0$  and  $\tau_1$  are fixed parameters. If  $t > t^*$ , then it can be shown by direct computation that

$$f(t) - f(t^*) = -\tau_1 [t - t^*]. \quad (7.115)$$

Recall Problem 11 but now redefine the old joint cost function (7.56) as follows:

$$\begin{aligned} \ln C(\mathbf{w}, \mathbf{y}, t) &\equiv f(t) + \alpha_0 + \sum_{m=1}^M \alpha_m \ln w_m + \sum_{n=1}^N \beta_n \ln y_n \\ &+ (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln y_i \ln y_j + (1/2) \sum_{k=1}^M \sum_{m=1}^M \delta_{km} \ln w_k \ln w_m \\ &+ \sum_{m=1}^M \sum_{n=1}^N \phi_{mn} \ln w_m \ln y_n \end{aligned} \quad (7.116)$$

where  $f(t)$  is the linear spline function  $f(t)$  defined by (7.114) and the parameters on the right hand side of (7.116) satisfy the restrictions (7.57)-(7.64) above. Note that  $\ln C(\mathbf{w}, \mathbf{y}, t)$  is the sum of the linear spline function  $f(t)$  and a function that is quadratic in the variables  $\ln w_m$  and  $\ln y_n$ . Hence we can apply Diewert's Quadratic Identity and (7.115) above to show that if  $t > t^*$ , then

$$\begin{aligned} &\ln C(\mathbf{w}^t, \mathbf{y}^t, t) - \ln C(\mathbf{w}^{t^*}, \mathbf{y}^{t^*}, t^*) \quad t > t^* \\ &= -\tau_1 [t - t^*] \\ &+ (1/2) \sum_{n=1}^N \{[\partial \ln C(\mathbf{w}^{t^*}, \mathbf{y}^{t^*}, t^*) / \partial \ln y_n] + [\partial \ln C(\mathbf{w}^t, \mathbf{y}^t, t) / \partial \ln y_n]\} [\ln y_n^t - \ln y_n^{t^*}] \\ &+ (1/2) \sum_{m=1}^M \{[\partial \ln C(\mathbf{w}^{t^*}, \mathbf{y}^{t^*}, t^*) / \partial \ln w_m] + [\partial \ln C(\mathbf{w}^t, \mathbf{y}^t, t) / \partial \ln w_m]\} [\ln w_m^t - \ln w_m^{t^*}] \\ &= -\tau_1 [t - t^*] + k \ln Q_T(\mathbf{p}^{t^*}, \mathbf{p}^t, \mathbf{y}^{t^*}, \mathbf{y}^t) + \ln P_T(\mathbf{w}^{t^*}, \mathbf{w}^t, \mathbf{x}^{t^*}, \mathbf{x}^t) \end{aligned} \quad (7.117)$$

where  $Q_T(\mathbf{p}^{t^*}, \mathbf{p}^t, \mathbf{y}^{t^*}, \mathbf{y}^t)$  is the *Törnqvist* (1936)[373] (1937)[374] *quantity index* for output growth between periods  $t^*$  and  $t$  and  $P_T(\mathbf{w}^{t^*}, \mathbf{w}^t, \mathbf{x}^{t^*}, \mathbf{x}^t)$  is the *Törnqvist input price index* for input price growth between periods  $t^*$  and  $t$ . For  $t \leq t^*$ , we still have the following counterparts to equations (7.86):

$$\ln Q_T^*(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t) = -\tau_0 + k \ln Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t); \quad t \leq t^*. \quad (7.118)$$

Using (7.117), for  $t \geq t^* + 1$ , we have the following estimating equations:

$$\ln Q_T^*(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t) = -\tau_1 + k \ln Q_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t); \quad t \geq t^* + 1. \quad (7.119)$$

Thus changing rates of technical progress can be modeled using dummy variables in a simple regression model.

In the following section, we return to a discussion of the problems that the inequality (7.100) raises. Recall that this inequality showed that regressing output growth on input growth led to a direct measure of returns to scale,  $d$ , which was equal to or less than the indirect measure of returns to scale,  $1/b$ , where  $b$  was obtained by regressing input growth on output growth. Unfortunately, empirical

experience shows that there is usually a large difference between these two methods of estimating returns to scale, with the direct measure being close to one and the indirect measure usually being very much greater than one. Since both the input and output growth rates are generally measured with error, both estimates (for  $b$  and  $d$ ) will usually be biased downwards so that the direct measure of returns to scale,  $d$ , will usually be too low and the indirect measure,  $1/b$ , will be too big.

The point is this: if we have a single linear regression with two jointly dependent variables and we decide to estimate the structural parameters in the model by running two conditional regressions, one where  $y$  is the dependent variable and one where  $x$  is the dependent variable, and if we want a relatively large or small estimator for  $\alpha$  (in order to please a client for example), then we can strategically choose to run either (7.96) or (7.97) to achieve this objective. This is a very unsatisfactory state of affairs. Again, there is a lack of reproducibility due to the possibility that different applied economists will choose to run different conditional regressions.

In the following section, we ask whether the use of an instrumental variable method of estimation could eliminate these biases.

## 7.11 Can the Use of Instrumental Variables Lead to Better Estimates of Returns to Scale?

We will generalize slightly the problem studied in the previous section. Let  $\mathbf{X}$ ,  $\mathbf{Y}$  and  $\mathbf{z}$  be exogenous vectors of  $N$  variables measured without error<sup>\*83</sup> and suppose that these vectors satisfy the following exact relationship for some parameters  $\alpha$  and  $\beta$ :

$$\mathbf{Y} = \alpha\mathbf{X} + \mathbf{z}\beta. \quad (7.120)$$

Now suppose that  $\mathbf{X}$  and  $\mathbf{Y}$  cannot be observed precisely but observable estimates for these vectors are available, say  $\mathbf{y}$  and  $\mathbf{x}$ , and they satisfy the following equations:

$$\mathbf{y} = \mathbf{Y} + \mathbf{u}; \quad (7.121)$$

$$\mathbf{x} = \mathbf{X} + \mathbf{v} \quad (7.122)$$

where the independently distributed random variables  $\mathbf{u}$  and  $\mathbf{v}$  satisfy:

$$E\mathbf{u} = \mathbf{0}_N; \quad E\mathbf{u}\mathbf{u}^T = \sigma_u^2; \quad (7.123)$$

$$E\mathbf{v} = \mathbf{0}_N; \quad E\mathbf{v}\mathbf{v}^T = \sigma_v^2. \quad (7.124)$$

Substituting (7.121) and (7.122) into the exact model (7.120) leads to the following stochastic model:

$$\mathbf{y} = \mathbf{x}\alpha + \mathbf{z}\beta + \mathbf{e} \quad (7.125)$$

where  $\mathbf{e}$  is defined as

$$\mathbf{e} \equiv \alpha\mathbf{v} - \mathbf{u}. \quad (7.126)$$

Let  $\mathbf{w}$  be an exogenous vector of *instruments*. Premultiply both sides of (7.125) by the transpose of the  $N \times 2$  matrix  $[\mathbf{w}, \mathbf{z}]$  (so that we are choosing  $\mathbf{w}$  as the instrument vector for  $\mathbf{x}$  and  $\mathbf{z}$  as the

---

<sup>\*83</sup> In order to relate the model in this section to the models presented in the previous section, specialize the vector  $\mathbf{z}$  to be the vector of ones,  $\mathbf{1}_N$ . Note also that we have replaced the number of observations  $T$  by  $N$  in order to reduce confusion with the use of the symbol  $T$  to denote transposition of a vector. We have also changed our notation for  $\mathbf{X}$  and  $\mathbf{Y}$ .

instrument vector for  $\mathbf{z}$ ). Taking expectations of both sides of the resulting system of equations, we obtain the following system of 2 equations:<sup>\*84</sup>

$$\begin{bmatrix} \mathbf{w}^T \mathbf{X} & \mathbf{w}^T \mathbf{z} \\ \mathbf{z}^T \mathbf{X} & \mathbf{z}^T \mathbf{z} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbf{w}^T \mathbf{Y} \\ \mathbf{z}^T \mathbf{Y} \end{bmatrix}. \quad (7.127)$$

$$\begin{bmatrix} \alpha^* \\ \beta^* \end{bmatrix} \equiv \begin{bmatrix} \mathbf{w}^T \mathbf{x} & \mathbf{w}^T \mathbf{z} \\ \mathbf{z}^T \mathbf{x} & \mathbf{z}^T \mathbf{z} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{w}^T \mathbf{y} \\ \mathbf{z}^T \mathbf{y} \end{bmatrix}. \quad (7.128)$$

Comparing (7.128) with (7.127), it can be seen that we have replaced the unobserved vectors  $\mathbf{X}$  and  $\mathbf{Y}$  in (7.127) by the observed vectors  $\mathbf{x}$  and  $\mathbf{y}$  in (7.128). Using (7.120)-(7.124), it can be verified that  $\alpha^*$  and  $\beta^*$  are unbiased estimators for  $\alpha$  and  $\beta$ . Inverting the  $2 \times 2$  matrix in (7.128) leads to the following estimator for  $\alpha^*$ :

$$\begin{aligned} \alpha^* &= [\mathbf{w}^T \mathbf{y} - \mathbf{w}^T \mathbf{z} (\mathbf{z}^T \mathbf{z})^{-1} \mathbf{z}^T \mathbf{y}] / [\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mathbf{z} (\mathbf{z}^T \mathbf{z})^{-1} \mathbf{z}^T \mathbf{x}] \\ &= [\mathbf{w}^T \mathbf{y} - \mathbf{w}^T \mathbf{P} \mathbf{y}] / [\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mathbf{P} \mathbf{x}] \\ &= \mathbf{w}^T \mathbf{M} \mathbf{y} / \mathbf{w}^T \mathbf{M} \mathbf{x} \end{aligned} \quad (7.129)$$

where the projection matrices  $\mathbf{P}$  and  $\mathbf{M}$  are defined as follows:

$$\mathbf{P} \equiv \mathbf{z} (\mathbf{z}^T \mathbf{z})^{-1} \mathbf{z}^T; \quad (7.130)$$

$$\mathbf{M} \equiv \mathbf{I}_N - \mathbf{P}. \quad (7.131)$$

Once  $\alpha^*$  has been determined via (7.129), the second equation in (7.128) can be used to determine  $\beta^*$ :

$$\beta^* = [\mathbf{z}^T \mathbf{y} - \mathbf{z}^T \mathbf{x} \alpha^*] / \mathbf{z}^T \mathbf{z}. \quad (7.132)$$

Note that different choices of the vector of instruments  $\mathbf{w}$  affect  $\beta^*$  only by the effects of a change in  $\alpha^*$ . In particular, if  $\mathbf{z}^T \mathbf{x}$  is positive, then changing  $\mathbf{w}$  so that  $\alpha^*$  *increases* will *decrease*  $\beta^*$ .

The good feature of the instrumental variable estimator defined by (7.129) is that it is *symmetric in  $\mathbf{x}$  and  $\mathbf{y}$* ; i.e., if we looked at the *inverse regression* between  $\mathbf{x}$  and  $\mathbf{y}$  defined by

$$\mathbf{x} = \gamma \mathbf{y} + \mathbf{z} \delta + \mathbf{e}^*, \quad (7.133)$$

then using the matrix of instruments  $[\mathbf{w}, \mathbf{z}]$  on (7.133) would lead to the following estimator for  $\gamma$ :

$$\gamma^* = \mathbf{w}^T \mathbf{M} \mathbf{x} / \mathbf{w}^T \mathbf{M} \mathbf{y} = 1 / \alpha^*. \quad (7.134)$$

The model defined by (7.120) and the subsequent equations can readily be generalized to the case where  $\mathbf{z}$  is replaced by an exogenous  $N \times K$  matrix of variables  $\mathbf{Z}$  and the scalar parameter  $\beta$  is replaced by the vector of parameters,  $\boldsymbol{\beta} \equiv [\beta_1, \dots, \beta_K]^T$ . The model counterpart to (7.120) is now:

$$\mathbf{Y} = \alpha \mathbf{X} + \mathbf{Z} \boldsymbol{\beta} \quad (7.135)$$

where  $\mathbf{Y}$  and  $\mathbf{X}$  still satisfy assumptions (7.121)-(7.124). Substituting (7.121)-(7.124) into (7.135) leads to the following linear regression model:

$$\mathbf{y} = \mathbf{x} \alpha + \mathbf{Z} \boldsymbol{\beta} + \mathbf{e} \quad (7.136)$$

<sup>\*84</sup> Assuming that  $\mathbf{z} \neq \mathbf{0}_N$ , we require that  $\mathbf{w}^T \mathbf{M} \mathbf{x} \neq 0$  so that the inverse of the  $2 \times 2$  matrix exists. The projection matrix  $\mathbf{M}$  is defined below by (7.131).

where  $\mathbf{e}$  is defined as

$$\mathbf{e} \equiv \alpha \mathbf{v} - \mathbf{u}. \quad (7.137)$$

Again let  $\mathbf{w}$  be an exogenous vector of *instruments*. Premultiply both sides of (7.136) by the transpose of the  $N \times (K + 1)$  matrix  $[\mathbf{w}, \mathbf{Z}]$  (so that we are choosing  $\mathbf{w}$  as the instrument vector for  $\mathbf{x}$  and  $\mathbf{Z}$  as the instrument matrix for the matrix of exogenous variables  $\mathbf{Z}$  that are measured without error). Taking expectations of both sides of the resulting system of equations, we obtain the following system of  $1 + K$  equations:

$$\begin{aligned} \mathbf{w}^T \mathbf{Y} &= \mathbf{w}^T \mathbf{X} \alpha + \mathbf{w}^T \mathbf{Z} \beta; \\ \mathbf{Z}^T \mathbf{Y} &= \mathbf{Z}^T \mathbf{X} \alpha + \mathbf{Z}^T \mathbf{Z} \beta. \end{aligned} \quad (7.138)$$

Now replace  $\mathbf{Y}$  by  $\mathbf{y}$  and  $\mathbf{X}$  by  $\mathbf{x}$  in the above equations and replace  $\alpha$  and  $\beta$  by their instrumental variable estimators,  $\alpha^*$  and  $\beta^*$ , and we obtain the following  $1 + K$  equations:

$$\begin{aligned} \mathbf{w}^T \mathbf{x} \alpha^* + \mathbf{w}^T \mathbf{Z} \beta^* &= \mathbf{w}^T \mathbf{y}; \\ \mathbf{Z}^T \mathbf{x} \alpha^* + \mathbf{Z}^T \mathbf{Z} \beta^* &= \mathbf{Z}^T \mathbf{y}. \end{aligned} \quad (7.139)$$

Now solve equations (7.139) for  $\alpha^*$  and  $\beta^*$  and we obtain the following counterpart to (7.128):

$$\begin{bmatrix} \alpha^* \\ \beta^* \end{bmatrix} \equiv \begin{bmatrix} \mathbf{w}^T \mathbf{x} & \mathbf{w}^T \mathbf{Z} \\ \mathbf{Z}^T \mathbf{x} & \mathbf{Z}^T \mathbf{Z} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{w}^T \mathbf{y} \\ \mathbf{Z}^T \mathbf{y} \end{bmatrix}. \quad (7.140)$$

Using partitioned matrices, the inverse matrix in (7.140) is equal to:<sup>\*85</sup>

$$\begin{bmatrix} \mathbf{w}^T \mathbf{x} & \mathbf{w}^T \mathbf{Z} \\ \mathbf{Z}^T \mathbf{x} & \mathbf{Z}^T \mathbf{Z} \end{bmatrix}^{-1} = \frac{1}{\mathbf{w}^T \mathbf{M} \mathbf{x}} \begin{bmatrix} 1 & -\mathbf{w}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \\ -(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{w} & \mathbf{w}^T \mathbf{M} \mathbf{x} (\mathbf{Z}^T \mathbf{Z})^{-1} - (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{x} \mathbf{w}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \end{bmatrix} \quad (7.141)$$

where the  $N \times N$  projection matrix  $\mathbf{M}$  is now defined as follows:

$$\mathbf{M} \equiv \mathbf{I}_N - \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T. \quad (7.142)$$

**Problem 16** (a) Use (7.140) and (7.141) to show that:

$$\alpha^* = \mathbf{w}^T \mathbf{M} \mathbf{y} / \mathbf{w}^T \mathbf{M} \mathbf{x} \quad (7.143)$$

where  $\mathbf{M}$  is defined by (7.142).

(b) Once  $\alpha^*$  has been determined, show that:

$$\beta^* = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T [\mathbf{y} - \mathbf{x} \alpha^*]. \quad (7.144)$$

*Hint:* Use the last  $K$  equations in (7.139).

Note that again,  $\beta^*$  does not depend on the choice of the vector of instrumental variables  $\mathbf{w}$  except through the dependence of  $\alpha^*$  on  $\mathbf{w}$  via (7.143).

Again, it is easy to show that the instrumental variable estimator for  $\alpha$  defined by (7.143) is *symmetric in  $\mathbf{x}$  and  $\mathbf{y}$* ; i.e., if we looked at the *inverse regression* between  $\mathbf{x}$  and  $\mathbf{y}$  defined by

$$\mathbf{x} = \gamma \mathbf{y} + \mathbf{Z} \delta + \mathbf{e}^*, \quad (7.145)$$

<sup>\*85</sup> We assume that  $\mathbf{Z}$  is of full rank  $K < N$  so that  $(\mathbf{Z}^T \mathbf{Z})^{-1}$  exists. We also require that  $\mathbf{w}^T \mathbf{M} \mathbf{x} \neq 0$ .

then using the matrix of instruments  $[\mathbf{w}, \mathbf{Z}]$  on (7.145) would lead to the following estimator for  $\gamma$ :

$$\gamma^* = \mathbf{w}^T \mathbf{M}\mathbf{x} / \mathbf{w}^T \mathbf{M}\mathbf{y} = 1/\alpha^*. \quad (7.146)$$

However, the above theory gives no indication on how to choose the vector of instruments. If we happen to choose  $\mathbf{w}$  so that it is orthogonal to  $\mathbf{M}\mathbf{y}$ , then  $\mathbf{w}^T \mathbf{M}\mathbf{y} = 0$  and  $\alpha^* = 0$ . If we happen to choose a sequence of  $\mathbf{w}$ 's so that the limiting  $\mathbf{w}$  is orthogonal to  $\mathbf{M}\mathbf{x}$ , then we would obtain limiting estimates of  $\alpha^*$  that approached plus or minus infinity! *This illustrates the basic nonreproducibility property of instrumental variable estimation for finite samples: almost anything can happen, depending on the choice of the instrumental variable.*<sup>\*86</sup>

**Problem 17** Let  $\mathbf{A} = \mathbf{A}^T$  be a positive semidefinite  $N \times N$  symmetric matrix. Let  $\mathbf{x}$  and  $\mathbf{y}$  be  $N$  dimensional vectors. Show that the following generalization of the Cauchy Schwarz inequality holds:

$$(\mathbf{x}^T \mathbf{A}\mathbf{y})^2 \leq (\mathbf{x}^T \mathbf{A}\mathbf{x})(\mathbf{y}^T \mathbf{A}\mathbf{y}). \quad (a)$$

*Hint:* You may find the concept of a *square root matrix* for a positive semidefinite matrix helpful. From matrix algebra, we know that every symmetric matrix has the following eigenvalue-eigenvector decomposition with the following properties: there exist  $N \times N$  matrices  $\mathbf{U}$  and  $\mathbf{\Lambda}$  such that

$$\mathbf{U}^T \mathbf{A}\mathbf{U} = \mathbf{\Lambda}; \quad (b)$$

$$\mathbf{U}^T \mathbf{U} = \mathbf{I}_N \quad (c)$$

where  $\mathbf{\Lambda}$  is a diagonal matrix with the eigenvalues of  $\mathbf{A}$  on the main diagonal and  $\mathbf{U}$  is an orthonormal matrix. Note that  $\mathbf{U}$  is the inverse of  $\mathbf{U}^T$ . Hence premultiply both sides of (b) by  $\mathbf{U}$  and postmultiply both sides of (b) by  $\mathbf{U}^T$  in order to obtain the following equation:

$$\begin{aligned} \mathbf{A} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{U}^T \text{ where we use the assumption that } \mathbf{A} \text{ is positive semidefinite and we} \\ &\quad \text{define } \mathbf{\Lambda}^{1/2} \text{ to be a diagonal matrix with diagonal elements equal to} \\ &\quad \text{the nonnegative square roots of the diagonal elements of } \mathbf{\Lambda} \text{ (which} \\ &\quad \text{are the nonnegative eigenvalues of } \mathbf{A}, \lambda_1, \dots, \lambda_N. \\ &= \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^T \quad \text{using (c)} \\ &= \mathbf{B}^T\mathbf{B} \end{aligned} \quad (d)$$

where the  $N \times N$  *square root matrix*  $\mathbf{B}$  is defined as

$$\mathbf{B} \equiv \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^T. \quad (e)$$

Note that  $\mathbf{B}$  is symmetric so that

$$\mathbf{B} = \mathbf{B}^T \quad (f)$$

and thus we can also write  $\mathbf{A}$  as

$$\mathbf{A} = \mathbf{B}\mathbf{B}. \quad (g)$$

<sup>\*86</sup> By nonreproducibility, we mean that independent investigators will generally choose a different vector of instruments  $\mathbf{w}$ , thus leading to different estimators for  $\alpha^*$ . Put another way, we do not have a general theory on how to pick an instrumental variable which would be accepted by all applied economists working on the particular problem at hand.

**Problem 18** An  $N \times N$  matrix  $\mathbf{M}$  is a projection matrix if it satisfies the following 2 properties:

$$\mathbf{M} = \mathbf{M}^T; \tag{a}$$

$$\mathbf{M} = \mathbf{M}\mathbf{M}. \tag{b}$$

Show that the  $\mathbf{M}$  defined by (7.142),  $\mathbf{M} \equiv \mathbf{I}_N - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ , is a projection matrix.

**Problem 19** Let  $\mathbf{M}$  be an  $N \times N$  projection matrix. Show that the eigenvalues of  $\mathbf{M}$  must all equal 0 or 1.

*Hint:* From part (d) of problem 17, we have

$$\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{a}$$

where

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_N \tag{b}$$

and  $\mathbf{\Lambda}$  is a diagonal matrix with the eigenvalues of  $\mathbf{M}$  on the main diagonal. Now substitute (a) into 18 (b) to get:

$$\begin{aligned} \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\mathbf{\Lambda}\mathbf{\Lambda}\mathbf{U}^T \quad \text{using (b)}. \end{aligned} \tag{c}$$

Now premultiply both sides of (c) by  $\mathbf{U}^T$  and postmultiply both sides of (c) by  $\mathbf{U}$  to get

$$\mathbf{\Lambda} = \mathbf{\Lambda}\mathbf{\Lambda}. \tag{d}$$

Equations (d) say that each eigenvalue of  $\mathbf{M}$ , say  $\lambda_n$ , satisfies the equation

$$\lambda_n = \lambda_n\lambda_n; \quad n = 1, \dots, N. \tag{e}$$

**Problem 20** Use problems 17 and 19 to show that if  $\mathbf{M}$  is the projection matrix defined by (7.142), then the following generalized Cauchy Schwarz inequality is satisfied for any two  $N$  dimensional vectors  $\mathbf{x}$  and  $\mathbf{y}$ :

$$(\mathbf{x}^T\mathbf{M}\mathbf{y})^2 \leq (\mathbf{x}^T\mathbf{M}\mathbf{x})(\mathbf{y}^T\mathbf{M}\mathbf{y}). \tag{7.147}$$

**Problem 21** Consider the linear regression model  $\mathbf{y} = \mathbf{x}\alpha + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e}$  defined by (7.136). Let  $\hat{\alpha}$  be the least squares estimator for the parameter  $\alpha$ . Show that

$$\hat{\alpha} = \mathbf{x}^T\mathbf{M}\mathbf{y}/\mathbf{x}^T\mathbf{M}\mathbf{x} \tag{7.148}$$

where  $\mathbf{M}$  is defined by (7.142).

*Hint:* Use the normal equations for the least squares regression model (the first order conditions for the unconstrained least squares minimization problem) and the results in Problem 16 above. Thus if we use  $\mathbf{w} = \mathbf{x}$  as our vector of instruments, our instrumental variable estimator becomes the ordinary least squares estimator for  $\alpha$ , where we regress  $\mathbf{y}$  on  $\mathbf{x}$ .

**Problem 22** Consider the linear regression model  $\mathbf{x} = \gamma\mathbf{y} + \mathbf{Z}\boldsymbol{\delta} + \mathbf{e}^*$  defined by (7.145). Let  $\hat{\gamma}$  be the least squares estimator for the parameter  $\gamma$ .

(a) Show that

$$\hat{\gamma} = \mathbf{x}^T\mathbf{M}\mathbf{y}/\mathbf{y}^T\mathbf{M}\mathbf{y} \tag{7.149}$$

where  $\mathbf{M}$  is defined by (7.142).

(b) Thus since the reciprocal of  $\hat{\gamma}$  is an estimator for the parameter  $\alpha$ , we have the following estimator for  $\alpha$ :

$$\tilde{\alpha} \equiv 1/\hat{\gamma} = \mathbf{y}^T \mathbf{M} \mathbf{y} / \mathbf{x}^T \mathbf{M} \mathbf{y}. \quad (7.150)$$

Show that  $\tilde{\alpha}$  is also the instrumental variable estimator for  $\alpha$  in the model (7.136) when we choose the vector of instruments  $\mathbf{w} = \mathbf{y}$ .

*Hint:* Just use the results in Problem 16 and the fact that  $\mathbf{M}$  is a symmetric matrix.

**Problem 23** Suppose that  $\mathbf{x}^T \mathbf{M} \mathbf{y} > 0$  so that  $\mathbf{M} \mathbf{x}$  and  $\mathbf{M} \mathbf{y}$  are positively correlated; i.e., when we project the vectors  $\mathbf{x}$  and  $\mathbf{y}$  onto the subspace orthogonal to the subspace spanned by  $\mathbf{Z}$ , we find that  $(\mathbf{M} \mathbf{x})^T (\mathbf{M} \mathbf{y}) = \mathbf{x}^T \mathbf{M}^T \mathbf{M} \mathbf{y} = \mathbf{x}^T \mathbf{M} \mathbf{M} \mathbf{y} = \mathbf{x}^T \mathbf{M} \mathbf{y} > 0$ , so that these projection vectors are positively correlated.

(a) Show that:

$$\hat{\alpha} \equiv \mathbf{x}^T \mathbf{M} \mathbf{y} / \mathbf{x}^T \mathbf{M} \mathbf{x} \leq \mathbf{y}^T \mathbf{M} \mathbf{y} / \mathbf{x}^T \mathbf{M} \mathbf{y} \equiv \tilde{\alpha} \quad (7.151)$$

where  $\hat{\alpha}$  and  $\tilde{\alpha}$  were defined in problems 21 and 22.

(b) Derive a counterpart to (7.151) if  $\mathbf{x}^T \mathbf{M} \mathbf{y} < 0$ .

**Problem 24** When we have two separate estimators for a parameter, such as  $\hat{\alpha}$  and  $\tilde{\alpha}$  for  $\alpha$  in the last problem, it is natural to think that a better estimator is a symmetric average of the two estimators. Thus two natural averages are the geometric mean and arithmetic mean of  $\hat{\alpha}$  and  $\tilde{\alpha}$  defined by (7.152) and (7.153) respectively:<sup>\*87</sup>

$$\begin{aligned} \alpha_G &\equiv [\hat{\alpha} \tilde{\alpha}]^{1/2} = [(\mathbf{x}^T \mathbf{M} \mathbf{y} / \mathbf{x}^T \mathbf{M} \mathbf{x})(\mathbf{y}^T \mathbf{M} \mathbf{y} / \mathbf{x}^T \mathbf{M} \mathbf{y})]^{1/2} \\ &= \begin{cases} [\mathbf{y}^T \mathbf{M} \mathbf{y}]^{1/2} / [\mathbf{x}^T \mathbf{M} \mathbf{x}]^{1/2} & \text{if } \mathbf{x}^T \mathbf{M} \mathbf{y} > 0; \\ -[\mathbf{y}^T \mathbf{M} \mathbf{y}]^{1/2} / [\mathbf{x}^T \mathbf{M} \mathbf{x}]^{1/2} & \text{if } \mathbf{x}^T \mathbf{M} \mathbf{y} < 0; \end{cases} \end{aligned} \quad (7.152)$$

$$\alpha_A \equiv (1/2)[\hat{\alpha} + \tilde{\alpha}] = (1/2)[(\mathbf{x}^T \mathbf{M} \mathbf{y} / \mathbf{x}^T \mathbf{M} \mathbf{x}) + (\mathbf{y}^T \mathbf{M} \mathbf{y} / \mathbf{x}^T \mathbf{M} \mathbf{y})]. \quad (7.153)$$

However, we also have two separate estimators for  $\gamma \equiv 1/\alpha$ ; namely  $\hat{\gamma}$  defined by (7.149) and  $\tilde{\gamma} \equiv 1/\tilde{\alpha}$  defined as follows:

$$\hat{\gamma} \equiv \mathbf{x}^T \mathbf{M} \mathbf{y} / \mathbf{y}^T \mathbf{M} \mathbf{y}; \quad (7.154)$$

$$\tilde{\gamma} \equiv \mathbf{x}^T \mathbf{M} \mathbf{x} / \mathbf{x}^T \mathbf{M} \mathbf{y}. \quad (7.155)$$

The geometric mean and arithmetic mean of  $\hat{\gamma}$  and  $\tilde{\gamma}$  are respectively:

$$\begin{aligned} \gamma_G &\equiv [\hat{\gamma} \tilde{\gamma}]^{1/2} = [(\mathbf{x}^T \mathbf{M} \mathbf{y} / \mathbf{y}^T \mathbf{M} \mathbf{y})(\mathbf{x}^T \mathbf{M} \mathbf{x} / \mathbf{x}^T \mathbf{M} \mathbf{y})]^{1/2} \\ &= \begin{cases} [\mathbf{x}^T \mathbf{M} \mathbf{x}]^{1/2} / [\mathbf{y}^T \mathbf{M} \mathbf{y}]^{1/2} & \text{if } \mathbf{x}^T \mathbf{M} \mathbf{y} > 0; \\ -[\mathbf{x}^T \mathbf{M} \mathbf{x}]^{1/2} / [\mathbf{y}^T \mathbf{M} \mathbf{y}]^{1/2} & \text{if } \mathbf{x}^T \mathbf{M} \mathbf{y} < 0; \end{cases} \end{aligned} \quad (7.156)$$

$$\gamma_A \equiv (1/2)[\hat{\gamma} + \tilde{\gamma}] = (1/2)[(\mathbf{x}^T \mathbf{M} \mathbf{y} / \mathbf{y}^T \mathbf{M} \mathbf{y}) + (\mathbf{x}^T \mathbf{M} \mathbf{x} / \mathbf{x}^T \mathbf{M} \mathbf{y})]. \quad (7.157)$$

Show that  $\gamma_G = 1/\alpha_G$  but in general,  $\gamma_A \neq 1/\alpha_A$ . Thus, if we do average our estimates of  $\alpha$  or  $\gamma$ , it seems preferable to use a geometric average over an arithmetic average.

<sup>\*87</sup> We assume that  $\mathbf{x}^T \mathbf{M} \mathbf{y} \neq 0$ . It is always the case that  $\mathbf{x}^T \mathbf{M} \mathbf{x} \geq 0$  and  $\mathbf{y}^T \mathbf{M} \mathbf{y} \geq 0$  since  $\mathbf{M}$  is positive semidefinite. However, since we assume that  $\mathbf{x}^T \mathbf{M} \mathbf{y} \neq 0$ , it must be the case that  $\mathbf{x}^T \mathbf{M} \mathbf{x} > 0$  and  $\mathbf{y}^T \mathbf{M} \mathbf{y} > 0$ .

Our conclusion from the results derived in this section is that the use of an instrumental variable is not going to lead to an estimator for  $\alpha$  that will be universally accepted by other applied economists. Different choices for the vector of instruments  $\mathbf{w}$  will frequently lead to very different estimates for  $\alpha$ .

What should we do in practice when estimating returns to scale? Since input growth generally precedes output growth, it probably makes more sense to condition on input growth and choose input growth as the exogenous variable. Also output growth generally has a larger variance than input growth and so it is more likely that output growth equals a constant (close to 1) times input growth plus a random error term; i.e., the model that regresses output growth on input growth is more likely to satisfy the assumption in a linear regression that the exogenous variables be uncorrelated with the error term in the regression.

## 7.12 References

- Adcock, R. J. (1878), "A Problem in Least Squares", *Analyst [Annals of Mathematics]* 5, 53-54.
- Allais, M. (1947), *Economie et Intérêt*, Paris: Imprimerie Nationale.
- Allen, R.G.D. (1939), "The Assumptions of Linear Regression", *Economica (New Series)* 6, 191-201.
- Allen, R.C. (1983), "Collective Invention", *Journal of Economic Behavior and Organization* 4, 1-24.
- Arrow, K.J. (1962), "The Economic Implications of Learning by Doing", *The Review of Economic Studies* 29, 155-173.
- Arrow, K.J. (1969), "Classificatory Notes on the Production and Transmission of Technological Knowledge", *American Economic Review* 59 (May), 29-35.
- Babbage, C. (1835), *On the Economy of Machinery and Manufactures*, Fourth Edition, reprinted by A. M. Kelley, New York, 1965.
- Balk, B. M. (1995), "Axiomatic Price Index Theory: A Survey", *International Statistical Review* 63, 69-93.
- Bartelsman, E. J. (1995), "Of Empty Boxes: Returns to Scale Revisited," *Economics Letters* 49, 59-67.
- Basu, S. and J. G. Fernald (1997), "Returns to Scale in U.S. Production: Estimates and Implications", *Journal of Political Economy* 105, 249-283.
- Basu, S. and J. G. Fernald (2002), "Aggregate Productivity and Aggregate Technology", *European Economic Review* 46, 963-991.
- Baumol, W.J. (1952), "The Transactions Demand For Cash: An Inventory Theoretic Approach", *Quarterly Journal of Economics* 66, 545-556.
- Bates, W. (2001), *How Much Government? The effects of high government spending on economic performance*, New Zealand Business Roundtable, Wellington, August 2001.
- Baumol, W.J. (1952), "The Transactions Demand For Cash: An Inventory Theoretic Approach", *Quarterly Journal of Economics* 66, 545-556.
- Burgess, D. F. (1974), "A Cost Minimization Approach to Import Demand Equations," *Review of Economics and Statistics* 56 (2): 224-234.
- Caves, D., L.R. Christensen and W.E. Diewert (1982), "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity", *Econometrica* 50, 1392-1414.
- Christensen, L.R., D.W. Jorgenson and L.J. Lau (1971), "Conjugate Duality and the Transcendental Logarithmic Production Function," *Econometrica* 39, 255-256.
- Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton, New Jersey: Princeton University Press.

- Diewert, W. E. (1974), "Applications of Duality Theory", pp. 106-171 in *Frontiers of Quantitative Economics*, Volume 2, M. D. Intriligator and D. A. Kendrick (eds.), Amsterdam: North-Holland.
- Diewert, W.E. (1976), "Exact and Superlative Index Numbers", *Journal of Econometrics* 4, 114-145.
- Diewert, W.E. (1981), "The Comparative Statics of Industry Long Run Equilibrium", *The Canadian Journal of Economics* 14, 78-92.
- Diewert, W.E. (1983), "The Measurement of Waste within the Production Sector of an Open Economy", *Scandinavian Journal of Economics* 85, 159-179.
- Diewert, W.E. (1992), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *Journal of Productivity Analysis* 3, 211-248.
- Diewert, W.E. (1992a), "The Measurement of Productivity", *Bulletin of Economic Research* 44:3, 163-198.
- Diewert, W.E. (1992b), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *Journal of Productivity Analysis* 3, 211-248.
- Diewert, W.E. (1993), "The Early History of Price Index Research", pp. 33-65 in *Essays in Index Number Theory*, Volume 1, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North-Holland.
- Diewert, W.E. (1997), "Commentary on Mathew D. Shapiro and David W. Wilcox: Alternative Strategies for Aggregating Price in the CPI", *The Federal Reserve Bank of St. Louis Review*, Vol. 79:3, (May/June), 127-137.
- Diewert, W.E. (2001), "Productivity Growth and the Role of Government", Discussion Paper No. 01-13, Department of Economics, The University of British Columbia, Vancouver, Canada, V6T 1Z1. <http://www.econ.ubc.ca/discpapers/dp0113.pdf>
- Diewert, W.E. (2004), "A New Axiomatic Approach to Index Number Theory", Discussion Paper 04-05, Department of Economics, University of British Columbia, Vancouver, Canada, V6T 1Z1.
- Diewert, W.E. and K.J. Fox (1999), "Can Measurement Error Explain the Productivity Paradox?", *Canadian Journal of Economics* 32, 251-280. Also available at: <http://web.arts.ubc.ca/econ/diewert/hmpgdie.htm>
- Diewert, W.E. and K.J. Fox (2004), "On the Estimation of Returns to Scale, Technical Progress and Monopolistic Markups", Discussion Paper 04-09, Department of Economics, University of British Columbia, July.
- Diewert, W.E. and D. Lawrence, (1994), *The Marginal Costs of Taxation in New Zealand*, Report prepared for the New Zealand Business Roundtable by Swan Consultants, Canberra.
- Diewert, W.E. and D. Lawrence, (2002), "The Deadweight Costs of Capital Taxation in Australia", pp. 103-167 in *Efficiency in the Public Sector*, Kevin J. Fox (ed.), Boston: Kluwer Academic Publishers.
- Diewert, W.E. and C.J. Morrison (1986), "Adjusting Output and Productivity Indexes for Changes in the Terms of Trade", *Economic Journal* 96, 659-679.
- Diewert, W.E. and A.O. Nakamura (1999), "Benchmarking and the Measurement of Best Practice Efficiency: An Electricity Generation Application", *Canadian Journal of Economics* 32, 570-588.
- Diewert, W.E. and A.O. Nakamura (2003), "Index Number Concepts, Measures and Decompositions of Productivity Growth", *Journal of Productivity Analysis* 19, 127-159.
- Driffill, E.J. and H.S. Rosen (1983), "Taxation and Excess Burden: A Life Cycle Perspective", *International Economic Review* 24, 671-683.
- Dupor, B., L. Lochner, C. Taber, and M.B. Wittekind (1996), "Some Effects of Taxes on Schooling and Training", *American Economic Review* 86 (May), 340-346.
- Edgeworth, F.Y. (1888), "The Mathematical Theory of Banking", *Journal of the Royal Statistical Society* 51, 113-127.

- Eichhorn, W. (1978), *Functional Equations in Economics*, London: Addison-Wesley.
- Eichhorn, W. and J. Voeller (1976), *Theory of the Price Index*, Lecture Notes in Economics and Mathematical Systems, Vol. 140, Berlin: Springer-Verlag.
- Feldstein, M. (1996), "How Big Should Government Be?", Working Paper 5868, National Bureau of Economic Research, Cambridge, Massachusetts.
- Fisher, I. (1911), *The Purchasing Power of Money*, London: Macmillan.
- Fisher, I. (1922), *The Making of Index Numbers*, Boston: Houghton-Mifflin.
- Frisch, R. (1930), "Necessary and Sufficient Conditions Regarding the Form of an Index Number Which Shall Meet Certain of Fisher's Tests", *American Statistical Association Journal* 25, 397–406.
- Frisch, R. (1936), "Annual Survey of General Economic Theory: The Problem of Index Numbers", *Econometrica* 4, 1-39.
- Funke, H. and J. Voeller (1978), "A Note on the Characterisation of Fisher's Ideal Index", pp. 177–181 in *Theory and Applications of Economic Indices*, W. Eichhorn, R. Henn, O. Opitz, and R.W. Shephard (eds.), Würzburg: Physica-Verlag.
- Funke, H., and J. Voeller (1979), "Characterization of Fisher's Ideal Index by Three Reversal Tests", *Statistische Hefte* 20, 54–60.
- Fox, K.J. and U. Kohli (1998), "GDP Growth, Terms of Trade Effects and Total Factor Productivity", *The Journal of International Trade and Economic Development* 7:1, 87-110.
- Golub, G. H. and C. F. Van Loan (1980), "An Analysis of the Total Least Squares Problem", *Siam Journal of Numerical Analysis* 17, 883-893.
- Green, J.B. (1915), "The Perpetual Inventory in Practical Stores Operation", *The Engineering Magazine* 48, 879-888.
- Hadley, G. and T.M. Whitin (1963), *Analysis of Inventory Systems*, Englewood Cliffs, N.J.: Prentice-Hall.
- Hall, R.E. (1988), "The Relationship between Price and Marginal Cost in U. S. Industry", *Journal of Political Economy* 96, 921-947.
- Hall, R.E. (1990), "Invariance Properties of Solow's Productivity Residual", in *Growth, Productivity, Employment*, P. Diamond (ed.), Cambridge MA: MIT Press.
- Haltiwanger, J. (2000), "Aggregate Growth: What Have we Learned from Microeconomic Evidence?" Economics Department Working Paper No. 267, Paris: OECD.
- Harberger, Arnold (1998), "A Vision of the Growth Process," *American Economic Review* 88, 1-32.
- Harris, F.W. (1915), *Operations and Cost*, Chicago: A.W. Shaw Company.
- Harris, R.G. (1999), "Making a Case for Tax Cuts", paper prepared for the Business Council on National Issues *Global Agenda Initiative*.
- Harris, R.G. (2001), "Determinants of Canadian Productivity Growth: Issues and Prospects", Forthcoming in *Productivity Issues in a Canadian Context*, A. Sharpe and S. Rao (eds.), Montreal: McGill-Queen's Press.
- Hicks, J. (1969), *A Theory of Economic History*, London: Oxford university Press.
- Hicks, J. (1973), *Capital and Time: A Neo-Austrian Theory*, London: Oxford University Press.
- Jorgenson, D.W. and Z. Griliches (1967). "The Explanation of Productivity Change", *Review of Economic Studies* 34, 249–283.
- Jorgenson, D.W., and Z. Griliches (1972), "Issues of Growth Accounting: A Reply to Edward F. Denison", *Survey of Current Business* 55(5), part II, 65–94.
- Jorgenson, D.W. and M. Nishimizu (1978), "U.S. and Japanese Economic Growth, 1952–1974", *Economic Journal* 88, 707–726.

- Jorgenson, D.W. and K.-Y. Yun (1986), "Tax Policy and Capital Allocation", *Scandinavian Journal of Economics* 88, 355-377.
- Jorgenson, D.W. and K.-Y. Yun (1990), "Tax Reform and US Economic Growth", *Journal of Political Economy* 98(5), S151-S193.
- Jorgenson, D.W. and K.-Y. Yun (1991), *Tax Reform and the Cost of Capital*, Oxford: Clarendon Press.
- Kaldor, N. (1972), "The Irrelevance of Equilibrium Economics", *The Economic Journal* 82, 1237-1255.
- Kesselman, J.R. (1997), *General Payroll Taxes: Economics, Politics and Design*, Canadian Tax Paper No. 101, Toronto: The Canadian Tax Foundation.
- Kesselman, J.R. (2000), "Flat Taxes, Dual Taxes, Smart Taxes: Making the Best Choices", *Policy Matters*, Volume 1, no. 7, Montreal: The Institute for Research On Public Policy. Email: irpp@irpp.org
- Kohli, U. (1990), "Growth Accounting in the Open Economy: Parametric and Nonparametric Estimates", *Journal of Economic and Social Measurement* 16, 125-136.
- Kneller, R., M.F. Bleaney and N. Gemmell (1999), "Fiscal Policy and Growth: Evidence from OECD Countries", *Journal of Public Economics* 74:2, 171-190.
- Krugman, P. (1991), *Geography and Trade*, Cambridge, MA: The MIT Press.
- Laspeyres, E. (1871), "Die Berechnung einer mittleren Waarenpreissteigerung", *Jahrbücher für Nationalökonomie und Statistik* 16, 296-314.
- Lipsey, R.G. (2000), "Economies of Scale in Theory and Practice", unpublished paper available at: <http://www.sfu.ca/~rlipsey/res.html>
- Lipsey, R.G. and K. Carlaw (2000), "What does Total Factor Productivity Measure?", unpublished paper available at: <http://www.sfu.ca/~rlipsey/res.html>
- Madansky, A. (1959), "The Fitting of Straight Lines when both Variables are Subject to Error", *Journal of the American Statistical Association* 54, 173-206.
- Marshall, A. (1898), *Principles of Economics*, Fourth Edition (first edition 1890, eighth edition 1920), London: The Macmillan Co.
- Mintz, J.M. (1999), *Why Canada Must Undertake Business Tax Reform Soon*, Backgrounder, Toronto: C.D. Howe Institute.
- Morrison, C. and W.E. Diewert (1990), "Productivity Growth and Changes in the Terms of Trade in Japan and the United States", pp. 201-227 in *Productivity Growth in Japan and the United States*, C.R. Hulten (ed.), University of Chicago Press, Chicago.
- Nakajima, T., A. Nakamura and M. Nakamura (2002), "Japanese TFP Growth before and after the Financial Bubble: Japanese Manufacturing Industries", paper presented at the NBER, Cambridge MA, July 26, 2002.
- Nakajima, T., M. Nakamura and K. Yoshioka (1998), "An Index Number Method for Estimating Scale Economies and Technical Progress Using Time-Series of Cross-Section Data: Sources of Total Factor Productivity Growth for Japanese Manufacturing, 1964-1988", *Japanese Economic Review* 49, 310-334.
- Nakamura, A.O. and W.E. Diewert (2000), "Insurance for the Unemployed: Canadian Reforms and their Relevance for the United States", pp. 217-247 in *Long-Term Unemployment and Reemployment Policies*, L.J. Bassi and S.A. Woodbury (eds.), Stamford Connecticut: JAI Press.
- Nakamura, A.O. and P. Lawrence (1994), "Education, Training and Prosperity", John Deutsch Institute for the Study of Economic Policy (March), 235-279.
- Nordhaus, W.D. (1969), "Theory of Innovations: An Economic Theory of Technological Change", *American Economic Review* 59 (May), 18-28.

- Norman, R.G. and S. Bahiri (1972), *Productivity Measurement and Incentives*, Oxford: Butterworth-Heinemann.
- Paasche, H. (1874), "Über die Preisentwicklung der letzten Jahre nach den Hamburger Borsennotirungen", *Jahrbücher für Nationalökonomie und Statistik* 12, 168-178.
- Pierson, N.G. (1896), "Further Consideration on Index Numbers", *Economic Journal* 6, 127-131.
- Romer, P. (1994), "New Goods, Old Theory and the Welfare Costs of Trade Restrictions", *Journal of Development Economics* 43, 5-38.
- Samuelson, P.A. (1967), "The Monopolistic Competition Revolution", In *Monopolistic Competition Theory: Studies in Impact*, R.E. Kuenne (ed.), New York: John Wiley.
- Shephard, R.W. (1953), *Cost and Production Functions*, Princeton N.J.: Princeton University Press.
- Smith, A. (1963), *The Wealth of Nations*, Volume 1 (first published in 1776), Homewood, Illinois: Richard D. Irwin.
- The Economist* (2000), "New Zealand's Economy", London, December 2.
- Tobin, J. (1956), "The Interest Elasticity of Transactions Demand for Cash", *The Review of Economics and Statistics* 38, 241-247.
- Törnqvist, L. (1936), "The Bank of Finland's Consumption Price Index", *Bank of Finland Monthly Bulletin* 10, 1-8.
- Törnqvist, L. and E. Törnqvist (1937), "Vilket är förhållandet mellan finska markens ochsvenska kronans köpkraft?", *Ekonomiska Samfundets Tidskrift* 39, 1-39 reprinted as pp. 121-160 in *Collected Scientific Papers of Leo Törnqvist*, Helsinki: The Research Institute of the Finnish Economy, 1981.
- Walsh, B. (2000), "The Role of Tax Policy in Ireland's Economic Renaissance", *Canadian Tax Journal* 48:3, 658-673.
- Walsh, C.M. (1901), *The Measurement of General Exchange Value*, New York: Macmillan and Co.
- Walsh, C.M. (1921a), *The Problem of Estimation*, London: P.S. King & Son.
- Walsh, C. M. (1921b), "Discussion", *Journal of the American Statistical Association* 17, 537-544.
- Watson, W. (1999), "Labour Day, From the Front Porch", *Financial Post*, Toronto, Canada, September 8.
- Whitin, T.M. (1952), "Inventory Control in Theory and Practice", *Quarterly Journal of Economics* 66, 502-521.
- Whitin, T.M. (1957), *The Theory of Inventory Management*, Second edition, Princeton, N.J.: Princeton University Press.
- Young, A.A. (1928), "Increasing Returns and Economic Progress", *Economic Journal* 38, 527-542.
- Zeitsch, J. and D. Lawrence (1996), "Decomposing Economic Inefficiency in Base-Load Power Plants", *The Journal of Productivity Analysis* 7, 359-378.



## Chapter 8

# The Measurement of Capital

### 8.1 Introduction

“Capital (I am not the first to discover) is a very large subject, with many aspects; wherever one starts, it is hard to bring more than a few of them into view. It is just as if one were making pictures of a building; though it is the same building, it looks quite different from different angles.” John Hicks (1973; v)[226].

“Perhaps a more realistic motive for reading earlier writers is not to rediscover forgotten truths, but to gain a perspective of how present day ideas have evolved and, perhaps, by reading the original statements of important ideas, to see them more vividly and understand them more clearly.” Geoffrey Whittington (1980; 240)[400].

When a firm buys a durable capital input, it is not appropriate to charge the entire purchase price to the period when the input was purchased. If this is done, then profits will be *understated* in the period of purchase and *overstated* in subsequent periods when the durable input continues to contribute to production. Rather than charging the entire purchase price of the asset to the first period of use (and charging nothing for the subsequent periods of use), it would be more appropriate to charge a *rental price* or a *user cost* for the asset for each period that it is used. If there are market rental prices for the asset (and for the used asset as it ages), then these market rental prices could be used to price the services of the asset in each period. But frequently, such market rental prices do not exist and so other techniques must be used to price the period by period services of the asset. We will indicate some of these alternative techniques in this chapter.\*<sup>1</sup>

In section 8.2, we discuss some of the problems that occur when an economy is experiencing very high inflation. Under these conditions, it will be necessary for the national price statistician to shorten the accounting period (or give up price measurement altogether). We also discuss some problems relating to the beginning, middle and end of the period.

In section 8.3, we present the basic equations relating stocks and flows of capital assuming that data on the prices of vintages of a homogeneous capital good are available. This framework is not applicable under all circumstances\*<sup>2</sup> but it is a framework that will allow us to disentangle the effects of general price change, asset specific price change and depreciation.

Section 8.4 continues the theoretical framework that was introduced in section 8.3. We show how information on vintage asset prices, vintage rental prices and vintage depreciation rates are all

---

\*<sup>1</sup> The material in this chapter draws on Diewert and Lawrence (2000)[?] and Diewert (2001)[121] (2004a)[125] (2004b)[126] (2005a)[128] (2005b)[129].

\*<sup>2</sup> Most notably, our framework cannot deal with unique or one of a kind assets, which by definition, do not have vintages.

equivalent under certain assumptions; i.e., knowledge of any one of these three sequences or profiles is sufficient to determine the other two.

Section 8.5 discusses alternative sets of assumptions on nominal interest rates and anticipated asset price changes. We specify three different sets of assumptions that could be used in empirical implementations of the suggested methods.

Section 8.6 discusses the problems involved in aggregating over vintages of capital, both in forming capital stocks and capital services. Instead of the usual perpetual inventory method for aggregating over vintages, which assumes perfectly substitutable vintages of the same stock, we suggest the use of a superlative index number formula to do the aggregation.

Sections 8.7-8.10 show how the general algebra presented in sections 8.3 and 8.4 can be adapted to deal with four specific models of depreciation. The four models considered are the one hoss shay model, the geometric model of depreciation, straight line depreciation and the linear efficiency decline model.

Finally, inventory stocks are a type of capital input to production but the present System of National Accounts treatment of inventories and inventory change is somewhat confusing. Thus in Appendix 1, a theoretical framework that provides a unified treatment for measuring inventory change and the user cost of inventories is explained.\*<sup>3</sup> Appendix 2 gives some background information on the origins of the theoretical framework used in Appendix 1.

## 8.2 Inflation, the Length of the Accounting Period and the Measurement of Economic Activity

Our goal in this chapter is twofold: (1) to measure the price and quantity of the *stock* of reproducible capital held by a production unit (an establishment, a firm, an industry or an entire economy) at a *point* in time and (2) to measure the price and quantity of the *flow* of reproducible capital services utilized by a production unit over a *period* of time. In particular, we want to extend the procedures for measuring capital stocks and flows to cover situations where there is general price level change or *inflation*. In this section, we shall review some of the general measurement problems that arise when inflation is high.

When capital flows are measured, the normal period of time is either a year or a quarter. Under conditions of high inflation, the aggregation of homogeneous commodity flows within a quarter or a year is complicated by the fact that the within period transactions are valued at very different prices. The recent national income accounting literature explains the problem as follows:

“Conventional index number theory is mostly concerned with comparisons between *points* of time whereas, in national accounts, price and quantity comparisons have to be made between discrete *periods* of time. Significant changes in price and quantity flows may occur not only between different periods but also within a single accounting period, especially one as long as a year. Indeed, the central problem of accounting under high inflation is that prices are much higher at the end of the accounting period than at the beginning.” Peter Hill (1996; 11)[228].

“The underlying problem is not a traditional index number problem. It stems from the use of current value data as inputs into the calculation of indirect price or quantity measures under high inflation. Current accounts permit identical quantities of the same homogeneous product to be valued at very different prices during the course of the same year. Implicitly, quantities sold at higher prices later in the year are treated as if they were superior qualities when they

---

\*<sup>3</sup> This methodology is based on Diewert and Smith (1994)[150] and Diewert (2004b; 36)[126]. The accounting methodology can also be found in Diewert (2005b; 21-23)[129], Diewert, Mizobuchi and Nomura (2005)[147] and Diewert and Lawrence (2006)[144].

are not.” Peter Hill (1996; 12)[228].

“Under high inflation, the monetary value of flows of goods and services at different points of time within the same accounting period are not commensurate with each other because the unit of currency used as the *numeraire* is not stable. Adding together different quantities of the same good valued at different prices is equivalent, from a scientific point of view, to using different units of measurement for different sets of observations on the same variable. In the case of physical data, however, it is rather more obvious that adding quantities measured in grams to quantities measured in ounces is a futile procedure.” Peter Hill (1996; 32)[228].

“Before the preparation of the 1993 SNA, issues connected with high or significant inflation had not been dealt with at all in international recommendations concerning national accounts. Uneasiness especially with the recording of nominal interest had been often expressed, for instance in Europe and North America at the time of two digit inflation and above all in countries, like in Latin America, experiencing high or hyper inflation. In relation with the latter situations, uneasiness extended to the whole set of accounts, because, due to the significant rate of inflation within each year, annual accounts in current values could no longer be deemed homogeneous as regards the level of prices in each year. They combine intra-annual flows that are valued at very different prices and are not, strictly speaking, additive. The effect of the intra-annual change in the general price level can be neglected for the sake of simplicity only when the rate of inflation is low. When it is high, the meaning of annual accounts in current values becomes fuzzy.” André Vanoli (1998)[380].

“When inflation is high, the aggregation of flows from different periods becomes very much a case of ‘adding apples and bananas’— the flows at the end of the period will carry a much greater weight than the flows at the beginning of the period, so that the change on average will reflect development at the end of the period disproportionately. Annual national accounts at current prices become virtually meaningless and computation of national accounts at constant prices becomes very problematic.” Ezra Hadar and Soli Peleg (1998; 2)[205].

Of course, concern over the effects of general price level change has a much longer history in the general cost accounting literature; see Baxter (1984)[29], Tweedie and Whittington (1984)[377] and Whittington (1992)[401] for example.\*<sup>4</sup>

We now discuss in more detail the accounting problems caused by high inflation that are referred to in the above quotations. The basic problem is this: all discrete time economic theories and most of index number theory assumes that all of the transactions of a production unit in a homogeneous commodity within the accounting period can be represented by a *single price* and a *single quantity*. It is natural to let the single quantity be the *sum* of the quantities sold (in the case of an output) or the *sum* of the quantities purchased (in the case of an input). But then, if we want the single price times the single quantity to equal the *value* of transactions for the commodity in the period, the single price must equal the value of transactions divided by the sum of quantities purchased or sold; i.e., the single price must equal a *unit value*.\*<sup>5</sup> But when there is substantial inflation within the accounting period, unit values give a much higher weight to transactions that occur near the

\*<sup>4</sup> The inflation accounting literature extends back to Middleditch: “Today’s dollar is, then, a totally different unit from the dollar of 1897. As the general price level fluctuates, the dollar is bound to become a unit of different magnitude. To mix these units is like mixing inches and centimeters or measuring a field with a rubber tape-line.” Livingston Middleditch (1918; 114-115)[311].

\*<sup>5</sup> The early index number theorists Walsh (1901; 96)[389] (1921; 88)[390], Fisher (1922; 318)[188] and Davies (1924; 96)[67] all suggested unit values as the prices that should be inserted into a bilateral index number formula. Walsh nicely sums up the case for unit values as follows: “Some nice questions arise as to whether only what is consumed in the country, or only what is produced in it, or both together are to be counted; and also there are difficulties as to the single price quotation that is to be given at each period to each commodity, since this, too, must be an average. Throughout the country during the period a commodity is not sold at one price, nor even at one wholesale price in its principle market. Various quantities of it are sold at different prices, and the full value is obtained by adding all the sums spent (at the same stage in its advance towards the consumer),

end of the period compared to transactions that occurred near the beginning; it is as if the end of period transactions are being *artificially quality adjusted* to be more valuable than the beginning of the period transactions.

The obvious solution to this artificial implicit weighting problem is to choose the accounting period to be small enough so that the general inflation within the period is small enough to be ignored. This is precisely the solution suggested by the index number theorist Fisher\*<sup>6</sup> and the measurement economist Hicks: the length of the accounting period should be the Hicksian “week”:

“I shall define a week as that period of time during which variations in price can be neglected.”  
John R. Hicks (1946; 122)[222].

Thus it seems that there is a simple solution to the problem of constructing meaningful accounting period prices and quantities for homogeneous commodities when there is high inflation: simply shorten the accounting period!

Hill (1996)[228] however noted that there are at least three classes of problems associated with the above solution:

“In order to keep these issues in perspective, it is useful to summarise the problems created by continually shortening the accounting period.

1. The compilation of accounts for shorter time periods requires more information about the times at which various transactions take place. Enquiries may have to be conducted more frequently thereby creating additional costs for the data collectors. More burdens are also placed on the respondents supplying the information. In many cases, they may be unable to supply the necessary information because their own internal records and accounts do not permit them to do so, especially when they traditionally report their accounts for longer time periods, such as a year.

2. As production is a process which can extend over a considerable period of time, its measurement becomes progressively more difficult the shorter the accounting period. The problem is not confined to agriculture or forestry where many production processes take a year or more. The production of large fixed assets such as large ships, bridges, power stations, dams or the like can extend over several years. The output produced over shorter periods of time then has to be measured on the basis of work in progress completed each period. . . .

3. Because many transactions, especially large transactions, are not completed within the day, there are typically many receivables and payables outstanding at any given moment of time. They assume greater importance in relation to the flows as the accounting period is reduced. This makes it more difficult to reconcile the values of different flows in the accounts, especially if the two parties to the transaction perceive it as taking place at different times from each other and do not record it in the same way required by the system. . . . Peter Hill (1996; 34-35)[228].

Thus shortening the accounting period leads to increased costs for the statistical agency and the businesses being surveyed. Moreover, firm accounting is geared to years and quarters and it may

---

and the average price is found by dividing the total sum (or the full value) by the total quantities.” Correa Moylan Walsh (1921; 88)[390].

\*<sup>6</sup> “Essentially the same problem enters, however, whenever, as is usually the case, the data for prices and quantities with which we start are averages instead of being the original market quotations. Throughout this book, ‘the price’ of any commodity or ‘the quantity’ of it for any one year was assumed given. But what is such a price or quantity? Sometimes it is a single quotation for January 1 or July 1, but usually it is an average of several quotations scattered through the year. The question arises: On what principle should this average be constructed? The *practical* answer is *any* kind of average since, ordinarily, the variations during a year, so far, at least, as prices are concerned, are too little to make any perceptible difference in the result, whatever kind of average is used. Otherwise, there would be ground for subdividing the year into quarters or months until we reach a small enough period to be considered practically a point.” Irving Fisher (1922; 318)[188].

not be possible for production units to provide complete accounting information for periods shorter than a quarter. As the accounting period becomes shorter, it is less likely that production, shipment, billing and payment for the same commodity will all coincide within the accounting period. Also as the accounting period becomes shorter, work in progress will tend to become ever more important relative to final sales, creating difficult valuation problems.\*<sup>7</sup> Put another way, more and more inputs will shift from being intermediate inputs (inputs that are used up within the accounting period) to being durable inputs (inputs whose contribution to production extends over more than one period). In addition to these difficulties, there are others. For example, as the accounting period becomes shorter, transactions tend to become more erratic and sporadic. Many goods will not be sold in a supermarket in a particular day or week. Normal index number theory breaks down under these conditions: it is difficult to compare a positive amount of a good sold in one period with a zero amount sold in the next period.

A related difficulty is that many commodities are produced or demanded on a *seasonal* basis. If the accounting period is a year, then there are no seasonal commodity difficulties but as we shorten the period from a year, we will run into the problem of seasonal fluctuations in prices and quantities. In many cases, a seasonal commodity will not be available in all seasons and we again run into the problem of comparing positive values with zero values in the periods when the commodity is out of season. Even if the seasonal commodity does not disappear, the application of standard index number theory is not straightforward.\*<sup>8</sup>

Nevertheless, even in the face of the above difficulties, it seems that the only possible solution to the artificial implicit weighting problem that is generated by high inflation is to shorten the accounting period so that normal index number theory can be applied in order to construct meaningful economic aggregates.\*<sup>9</sup>

In addition to the above general problems associated with economic measurement of flow variables under conditions of high inflation, there are some additional problems associated with the measurement of capital. These additional problems are associated with the stock and flow aspects of capital. We will conclude this section by explaining these problems.

Given an accounting period of some predetermined length, we can associate with it at least three separate points in time:

- The beginning of the accounting period;
- The middle of the accounting period; and
- The end of the accounting period.

In interpreting the national accounts or the accounts of a business unit, we generally think of all flow variables as being concentrated in the middle of the period. If we follow this convention in the context of high inflation, then we require one (nominal) interest rate to index the value of money or financial capital going from the beginning of the period to the middle of the period and we require another (nominal) interest rate to index the value of money going from the middle of the period to the end of the period. Given these two interest rates, we could construct *centered user costs of capital* for each type of reproducible capital, which would be the appropriate flow variables that would match up with the other flow variables in the production accounts of the production unit.

---

\*<sup>7</sup> There are very few price indexes for work in progress! This is to be expected since there are very few transactions involving partially completed products.

\*<sup>8</sup> Hadar and Peleg (1998; 5)[205] comment on the importance of seasonal adjustment procedures in the context of high inflation: "As a by-product of the emphasis on quarterly estimates at constant prices the seasonal adjustment got large attention and many resources were spent to improve the adjustment." Diewert (1996)[114] (1998)[117] (1999)[118] reviews possible approaches to the problems involved in treating seasonal commodities (and suggests solutions) when there is high inflation.

\*<sup>9</sup> Our discussion in the previous paragraph indicates that this cannot be done if the economy is experiencing a hyperinflation. Thus meaningful economic measurement becomes impossible under very high inflation. This is a hidden cost of inflation that is not discussed very much in the literature on the costs of inflation.

However, in order to reduce notational complexity, we do not construct centered user costs in what follows. Instead, for each type of asset, we construct either a *beginning of the period user cost* (which measures the cost of using the asset for the period under consideration from the perspective of the price level prevailing at the beginning of the period) or an *end of the period user cost* (which measures the cost of using the asset for the period under consideration from the perspective of the price level prevailing at the end of the period). Of course, armed with a knowledge of the appropriate half period interest rates, it is easy to convert these “bookend” user costs into centered user costs.

In the following section, we explain the fundamental equations relating stocks and flows of capital.

### 8.3 The Fundamental Equations Relating Stocks and Flows of Capital

Before we begin with our algebra, it seems appropriate to explain why accounting for the contribution of capital to production is more difficult than accounting for the contributions of labour or materials. The main problem is that when a reproducible capital input is purchased for use by a production unit at the beginning of an accounting period, we cannot simply charge the entire purchase cost to the period of purchase. Since the benefits of using the capital asset extend over more than one period, the initial purchase cost must be distributed somehow over the useful life of the asset. This is *the fundamental problem of accounting*.<sup>\*10</sup> Hulten (1990)[244] explains the consequences for accountants of the durability of capital as follows:

“Durability means that a capital good is productive for two or more time periods, and this, in turn, implies that a distinction must be made between the value of using or renting capital in any year and the value of owning the capital asset. This distinction would not necessarily lead to a measurement problem if the capital services used in any given year were paid for in that year; that is, if all capital were rented. In this case, transactions in the rental market would fix the price and quantity of capital in each time period, much as data on the price and quantity of labor services are derived from labor market transactions. But, unfortunately, much capital is utilized by its owner and the transfer of capital services between owner and user results in an implicit rent typically not observed by the statistician. Market data are thus inadequate for the task of directly estimating the price and quantity of capital services, and this has led to the development of indirect procedures for inferring the quantity of capital, like the perpetual inventory method, or to the acceptance of flawed measures, like book value.” Charles R. Hulten (1990; 120-121)[244].

<sup>\*10</sup> “The difficulty of imputing expenses to individual sales or even to the gross earnings of the accounting period, the month or year, is an ever present problem for the accountant in the periodic determination of enterprise income. The longer the period for which the income is to be determined, the smaller the relative amount of error. Absolute accuracy can be attained only when the venture is completed and the enterprise terminated.” William T Crandell (1935; 388-389)[61].

“Early enterprises and partners working in the main in isolated trading ventures, needed only an irregular determination of profit. But before the business corporation had been very long in operation it was evident that it needed to be treated as a continuing enterprise. For example, calculating dividends by separate voyages was found impractical in the East India Company by 1660. Profit calculation therefore became a matter of periodic estimates in place of the known results of completed ventures.” A.C. Littleton (1933; 270)[294].

“The third convention is that of the annual accounting period. It is this convention which is responsible for most of the difficult accounting problems. Without this convention, accounting would be a simple matter of recording completed and fully realized transactions: an act of primitive simplicity.” Stephen Gilman (1939; 26)[198].

“All the problems of income measurement are the result of our desire to attribute income to arbitrarily determined short periods of time. Everything comes right in the end; but by then it is too late to matter.” David Solomons (1961; 378)[362]. Note that these authors do not mention the additional complications that are due to the fact that future revenues and costs must be discounted to yield values that are equivalent to present dollars.

The value of an asset at the beginning of an accounting period is equal to the discounted stream of future rental payments that the asset is expected to yield. Thus the *stock value* of the asset is equal to the discounted future *service flows*<sup>\*11</sup> that the asset is expected to yield in future periods. Let the price of a new capital input purchased at the beginning of period  $t$  be  $P_0^t$ . In a noninflationary environment, it can be assumed that the (potentially observable) sequence of (cross sectional) vintage rental prices prevailing at the beginning of period  $t$  can be expected to prevail in future periods. Thus in this no general inflation case, there is no need to have a separate notation for future expected rental prices for a new asset as it ages. However, in an inflationary environment, it is necessary to distinguish between the observable rental prices for the asset at different ages at the beginning of period  $t$  and future *expected* rental prices for assets of various ages.<sup>\*12</sup> Thus let  $f_0^t$  be the (observable) rental price of a new asset at the beginning of period  $t$ , let  $f_1^t$  be the (observable) rental price of a one period old asset at the beginning of period  $t$ , let  $f_2^t$  be the (observable) rental price of a 2 period old asset at the beginning of period  $t$ , etc. Then the *fundamental equation* relating the *stock value of a new asset* at the beginning of period  $t$ ,  $P_0^t$ , to the sequence of *cross sectional rental prices for assets of age  $n$*  prevailing at the beginning of period  $t$ ,  $\{f_n^t : n = 0, 1, 2, \dots\}$  is<sup>\*13</sup>:

$$P_0^t = f_0^t + [(1 + i_1^t)/(1 + r_1^t)]f_1^t + [(1 + i_1^t)(1 + i_2^t)/(1 + r_1^t)(1 + r_2^t)]f_2^t + \dots \quad (8.1)$$

In the above equation,<sup>\*14</sup>  $1 + i_1^t$  is the *rental price escalation factor* that is *expected* to apply to a one period old asset going from the beginning of period  $t$  to the end of period  $t$  (or equivalently, to the beginning of period  $t + 1$ ),  $(1 + i_1^t)(1 + i_2^t)$  is the *rental price escalation factor* that is *expected* to apply to a 2 period old asset going from the beginning of period  $t$  to the beginning of period  $t + 2$ , etc. Thus the  $i_n^t$  are *expected rates of price change for used assets of varying ages  $n$*  that are formed at the beginning of period  $t$ . The term  $1 + r_1^t$  is the discount factor that makes a dollar received at the beginning of period  $t$  equivalent to a dollar received at the beginning of period  $t + 1$ , the term  $(1 + r_1^t)(1 + r_2^t)$  is the discount factor that makes a dollar received at the beginning of period  $t$  equivalent to a dollar received at the beginning of period  $t + 2$ , etc. Thus the  $r_n^t$  are one period *nominal interest rates* that represent the *term structure of interest rates* at the beginning of period  $t$ .<sup>\*15</sup>

<sup>\*11</sup> Walras (1954)[387] (first edition published in 1874) was one of the earliest economists to state that capital stocks are demanded because of the future flow of services that they render. Although he was perhaps the first economist to formally derive a user cost formula as we shall see, he did not work out the explicit discounting formula that Böhm-Bawerk (1891; 342)[40] was able to derive.

<sup>\*12</sup> Note that these future expected rental prices are not generally observable due to the lack of futures markets for these future period rentals of the assets of varying ages.

<sup>\*13</sup> The sequence of (cross sectional) vintage rental prices  $\{f_n^t\}$  is called the *age-efficiency profile* of the asset.

<sup>\*14</sup> It should be noted that Irving Fisher (1897; 365)[183] seemed to be well aware of the complexities that are imbedded in equation (8.1): "There is not space here to discuss the theory in greater detail, nor to apply it to economic problems. A full treatment would take account of the various standards in which income is or may be expressed, of the case in which the rates of interest at different dates and for different periods does not remain constant, of the fact that the services of capital which are discounted in its value are only *expected* services, not those which actually materialise, and of the consequent discrepancy between income anticipated and income realised, of the propriety or impropriety of including man himself as a species of income-bearing capital, and so on."

<sup>\*15</sup> Peter Hill has noted a major problem with the use of equation (8.1) as the starting point of our discussion: namely, *unique assets* will by definition not have used versions of the same asset in the marketplace during the current period and so the cross sectional rental prices  $f_n^t$  for assets of age  $n$  in period  $t$  will not exist for these assets! In this case, the  $f_n^t$  should be interpreted as expected future rentals that the unique asset is expected to generate at today's prices. The  $(1 + i_n^t)$  terms then summarize expectations about the amount of asset specific price change that is expected to take place. This reinterpretation of equation (8.1) is more fundamental but we chose not to make it our starting point because it does not lead to a completely objective method for national statisticians to form reproducible estimates of these future rental payments. However, in many situations (e.g., the valuation of a new movie), the statistician will be forced to attempt to implement Hill's (2000)[230] more general model.

We now generalize equation (8.1) to relate the *stock value of an  $n$  period old asset* at the beginning of period  $t$ ,  $P_n^t$ , to the sequence of *cross sectional vintage rental prices* prevailing at the beginning of period  $t$ ,  $\{f_n^t\}$ ; thus for  $n = 0, 1, 2, \dots$ , we assume:

$$P_n^t = f_n^t + [(1 + i_1^t)/(1 + r_1^t)]f_{n+1}^t + [(1 + i_1^t)(1 + i_2^t)/(1 + r_1^t)(1 + r_2^t)]f_{n+2}^t + \dots \quad (8.2)$$

Thus older assets discount fewer terms in the above sum; i.e., as  $n$  increases by one, we have one less term on the right hand side of (8.2). However, note that we are applying the same price escalation factors  $(1 + i_1^t), (1 + i_1^t)(1 + i_2^t), \dots$ , to escalate the cross sectional rental prices prevailing at the beginning of period  $t$ ,  $f_1^t, f_2^t, \dots$ , and to form estimates of future expected rental prices for each vintage of the capital stock that is in use at the beginning of period  $t$ .

The rental prices prevailing at the beginning of period  $t$  for assets of various ages,  $f_0^t, f_1^t, \dots$  are potentially observable.\*<sup>16</sup> These cross section rental prices reflect the relative efficiency of the various vintages of the capital good that are still in use at the beginning of period  $t$ . It is assumed that these rentals are paid (explicitly or implicitly) by the users at the beginning of period  $t$ . Note that the sequence of asset stock prices for various ages at the beginning of period  $t$ ,  $P_0^t, P_1^t, \dots$  is not affected by general inflation provided that the general inflation affects the expected asset rates of price change  $i_n^t$  and the nominal interest rates  $r_n^t$  in a proportional manner. We will return to this point later.

The physical productivity characteristics of a unit of capital of each age are determined by the sequence of cross sectional rental prices. Thus a brand new asset is characterized by the vector of current rental prices for assets of various ages,  $f_0^t, f_1^t, f_2^t, \dots$ , which are interpreted as “physical” contributions to output that the new asset is expected to yield during the current period  $t$  (this is  $f_0^t$ ), the next period (this is  $f_1^t$ ), and so on. An asset which is one period old at the start of period  $t$  is characterized by the vector  $f_1^t, f_2^t, \dots$ , etc.\*<sup>17</sup>

We have not explained how the expected rental price rates of price change  $i_n^t$  are to be estimated. We shall deal with this problem in section 8.5 below. However, it should be noted that there is no guarantee that our expectations about the future course of rental prices are correct.

At this point, we make some simplifying assumptions about the expected rates of rental price change for future periods  $i_n^t$  and the interest rates  $r_n^t$ . We assume that these anticipated specific price change escalation factors at the beginning of each period  $t$  are all equal; i.e., we assume:

$$i_n^t = i^t; \quad n = 1, 2, \dots \quad (8.3)$$

We also assume that the term structure of (nominal) interest rates at the beginning of each period  $t$  is constant; i.e., we assume:

$$r_n^t = r^t; \quad n = 1, 2, \dots \quad (8.4)$$

However, note that as the period  $t$  changes,  $r^t$  and  $i^t$  can change.

Using assumptions (8.3) and (8.4), we can rewrite the system of equations (8.2), which relate the sequence or profile of *stock prices* of age  $n$  at the beginning of period  $t$   $\{P_n^t\}$  to the sequence or profile

\*<sup>16</sup> This is the main reason that we use this escalation of cross sectional rental prices approach to capital measurement rather than the more fundamental discounted future expected rentals approach advocated by Hill.

\*<sup>17</sup> Triplett (1996; 97)[375] used this characterization for capital assets of various vintages.

of (cross sectional) *rental prices* for assets of age  $n$  at the beginning of period  $t$   $\{f_n^t\}$ , as follows:

$$\begin{aligned} P_0^t &= f_0^t + [(1+i^t)/(1+r^t)]f_1^t + [(1+i^t)/(1+r^t)]^2 f_2^t + [(1+i^t)/(1+r^t)]^3 f_3^t + \dots \\ P_1^t &= f_1^t + [(1+i^t)/(1+r^t)]f_2^t + [(1+i^t)/(1+r^t)]^2 f_3^t + [(1+i^t)/(1+r^t)]^3 f_4^t + \dots \\ P_2^t &= f_2^t + [(1+i^t)/(1+r^t)]f_3^t + [(1+i^t)/(1+r^t)]^2 f_4^t + [(1+i^t)/(1+r^t)]^3 f_5^t + \dots \\ &\dots \\ P_n^t &= f_n^t + [(1+i^t)/(1+r^t)]f_{n+1}^t + [(1+i^t)/(1+r^t)]^2 f_{n+2}^t + [(1+i^t)/(1+r^t)]^3 f_{n+3}^t + \dots \end{aligned} \quad (8.5)$$

On the left hand side of equations (8.5), we have the sequence of period  $t$  asset prices by age starting with the price of a new asset,  $P_0^t$ , moving to the price of an asset that is one period old at the start of period  $t$ ,  $P_1^t$ , then moving to the price of an asset that is 2 periods old at the start of period  $t$ ,  $P_2^t$ , and so on. On the right hand side of equations (8.5), the first term in each equation is a member of the sequence of rental prices by age of asset that prevails in the market (if such markets exist) at the beginning of period  $t$ . Thus  $f_0^t$  is the rent for a new asset,  $f_1^t$  is the rent for an asset that is one period old at the beginning of period  $t$ ,  $f_2^t$  is the rent for an asset that is 2 periods old, and so on. This sequence of current market rental prices for the assets of various vintages is then extrapolated out into the future using the anticipated price escalation rates  $(1+i^t)$ ,  $(1+i^t)^2$ ,  $(1+i^t)^3$ , etc. and then these future expected rentals are discounted back to the beginning of period  $t$  using the nominal discount factors  $(1+r^t)$ ,  $(1+r^t)^2$ ,  $(1+r^t)^3$ , etc. Note that given the period  $t$  expected asset inflation rate  $i^t$  and the period  $t$  nominal discount rate  $r^t$ , we can go from the (cross sectional) sequence of vintage rental prices  $\{f_n^t\}$  to the (cross sectional) sequence of vintage asset prices  $\{P_n^t\}$  using equations (8.5). We shall show below how this procedure can be reversed; i.e., we shall show how given the sequence of cross sectional asset prices, we can construct estimates for the sequence of cross sectional rental prices.

It seems that Böhm-Bawerk was the first economist to use the above method for relating the future service flows of a durable input to its stock price:

“If the services of the durable good be exhausted in a short space of time, the individual services, provided that they are of the same quality— which, for simplicity’s sake, we assume— are, as a rule, equal in value, and the value of the material good itself is obtained by multiplying the value of one service by the number of services of which the good is capable. But in the case of many durable goods, such as ships, machinery, furniture, land, the services rendered extend over long periods, and the result is that the later services cannot be rendered, or at least cannot be rendered in a normal economic way, before a long time has expired. As a consequence, the value of the more distant material services suffers the same fate as the value of future goods. A material service, which, technically, is exactly the same as a service of this year, but which cannot be rendered before next year, is worth a little less than this year’s service; another similar service, but obtainable only after two years, is, again, a little less valuable, and so on; the values of the remote services decreasing with the remoteness of the period at which they can be rendered. Say that this year’s service is worth 100, then next year’s service— assuming a difference of 5 % per annum— is worth in today’s valuation only 95.23; the third year’s service is worth only 90.70; the fourth year’s service, 86.38; the fifth, sixth and seventh year’s services, respectively, worth 82.27, 78.35, 74.62 of present money. The value of the durable good in this case is not found by multiplying the value of the current service by the total number of services, but is represented by a sum of services decreasing in value.” Eugen von Böhm-Bawerk (1891; 342)[40].

Böhm-Bawerk (1891; 342)[40] considered a special case of (8.5) where all service flows  $f_n$  were equal to 100 for  $n = 0, 1, \dots, 6$  and equal to 0 thereafter, where the asset inflation rate was expected to be

0 and where the interest rate  $r$  was equal to .05 or 5 %.\*<sup>18</sup> This is a special case of what has come to be known as the *one hoss shay model* and we shall consider it in more detail in section 8.7.

Note that equations (8.5) can be rewritten as follows: \*<sup>19</sup>

$$\begin{aligned} P_0^t &= f_0^t + [(1+i^t)/(1+r^t)]P_1^t; \\ P_1^t &= f_1^t + [(1+i^t)/(1+r^t)]P_2^t; \\ P_2^t &= f_2^t + [(1+i^t)/(1+r^t)]P_3^t; \\ &\dots \\ P_n^t &= f_n^t + [(1+i^t)/(1+r^t)]P_{n+1}^t; \dots \end{aligned} \quad (8.6)$$

The first equation in (8.6) says that the value of a new asset at the start of period  $t$ ,  $P_0^t$ , is equal to the rental that the asset can earn in period  $t$ ,  $f_0^t$ , \*<sup>20</sup> plus the expected asset value of the capital good at the end of period  $t$ ,  $(1+i^t)P_1^t$ , but this expected asset value must be divided by the discount factor,  $(1+r^t)$ , in order to convert this future value into an equivalent beginning of period  $t$  value. \*<sup>21</sup>

Now it is straightforward to solve equations (8.6) for the sequence of period  $t$  cross sectional rental prices,  $\{f_n^t\}$ , in terms of the cross sectional asset prices,  $\{P_n^t\}$ :

$$\begin{aligned} f_0^t &= P_0^t - [(1+i^t)/(1+r^t)]P_1^t = (1+r^t)^{-1}[P_0^t(1+r^t) - (1+i^t)P_1^t] \\ f_1^t &= P_1^t - [(1+i^t)/(1+r^t)]P_2^t = (1+r^t)^{-1}[P_1^t(1+r^t) - (1+i^t)P_2^t] \\ f_2^t &= P_2^t - [(1+i^t)/(1+r^t)]P_3^t = (1+r^t)^{-1}[P_2^t(1+r^t) - (1+i^t)P_3^t] \\ &\dots \\ f_n^t &= P_n^t - [(1+i^t)/(1+r^t)]P_{n+1}^t = (1+r^t)^{-1}[P_n^t(1+r^t) - (1+i^t)P_{n+1}^t]; \dots \end{aligned} \quad (8.7)$$

Thus equations (8.5) allow us to go from the sequence of rental prices by age  $n$   $\{f_n^t\}$  to the sequence of asset prices by age  $n$   $\{P_n^t\}$  while equations (8.7) allow us to reverse the process.

Equations (8.7) can be derived from elementary economic considerations. Consider the first equation in (8.7). Think of a production unit as purchasing a unit of the new capital asset at the beginning of period  $t$  at a cost of  $P_0^t$  and then using the asset throughout period  $t$ . However, at the end of period  $t$ , the producer will have a depreciated asset that is expected to be worth  $(1+i^t)P_1^t$ . Since this offset to the initial cost of the asset will only be received at the end of period  $t$ , it must be divided by  $(1+r^t)$  to express the benefit in terms of beginning of period  $t$  dollars. Thus the expected net cost of *using* the new asset for period  $t$  \*<sup>22</sup> is  $P_0^t - [(1+i^t)/(1+r^t)]P_1^t$ .

The above equations assume that the actual or implicit period  $t$  rental payments  $f_n^t$  for assets of different ages  $n$  are made at the *beginning* of period  $t$ . It is sometimes convenient to assume that the

\*<sup>18</sup> Böhm-Bawerk (1891; 343)[40] went on and constructed the sequence of vintage asset prices using his special case of equations (8.5).

\*<sup>19</sup> Christensen and Jorgenson (1969; 302)[52] do this for the geometric depreciation model except that they assume that the rental is paid at the end of the period rather than the beginning. Variants of the system of equations (8.6) were derived by Christensen and Jorgenson (1973)[54], Jorgenson (1989; 10)[255], Hulten (1990; 128)[244] and Diewert and Lawrence (2000; 276)[142]. Irving Fisher (1908; 32-33)[184] also derived these equations in words.

\*<sup>20</sup> Note that we are implicitly assuming that the rental is paid to the owner at the beginning of period  $t$ .

\*<sup>21</sup> Another way of interpreting say the first equation in (8.6) runs as follows: the purchase cost of a new asset  $P_0^t$  less the rental  $f_0^t$  (which is paid immediately at the beginning of period  $t$ ) can be regarded as an investment, which must earn the going rate of return  $r^t$ . Thus we must have  $[P_0^t - f_0^t](1+r^t) = (1+i^t)P_1^t$  which is the (expected) value of the asset at the end of period  $t$ . This line of reasoning can be traced back to Walras (1954; 267)[387].

\*<sup>22</sup> This explains why the rental prices  $f_n^t$  are sometimes called *user costs*. This derivation of a user cost was used by Diewert (1974; 504)[80], (1980; 472-473)[89], (1992a; 194)[104] and by Hulten (1996; 155)[245].

rental payments are made at the *end* of each accounting period. Thus we define the *end of period  $t$  rental price or user cost* for an asset that is  $n$  periods old at the beginning of period  $t$ ,  $u_n^t$ , in terms of the corresponding *beginning of period  $t$  rental price*  $f_n^t$  as follows:

$$u_n^t \equiv (1 + r^t)f_n^t; \quad n = 0, 1, 2, \dots \quad (8.8)$$

Thus if the rental payment is made at the end of the period instead of the beginning, then the beginning of the period rental  $f_n^t$  must be escalated by the interest rate factor  $(1 + r^t)$  in order to obtain the end of the period user cost  $u_n^t$ .<sup>\*23</sup>

Using equations (8.8) and the second set of equations in (8.7), it can readily be shown that the sequence of end of period  $t$  user costs  $\{u_n^t\}$  can be defined in terms of the period  $t$  sequence of asset prices by age  $\{P_n^t\}$  as follows:

$$\begin{aligned} u_0^t &= P_0^t(1 + r^t) - (1 + i^t)P_1^t \\ u_1^t &= P_1^t(1 + r^t) - (1 + i^t)P_2^t \\ u_2^t &= P_2^t(1 + r^t) - (1 + i^t)P_3^t \\ &\dots \\ u_n^t &= P_n^t(1 + r^t) - (1 + i^t)P_{n+1}^t; \dots \end{aligned} \quad (8.9)$$

Equations (8.9) can also be given a direct economic interpretation. Consider the following explanation for the user cost for a new asset,  $u_0^t$ . At the end of period  $t$ , the business unit expects to have an asset worth  $(1 + i^t)P_1^t$ . Offsetting this benefit is the beginning of the period asset purchase cost,  $P_0^t$ . However, in addition to this cost, the business must charge itself either the explicit interest cost that occurs if money is borrowed to purchase the asset or the implicit opportunity cost of the equity capital that is tied up in the purchase. Thus offsetting the end of the period benefit  $(1 + i^t)P_1^t$  is the initial purchase cost and opportunity interest cost of the asset purchase,  $P_0^t(1 + r^t)$ , leading to an end of period  $t$  net cost of  $P_0^t(1 + r^t) - (1 + i^t)P_1^t$  or  $u_0^t$ .

It is interesting to note that in both the accounting and financial management literature of the past century, there was a reluctance to treat the opportunity cost of *equity capital* tied up in capital inputs as a genuine cost of production.<sup>\*24</sup> However, more recently, there is an acceptance of an imputed interest charge for equity capital as a genuine cost of production.<sup>\*25</sup>

In the following section, we will relate the asset price profiles  $\{P_n^t\}$  and the user cost profiles  $\{u_n^t\}$  to *depreciation profiles*. However, before turning to the subject of depreciation, it is important to stress that the analysis presented in this section is based on a number of restrictive assumptions, particularly on future price expectations. Moreover, we have not explained how these asset price expectations are formed and we have not explained how the period  $t$  nominal interest rate is to be estimated (we will address these topics in section 8.5 below). We have not explained what should be done if the sequence of second hand asset prices  $\{P_n^t\}$  is not available and the sequences of vintage rental prices or user costs,  $\{f_n^t\}$  or  $\{u_n^t\}$ , are also not available (we will address this problem in later sections as well). We have also assumed that asset values and user costs are independent of how

<sup>\*23</sup> It is interesting that Böhm-Bawerk (1891; 343)[40] carefully distinguished between rental payments made at the beginning or end of a period: "These figures are based on the assumption that the whole year's utility is obtained all at once, and, indeed, obtained in anticipation at the beginning of the year; e.g., by hiring the good at a year's interest of 100 payable on each 1st January. If, on the other hand, the year's use can only be had at the end of the year, a valuation undertaken at the beginning of the year will show figures not inconsiderably lower. . . . That the figures should alter according as the date of the valuation stands nearer or farther from the date of obtaining the utility, is an entirely natural thing, and one quite familiar in financial life."

<sup>\*24</sup> This literature is reviewed in Diewert and Fox (1999; 271-274)[137].

<sup>\*25</sup> Stern Stewart & Co. has popularized the idea of charging for the opportunity cost of equity capital and has called the resulting income concept, EVA, Economic Value Added.

intensively the assets are used. Finally, we have not modeled uncertainty (about future prices and the useful lives of assets) and attitudes towards risk on the part of producers. Thus the analysis presented in this chapter is only a start on the difficult problems associated with measuring capital input.

## 8.4 Cross Sectional Depreciation Profiles

Recall that in the previous section,  $P_n^t$  was defined to be the price of an asset that was  $n$  periods old at the beginning of period  $t$ . Generally, the decline in asset value as we go from one age to the next older age is called *depreciation*. More precisely, we define the *cross sectional depreciation*  $D_n^t$ <sup>\*26</sup> of an asset that is  $n$  periods old at the beginning of period  $t$  as

$$D_n^t \equiv P_n^t - P_{n+1}^t; \quad n = 0, 1, 2, \dots \quad (8.10)$$

Thus  $D_n^t$  is the value of an asset that is  $n$  periods old at the beginning of period  $t$ ,  $P_n^t$ , minus the value of an asset that is  $n + 1$  periods old at the beginning of period  $t$ ,  $P_{n+1}^t$ .<sup>\*27</sup>

Obviously, given the sequence of period  $t$  used asset prices  $\{P_n^t\}$ , we can use equations (8.10) to determine the period  $t$  sequence of declines in asset values by age  $n$ ,  $\{D_n^t\}$ . Conversely, given the period  $t$  cross sectional depreciation sequence or *profile*,  $\{D_n^t\}$ , we can determine the period  $t$  asset prices by age by adding up amounts of depreciation:

$$\begin{aligned} P_0^t &= D_0^t + D_1^t + D_2^t + \dots \\ P_1^t &= D_1^t + D_2^t + D_3^t + \dots \\ &\dots \\ P_n^t &= D_n^t + D_{n+1}^t + D_{n+2}^t + \dots \end{aligned} \quad (8.11)$$

Rather than working with first differences of asset prices by age, it is more convenient to reparameterize the pattern of cross sectional depreciation by defining the *period  $t$  depreciation rate*  $\delta_n^t$  for an asset that is  $n$  periods old at the start of period  $t$  as follows:

$$\delta_n^t \equiv 1 - [P_{n+1}^t / P_n^t] = D_n^t / P_n^t; \quad n = 0, 1, 2, \dots \quad (8.12)$$

In the above definitions, we require  $n$  to be such that  $P_n^t$  is positive.<sup>\*28</sup>

<sup>\*26</sup> This terminology is due to Hill (1999)[229] who used this terminology to distinguish the decline in second hand asset values due to aging (cross sectional depreciation) from the decline in an asset value over a period of time (time series depreciation). Triplett (1996; 98-99)[375] uses the cross sectional definition of depreciation and shows that it is equal to the concept of capital consumption in the national accounts but he does this under the assumption of no expected real asset inflation.

<sup>\*27</sup> Of course, the objections to the use of second hand market data to determine depreciation rates are very old: "We readily agree that where a market is sufficiently large, generally accessible, and continuous over time, it serves to coordinate a large number of subjective estimates and thus may impart a moment of (social) objectivity to value relations based on prices forced on it. But it can hardly be said that the second-hand market for industrial equipment, which would be the proper place for the determination of the value of capital goods which have been in use, satisfies these requirements, and that its valuations are superior to intra-enterprise valuation." L.M. Lachmann (1941; 376-377)[284]. "Criticism has also been voiced about the viability of used asset market price data as an indicator of in use asset values. One argument, drawing on the Akerlof Lemons Model, is that assets resold in second hand markets are not representative of the underlying population of assets, because only poorer quality units are sold when used. Others express concerns about the thinness of resale markets, believing that it is sporadic in nature and is dominated by dealers who under-bid." Charles R. Hulten and Frank C. Wykoff (1996; 17-18)[248].

<sup>\*28</sup> This definition of depreciation dates back to Hicks (1939)[219] at least and was used extensively by Hulten and Wykoff (1981a)[246] (1981b)[247], Diewert (1974; 504)[80] and Hulten (1990; 128)[244] (1996; 155)[245]: "If there is a perfect second hand market for the goods in question, so that a market value can be assessed for them

Obviously, given the sequence of period  $t$  asset prices by age  $n$ ,  $\{P_n^t\}$ , we can use equations (8.12) to determine the period  $t$  sequence of *cross sectional depreciation rates* by age,  $\{\delta_n^t\}$ . Conversely, given the cross sectional sequence of period  $t$  depreciation rates,  $\{\delta_n^t\}$ , as well as the price of a new asset in period  $t$ ,  $P_0^t$ , we can determine the period  $t$  asset prices by age as follows:

$$\begin{aligned} P_1^t &= (1 - \delta_0^t)P_0^t \\ P_2^t &= (1 - \delta_0^t)(1 - \delta_1^t)P_0^t \\ &\dots \\ P_n^t &= (1 - \delta_0^t)(1 - \delta_1^t) \cdots (1 - \delta_{n-1}^t)P_0^t; \dots \end{aligned} \quad (8.13)$$

The interpretation of equations (8.13) is straightforward. At the beginning of period  $t$ , a new capital good is worth  $P_0^t$ . An asset of the same type but which is one period older at the beginning of period  $t$  is less valuable by the amount of depreciation  $\delta_0^t P_0^t$  and hence is worth  $(1 - \delta_0^t)P_0^t$ , which is equal to  $P_1^t$ . An asset which is two periods old at the beginning of period  $t$  is less valuable than a one period old asset by the amount of depreciation  $\delta_1^t P_1^t$  and hence is worth  $P_2^t = (1 - \delta_1^t)P_1^t$  which is equal to  $(1 - \delta_1^t)(1 - \delta_0^t)P_0^t$  using the first equation in (8.13) and so on. Suppose  $L - 1$  is the first integer which is such that  $\delta_{L-1}^t$  is equal to one. Then  $P_n^t$  equals zero for all  $n \geq L$ ; i.e., at the end of  $L$  periods of use, the asset no longer has a positive rental value. If  $L = 1$ , then a new asset of this type delivers all of its services in the first period of use and the asset is in fact a nondurable asset.

Now substitute equations (8.12) into equations (8.9) in order to obtain the following formulae for the sequence of the *end of the period user costs* by age  $n$   $\{u_n^t\}$  in terms of the price of a new asset at the beginning of period  $t$ ,  $P_0^t$ , and the sequence of cross sectional depreciation rates,  $\{\delta_n^t\}$ :

$$\begin{aligned} u_0^t &= [(1 + r^t) - (1 + i^t)(1 - \delta_0^t)]P_0^t \\ u_1^t &= (1 - \delta_0^t)[(1 + r^t) - (1 + i^t)(1 - \delta_1^t)]P_0^t \\ &\dots \\ u_n^t &= (1 - \delta_0^t) \cdots (1 - \delta_{n-1}^t)[(1 + r^t) - (1 + i^t)(1 - \delta_n^t)]P_0^t; \dots \end{aligned} \quad (8.14)$$

Thus given  $P_0^t$  (the beginning of period  $t$  price of a new asset),  $i^t$  (the new asset inflation rate that is expected at the beginning of period  $t$ ),  $r^t$  (the one period nominal interest rate that the business unit faces at the beginning of period  $t$ ) and given the sequence of cross sectional depreciation rates by age prevailing at the beginning of period  $t$  (the  $\delta_n^t$ ), then we can use equations (8.14) to calculate the sequence of end of the period user costs for period  $t$ , the  $u_n^t$ . Of course, given the  $u_n^t$ , we can use equations (8.8) to calculate the beginning of the period user costs (the  $f_n^t$ ) and then use the  $f_n^t$  to calculate the sequence of asset prices  $P_n^t$  using equations (8.5) and finally, given the  $P_n^t$ , we can use equations (8.12) in order to calculate the sequence of depreciation rates, the  $\delta_n^t$ . Thus *given any one of these sequences or profiles, all of the other sequences are completely determined*. This means that assumptions about depreciation rates, the pattern of user costs by age or the pattern of asset prices by age *cannot be made independently of each other*.<sup>\*29</sup>

---

with precision, corresponding to each particular degree of wear, then the value-loss due to consumption can be exactly measured. . .” John R. Hicks (1939; 176)[219]. Current cost accountants have also advocated the use of second hand market data (when available) to calculate “objective” depreciation rates: “But as a practical matter the quantification and valuation of asset services used is not a simple matter and we must fall back on estimated patterns as a basis for current cost as well as historic cost depreciation. For those fixed assets which have active second hand markets the problem is not overly difficult. A pattern of service values can be obtained at any time by comparing the market values of different ages or degrees of use. The differences so obtained, when related to the value of a new asset, yield the proportions of asset value which are normally used up or foregone in the various stages of asset life.” Edgar O. Edwards and Philip W. Bell (1961; 175)[168].

<sup>\*29</sup> This point was first made explicitly by Jorgenson and Griliches (1967; 257)[258]: “An almost universal conceptual error in the measurement of capital input is to confuse the aggregation of capital stock with the aggregation

It is useful to look more closely at the first equation in (8.14), which expresses the user cost or rental price of a new asset at the beginning of period  $t$ ,  $u_0^t$ , in terms of the depreciation rate  $\delta_0^t$ , the one period nominal interest rate  $r^t$ , the new asset inflation rate  $i^t$  that is expected to prevail at the beginning of period  $t$  and the beginning of period  $t$  price for a new asset,  $P_0^t$ :

$$u_0^t = [(1 + r^t) - (1 + i^t)(1 - \delta_0^t)]P_0^t = [r^t - i^t + (1 + i^t)\delta_0^t]P_0^t. \quad (8.15)$$

Thus the user cost of a new asset  $u_0^t$  that is purchased at the beginning of period  $t$  (and the actual or imputed rental payment is made at the end of the period) is equal to  $r^t - i^t$  (a nominal interest rate minus an asset inflation rate which can be loosely interpreted<sup>\*30</sup> as a *real interest rate*) times the initial asset cost  $P_0^t$  plus  $(1 + i^t)\delta_0^t P_0^t$  which is *depreciation* on the asset at beginning of the period prices,  $\delta_1^t P_0^t$ , times *the inflation escalation factor*,  $(1 + i^t)$ .<sup>\*31</sup> If we further assume that the expected asset inflation rate is 0, then (8.15) further simplifies to:

$$u_0^t = [r^t + \delta_0^t]P_0^t. \quad (8.16)$$

Under these assumptions, the user cost of a new asset is equal to the interest rate plus the depreciation rate times the initial purchase price.<sup>\*32</sup> This is essentially the user cost formula that was obtained by Walras in 1874:

“Let  $P$  be the price of a capital good. Let  $p$  be its gross income, that is, the price of its service inclusive of both the depreciation charge and the insurance premium. Let  $\mu P$  be the portion of this income representing the depreciation charge and  $\nu P$  the portion representing the insurance premium. What remains of the gross income after both charges have been deducted,  $\pi = p - (\mu + \nu)P$ , is the *net income*.

We are now able to explain the differences in gross incomes derived from various capital goods having the same value, or conversely, the differences in values of various capital goods yielding the same gross incomes. It is, however, readily seen that the values of capital goods are rigorously proportional to their net incomes. At least this would have to be so under certain normal and ideal conditions when the market for capital goods is in equilibrium. Under equilibrium conditions the ratio  $[p - (\mu + \nu)P]/P$ , or the rate of net income, is the same for all capital goods. Let  $i$  be this common ratio. When we determine  $i$ , we also determine the prices of all landed capital, personal capital and capital goods proper by virtue of the equation  $p - (\mu + \nu)P = iP$  or  $P = p/[i + \mu + \nu]$ .” Léon Walras (1954; 268-269)[387].

However, the basic idea that a durable input should be charged a period price that is equal to a depreciation term plus a term that would cover the cost of financial capital goes back much further<sup>\*33</sup>. For example, consider the following quotation from Babbage:

---

of capital service.” See also Jorgenson and Griliches (1972; 81-87)[260]. Much of the above algebra for switching from one method of representing vintage capital inputs to another was first developed by Christensen and Jorgenson (1969; 302-305)[52] (1973)[54] for the geometrically declining depreciation model. The general framework for an internally consistent treatment of capital services and capital stocks in a set of vintage accounts was set out by Jorgenson (1989)[255], Hulten (1990; 127-129)[244] (1996; 152-160)[245] and Diewert and Lawrence (2000)[142].

<sup>\*30</sup> We will provide a more precise definition of a real interest rate later.

<sup>\*31</sup> This formula was obtained by Christensen and Jorgenson (1969; 302)[52] for the geometric model of depreciation but it is valid for any depreciation model. Griliches (1963; 120)[203] also came very close to deriving this formula in words: “In a perfectly competitive world the annual rent of a machine would equal the marginal product of its services. The rent itself would be determined by the interest costs on the investment, the deterioration in the future productivity of the machine due to current use, and the expected change in the price of the machine (obsolescence).”

<sup>\*32</sup> Using equations (8.13) and (8.14) and the assumption that the asset inflation rate  $i^t = 0$ , it can be shown that the user cost of an asset that is  $n$  periods old at the start of period  $t$  can be written as  $u_n^t = [r^t + \delta_n^t]P_n^t$  where  $P_n^t$  is the beginning of period  $t$  second hand market price for the asset.

<sup>\*33</sup> Solomons (1968; 9-17)[363] indicates that interest was regarded as a cost for a durable input in much of the nineteenth century accounting literature. The influential book by Garcke and Fells (1893)[197] changed this.

“Machines are, in some trades, let out to hire, and a certain sum is paid for their use, in the manner of rent. This is the case amongst the frame-work knitters: and Mr. Hensen, in speaking of the rate of payment for the use of their frames, states, that the proprietor receives such a rent that, besides paying the full interest for his capital, he clears the value of his frame in nine years. When the rapidity with which improvements succeed each other is considered, this rent does not appear exorbitant. Some of these frames have been worked for thirteen years with little or no repair.” Charles Babbage (1835; 287)[16].

Babbage did not proceed further with the user cost idea. Walras seems to have been the first economist who formalized the idea of a user cost into a mathematical formula. However, the early industrial engineering literature also independently came up with the user cost idea; Church described how the use of a machine should be charged as follows:

“No sophistry is needed to assume that these charges are in the nature of rents, for it might easily happen that in a certain building a number of separate little shops were established, each containing one machine, all making some particular part or working on some particular operation of the same class of goods, but each shop occupied, not by a wage earner, but by an independent mechanic, who rented his space, power and machinery, and sold the finished product to the lessor. Now in such a case, what would be the shop charges of these mechanics? Clearly they would comprise as their chief if not their only item, just the rent paid. And this rent would be made up of: (1) Interest. (2) Depreciation. (3) Insurance. (4) Profit on the capital involved in the building, machine and power-transmitting and generating plant. There would also most probably be a separate charge for power according to the quantity consumed. Exclude the item of profit, which is not included in the case of a shop charge, and we find that we have approached most closely to the new plan of reducing any shop into its constituent production centres. No one would pretend that there was any insuperable difficulty involved in fixing a just rent for little shops let out in this plan.” A. Hamilton Church (1901; 907-908)[56].

“A production centre is, of course, either a mechanic, or a bench at which a hand craftsman works. Each of these is in the position of a little shop carrying on one little special industry, paying rent for the floor space occupied, interest for the capital involved, depreciation for the wear and tear, and so on, *quite independently of what may be paid by other production centres* in the same shop.” A. Hamilton Church (1901; 734)[56].

Church was well aware of the importance of determining the “right” rate to be charged for the use of a machine in a multiproduct enterprise. This information is required not only to price products appropriately but to determine whether an enterprise should make or purchase a particular commodity. Babbage and Canning were also aware of the importance of determining the right machine rate charge.\*<sup>34</sup>

---

\*<sup>34</sup> Under moderate inflation, the difficulties with traditional cost accounting based on historical cost and no proper allowance for the opportunity of capital, the proper pricing of products becomes very difficult. Diewert and Fox (1999; 271-274)[137] argued that this factor contributed to the great productivity slowdown that started around 1973 and persisted to the early 1990’s. The traditional method of cost accounting can be traced back to a book first published in 1887 by the English accountants, Garcke and Fells, who suggested allocating the “indirect costs” of producing a good proportionally to the amount of labour and materials costs used to make the item: “In some establishments the direct expenditures in wages and materials only is considered to constitute the cost; and no attempt is made to allocate to the various working or stock orders any portion of the indirect expenses. Under this system the difference between the sum of the wages and materials expended on the articles and their selling price constitutes the gross profit, which is carried in the aggregate to the credit of profit and loss, the indirect factory expenses already referred to, together with the establishment expenses and depreciation, being particularised on the debit side of that account. This method has certainly simplicity in its favour, but a more efficient check upon the indirect expenses would be obtained by establishing a relation between them and the direct expenses. This may be done by distributing all the indirect expenses, such as wages of foremen, rent of factory, fuel, lighting, heating, and cleaning, etc. (but not the salaries of clerks, office rent, stationery and other establishment charges to be referred to later), over the various jobs, as a percentage, either upon the

“The great competition introduced by machinery, and the application of the principle of the subdivision of labour, render it necessary for each producer to be continually on the watch, to discover improved methods by which the cost of the article he manufactures may be reduced; and, with this view, it is of great importance to know the precise expense of every process, as well as of the wear and tear of machinery which is due to it.” Charles Babbage (1835; 203)[16].

“The question of ‘adequate’ rates of depreciation, in the sense that they will ultimately adjust the valuations to the realities, is often discussed as though it had no effect upon ultimate profit at all. Of some modes of valuing, it is said that they tend to overvalue some assets and to undervalue others, but the aggregate of book values found is nearly right. If the management pay no attention at all to the unit costs implied in such valuations, no harm is done. But if the cost accountant gives effect to these individually bad valuations through a machine-rate burden charge, and if the selling policy has regard for apparent unit profits, the valuation may lead to the worst rather than to the best possible policy.” John B. Canning (1929; 259-260)[44].

The above equations relating asset prices by age  $P_n^t$ , beginning of the period user costs by age  $f_n^t$ , end of the period user costs by age  $u_n^t$  and the (cross sectional) depreciation rates by age  $\delta_n^t$  are the fundamental ones that we will specialize in subsequent sections in order to measure both wealth capital stocks and capital services under conditions of inflation. In the following section, we shall consider several options that could be used in order to determine empirically the interest rates  $r^t$  and the asset inflation rates  $i^t$ .

## 8.5 The Empirical Determination of Interest Rates and Asset Inflation Rates

We consider initially three broad approaches<sup>\*35</sup> to the determination of the nominal interest rate  $r^t$  that is to be used to discount future period value flows by the business units in the aggregate under consideration:

- Use the ex post rate of return that will just make the sum of the user costs exhaust the gross operating surplus of the production sectors in the aggregate under consideration.

---

wages expended upon the jobs respectively, or upon the cost of both wages and materials.” Emile Garcke and John Manger Fells (1893; 70-71)[197]. Compare this rather crude approach to cost accounting to the masterful analysis of Church! Garcke and Fells endorsed the idea that depreciation was an admissible item of cost that should be allocated in proportion to the prime cost (i.e., labour and materials cost) of manufacturing an article but they explicitly ruled out interest as a cost: “The item of Depreciation may, for the purpose of taking out the cost, simply be included in the category of the indirect expenses of the factory, and be distributed over the various enterprises in the same way as those expenses may be allocated; or it may be dealt with separately and more correctly in the manner already alluded to and hereafter to be fully described. The establishment expenses and interest on capital should not, however, in any case form part of the cost of production. There is no advantage in distributing these items over the various transactions or articles produced. They do not vary proportionately with the volume of business. . . . The establishment charges are, in the aggregate, more or less constant, while the manufacturing costs fluctuate with the cost of labour and the price of material. To distribute the charges over the articles manufactured would, therefore, have the effect of disproportionately reducing the cost of production with every increase, and the reverse with every diminution, of business. Such a result is greatly to be deprecated, as tending to neither economy of management nor to accuracy in estimating for contracts. The principles of a business can always judge what percentage of gross profit upon cost is necessary to cover fixed establishment charges and interest on capital.” Emile Garcke and John Manger Fells (1893; 72-73)[197]. The aversion of accountants to include interest as a cost can be traced back to this quotation.

<sup>\*35</sup> Other methods for determining the appropriate interest rates that should be inserted into user cost formulae are discussed by Harper, Berndt and Wood (1989)[214] and in Chapter 5 of Schreyer (2001)[352]. Harper, Berndt and Wood (1989)[214] evaluate empirically 5 alternative rental price formulae using geometric depreciation but making different assumptions about the interest rate and the treatment of asset price change. They show that the choice of formula matters.

- Use an aggregate of nominal interest rates that the production sectors in the aggregate might be facing at the beginning of each period.
- Take a fixed real interest rate and add to it actual ex post consumer price inflation or anticipated consumer price inflation.

The first approach was used for the entire private production sector of the economy by Jorgenson and Griliches (1967; 267)[258] and for various sectors of the economy by Christensen and Jorgenson (1969; 307)[52]. It is also widely used by statistical agencies. It has the advantage that the value of output for the sector will exactly equal the value of input in a consistent accounting framework. It has the disadvantages that it is subject to measurement error and it is an *ex post rate of return* which may not reflect the economic conditions facing producers at the beginning of the period.

The second approach suffers from aggregation problems. There are many interest rates in an economy at the beginning of an accounting period and the problem of finding the “right” aggregate of these rates is not a trivial one.

The third approach works as follows. Let the consumer price index for the economy at the beginning of period  $t$  be  $c^t$  say. Then the ex post general consumer inflation rate for period  $t$  is  $\rho^t$  defined as:

$$1 + \rho^t \equiv c^{t+1}/c^t. \quad (8.17)$$

Let the production units under consideration face the real interest rate  $r^{*t}$ . Then by the Fisher (1896)[182] effect, the relevant nominal interest rate that the producers face should be approximately equal to  $r^t$  defined as follows:

$$r^t \equiv (1 + r^{*t})(1 + \rho^t) - 1. \quad (8.18)$$

The Australian Bureau of Statistics assumes that producers face a real interest rate of 4 %. This is consistent with long run observed economy wide real rates of return for most OECD countries which fall in the 2 to 5 per cent range. Thus following the example of the ABS, we could choose the real rate of return to be 4 % per annum; i.e., we assume that the nominal rate  $r^t$  is defined by (8.18) with the real rate defined by

$$r^{*t} \equiv .04 \quad (8.19)$$

assuming that the accounting period chosen is a year.\*<sup>36</sup>

We turn now to the determination of the asset inflation rates, the  $i^t$ , which appear in most of the formulae derived in the preceding sections of this chapter. There are three broad approaches that can be used in this context:

- Use actual ex post asset inflation rates over each period.
- Assume that each asset inflation rate is equal to the general inflation rate for each period.
- Estimate anticipated asset inflation rates for each period.

The problem with the first alternative is that ex post asset inflation rates tend to be very volatile and including actual ex post asset inflation rates in the user cost formula will tend to lead to very volatile user costs or even negative user costs. If our intention is to construct user costs that approximate market rental rates for the assets under consideration, then it usually will not be appropriate to use ex post asset inflation rates in the user cost formula.

When we assume that each asset inflation rate is equal to the general inflation rate  $\rho^t$  defined by (8.17), the equations presented earlier simplify. Thus if we replace  $1 + i^t$  by  $1 + \rho^t$  and  $1 + r^t$  by  $(1 + r^*)(1 + \rho^t)$ , equations (8.5), which relate the vintage asset prices  $P_n^t$  to the vintage rental prices

---

\*<sup>36</sup> If we are in a high inflation situation so that the accounting period becomes a quarter or a month, then  $r^{*t}$  must be chosen to be appropriately smaller.

$f_n^t$ , become:

$$\begin{aligned}
 P_0^t &= f_0^t + [1/(1+r^*)]f_1^t + [1/(1+r^*)]^2 f_2^t + [1/(1+r^*)]^3 f_3^t + \dots \\
 P_1^t &= f_1^t + [1/(1+r^*)]f_2^t + [1/(1+r^*)]^2 f_3^t + [1/(1+r^*)]^3 f_4^t + \dots \\
 &\dots \\
 P_n^t &= f_n^t + [1/(1+r^*)]f_{n+1}^t + [1/(1+r^*)]^2 f_{n+2}^t + [1/(1+r^*)]^3 f_{n+3}^t + \dots
 \end{aligned} \tag{8.20}$$

Note that only the constant real interest rate  $r^*$  appears in these equations.

If we replace  $1+i^t$  by  $1+\rho^t$  and  $1+r^t$  by  $(1+r^*)(1+\rho^t)$ , equations (8.14), which relate *the end of period vintage user costs*  $u_n^t$  to the vintage depreciation rates  $\delta_n^t$ , become:

$$\begin{aligned}
 u_0^t &= (1+\rho^t)[(1+r^*) - (1-\delta_0^t)]P_0^t = (1+\rho^t)[r^* + \delta_0^t]P_0^t \\
 u_1^t &= (1+\rho^t)(1-\delta_0^t)[(1+r^*) - (1-\delta_1^t)]P_0^t \\
 &= (1+\rho^t)(1-\delta_0^t)[r^* + \delta_1^t]P_0^t \\
 &\dots \\
 u_n^t &= (1+\rho^t)(1-\delta_0^t) \dots (1-\delta_{n-1}^t)[(1+r^*) - (1-\delta_n^t)]P_0^t \\
 &= (1+\rho^t)(1-\delta_0^t) \dots (1-\delta_{n-1}^t)[r^* + \delta_n^t]P_0^t.
 \end{aligned} \tag{8.21}$$

Now use equations (8.8) and  $1+r^t = (1+r^*)(1+\rho^t)$  and substitute into (8.21) to obtain the following equations, which relate the *beginning of period vintage user costs*  $f_n^t$  to the vintage depreciation rates  $\delta_n^t$ :

$$\begin{aligned}
 f_0^t &= (1+r^*)^{-1}[r^* + \delta_0^t]P_0^t \\
 f_1^t &= (1+r^*)^{-1}(1-\delta_0^t)[r^* + \delta_1^t]P_0^t \\
 &\dots \\
 f_n^t &= (1+r^*)^{-1}(1-\delta_0^t) \dots (1-\delta_{n-1}^t)[r^* + \delta_n^t]P_0^t.
 \end{aligned} \tag{8.22}$$

Note that only the constant real interest rate  $r^*$  appears in equations (8.22) but equations (8.21) also have the general inflation rate  $(1+\rho^t)$  as a multiplicative factor.

As mentioned above, in our third class of assumptions about asset inflation rates, we want to estimate *anticipated inflation rates* and use these estimates as our  $i^t$  in the various formulae we have exhibited. Unfortunately, there are any number of forecasting methods that could be used to estimate the anticipated inflation asset inflation rates. Alternatively, one could take a somewhat different approach than the pure forecasting one: one could simply *smooth* the observed ex post inflation rates and use these smoothed rates as our estimates of anticipated asset inflation rates.<sup>\*37</sup> However, there are a wide variety of smoothing methods and so again, we run into the *lack of reproducibility* problem.

To summarize our discussion on choosing interest rates and asset inflation rates to go into a user cost formula: there are many plausible alternatives and economists and statisticians have not been able to agree on “best” alternatives to use in practice.

In the next section, we turn our attention to the problem of aggregating across vintages of the same capital good.

<sup>\*37</sup> Unfortunately, different analysts may choose different smoothing methods so there may be a problem of a lack of reproducibility in our estimating procedures. Harper, Berndt and Wood (1989; 351)[214] note that the use of time series techniques to smooth ex post asset inflation rates and the use of such estimates as anticipated price change dates back to Epstein (1977)[172].

## 8.6 Aggregation over Vintages of a Capital Good

In previous sections, we have discussed the beginning of period  $t$  stock price  $P_n^t$  of an asset that is  $n$  periods old and the corresponding beginning and end of period user costs,  $f_n^t$  and  $u_n^t$ . The stock prices are relevant for the construction of *real wealth measures* of capital and the user costs are relevant for the construction of *capital services measures*. We now address the problems involved in obtaining quantity series that will match up with these prices.

Let the period  $t - 1$  investment in a homogeneous asset for the sector of the economy under consideration be  $I^{t-1}$ . We assume that the starting capital stock for a new unit of capital stock at the beginning of period  $t$  is  $K_0^t$  and this stock is equal to the new investment in the asset in the previous period; i.e., we assume:

$$K_0^t \equiv I^{t-1}. \quad (8.23)$$

Essentially, we are assuming that the length of the period is short enough so that we can neglect any contribution of investment to current production; a new capital good becomes productive only in the period immediately following its construction. In a similar manner, we assume that the capital stock available of an asset that is  $n$  periods old at the start of period  $t$  is  $K_n^t$  and this stock is equal to the gross investment in this asset class during period  $t - n - 1$ ; i.e., we assume:

$$K_n^t \equiv I^{t-n-1}; \quad n = 0, 1, 2, \dots \quad (8.24)$$

Given these definitions, the value of the capital stock in the given asset class for the sector of the economy under consideration (*the wealth capital stock*) at the start of period  $t$  is

$$\begin{aligned} W^t &\equiv P_0^t K_0^t + P_1^t K_1^t + P_2^t K_2^t + \dots \\ &= P_0^t I^{t-1} + P_1^t I^{t-2} + P_2^t I^{t-3} + \dots \quad \text{using (8.24)}. \end{aligned} \quad (8.25)$$

Turning now to the capital services quantity, we assume that the quantity of services that an asset of a particular vintage at a point in time is proportional (or more precisely, is equal) to the corresponding stock. Thus we assume that the quantity of services provided in period  $t$  by a unit of the capital stock that is  $n$  periods old at the start of period  $t$  is  $K_n^t$  defined by (8.24) above. Given these definitions, the value of capital services for all vintages of asset in the given asset class for the sector of the economy under consideration (*the productive services capital stock*) during period  $t$  using the end of period user costs  $u_n^t$  defined by equations (8.8) above is

$$\begin{aligned} S^t &\equiv u_0^t K_0^t + u_1^t K_1^t + u_2^t K_2^t + \dots \\ &= u_0^t I^{t-1} + u_1^t I^{t-2} + u_2^t I^{t-3} + \dots \quad \text{using (8.24)}. \end{aligned} \quad (8.26)$$

Now we are faced with the problem of decomposing the value aggregates  $W^t$  and  $S^t$  defined by (8.25) and (8.26) into separate price and quantity components. If we assume that each new unit of capital lasts only a finite number of periods,  $L$  say, then we can solve this value decomposition problem using normal index number theory. Thus define the period  $t$  vintage stock price and quantity vectors,  $\mathbf{P}^t$  and  $\mathbf{K}^t$  respectively, as follows:

$$\mathbf{P}^t \equiv [P_0^t, P_1^t, \dots, P_{L-1}^t]; \quad \mathbf{K}^t \equiv [K_0^t, K_1^t, \dots, K_{L-1}^t] = [I^{t-1}, I^{t-2}, \dots, I^{t-L-1}]; \quad t = 0, 1, \dots, T. \quad (8.27)$$

Fixed base or chain indexes may be used to decompose value ratios into price change and quantity change components. In empirical work involving annual data, it is preferable to use the chain

principle.<sup>\*38</sup> Thus the value of the capital stock in period  $t$ ,  $W^t$ , relative to its value in the preceding period,  $W^{t-1}$ , has the following index number decomposition:

$$W^t/W^{t-1} = P(\mathbf{P}^{t-1}, \mathbf{P}^t, \mathbf{K}^{t-1}, \mathbf{K}^t)Q(\mathbf{P}^{t-1}, \mathbf{P}^t, \mathbf{K}^{t-1}, \mathbf{K}^t); \quad t = 1, 2, \dots, T \quad (8.28)$$

where  $P$  and  $Q$  are *bilateral price and quantity indexes* respectively.

In a similar manner, we define the period  $t$  *vintage end of the period user cost price and quantity vectors*,  $\mathbf{u}^t$  and  $\mathbf{K}^t$  respectively, as follows:

$$\mathbf{u}^t \equiv [u_0^t, u_1^t, \dots, u_{L-1}^t]; \quad \mathbf{K}^t \equiv [K_0^t, K_1^t, \dots, K_{L-1}^t] = [I^{t-1}, I^{t-2}, \dots, I^{t-L-1}]; \quad t = 0, 1, \dots, T. \quad (8.29)$$

We ask that the value of capital services in period  $t$ ,  $S^t$ , relative to its value in the preceding period,  $S^{t-1}$ , has the following index number decomposition:

$$S^t/S^{t-1} = P(\mathbf{u}^{t-1}, \mathbf{u}^t, \mathbf{K}^{t-1}, \mathbf{K}^t)Q(\mathbf{u}^{t-1}, \mathbf{u}^t, \mathbf{K}^{t-1}, \mathbf{K}^t); \quad t = 1, 2, \dots, T \quad (8.30)$$

where again  $P$  and  $Q$  are *bilateral price and quantity indexes* respectively.

There is now the problem of choosing the functional form for either the price index  $P$  or the quantity index  $Q$ .<sup>\*39</sup> For most empirical work, we recommend the *Fisher* (1922)[188] *ideal price and quantity indexes*. These indexes appear to be “best” from the axiomatic viewpoint<sup>\*40</sup> and can also be given strong economic justifications.<sup>\*41</sup>

It should be noted that our use of an index number formula to aggregate over ages both stocks and services is more general than the usual aggregation over age procedures, which essentially assume that the different age of the same capital good are perfectly substitutable so that linear aggregation techniques can be used.<sup>\*42</sup> However, as we shall see in subsequent sections, the more general method of aggregation suggested here frequently reduces to the traditional linear method of aggregation provided that the prices by age all move in strict proportion over time.

Many researchers and statistical agencies relax the assumption that an asset lasts only a fixed number of periods,  $L$  say, and make assumptions about the distribution of retirements around the average service life,  $L$ . Usually, this extra degree of generality will not make much difference. However, the simultaneous retirement assumption can readily be relaxed (at the cost of much additional computational complexity) using the following methodology developed by Hulten:

“We have thus far taken the date of retirement  $T$  to be the same for all assets in a given cohort (all assets put in place in a given year). However, there is no reason for this to be true, and the theory is readily extended to allow for different retirement dates. A given cohort can be broken into components, or subcohorts, according to date of retirement and a separate  $T$  assigned to each. Each subcohort can then be characterized by its own efficiency sequence, which depends among other things on the subcohort’s useful life  $T_i$ .” Charles R. Hulten (1990; 125)[244].

We now have all of the pieces that are required in order to decompose the capital stock of an asset class and the corresponding capital services into price and quantity components. However, in order

<sup>\*38</sup> Given smoothly trending price and quantity data, the use of chain indexes will tend to reduce the differences between Paasche and Laspeyres indexes compared to the corresponding fixed base indexes and so chain indexes are generally preferred; see Diewert (1978; 895)[85] for a discussion.

<sup>\*39</sup> Obviously, given one of these functional forms, we may use (8.28) to determine the other.

<sup>\*40</sup> See Diewert (1992b; 214-223)[105].

<sup>\*41</sup> See Diewert (1976; 129-134)[82].

<sup>\*42</sup> This more general form of aggregation was first suggested by Diewert and Lawrence (2000)[142]. For descriptions of the more traditional linear method of aggregation, see Jorgenson (1989; 4)[255] or Hulten (1990; 121-127)[244] (1996; 152-165)[245].

to construct price and quantity components for capital services, we need information on the relative efficiencies  $f_n^t$  of the various ages of the capital input or equivalently, we need information on cross sectional depreciation rates by age  $\delta_n^t$  in order to use (8.30) above. The problem is that frequently, we do not have accurate information on either of these series and so instead, what is often done is that a standard asset life  $L$  is *assumed* and we make additional *assumptions* on the either the pattern of efficiencies or depreciation rates by age. Thus in a sense, this strategy follows the same somewhat mechanical strategy that was used by the early cost accountants:<sup>\*43</sup>

“The function of depreciation is recognized by most accountants as the provision of a means for spreading equitably the cost of comparatively long lived assets. Thus if a building will be of use during twenty years of operations, its cost should be recognized as operating expense, not of the first year, nor the last, but of all twenty years. Various methods may be proper in so allocating cost. The method used, however, is unimportant in this connection. The important matter is that at the time of abandonment, the cost of the asset shall as nearly as possible have been charged off as expense, under some systematic method.” M.B. Daniels (1933; 303)[63].

However, our suggested mechanical strategy is more complex than that used by early accountants in that we translate assumptions about the pattern of cross section depreciation rates into implications for the pattern of rental prices and asset prices by age, taking into account the complications induced by discounting and expected future asset price changes.

In the following sections, we will consider 4 different sets of assumptions and show how the resulting aggregate capital stocks and services could be constructed.<sup>\*44</sup>

## 8.7 The One Hoss Shay Model of Efficiency and Depreciation

In section 8.3 above, we noted that Böhm-Bawerk (1891; 342)[40] postulated that an asset would yield a constant level of services throughout its useful life of  $L$  years and then collapse in a heap to yield no services thereafter. This has come to be known as the one hoss shay or light bulb model of depreciation. Hulten noted that this pattern of relative efficiencies has the most intuitive appeal:

“Of these patterns, the one hoss shay pattern commands the most intuitive appeal. Casual experience with commonly used assets suggests that most assets have pretty much the same level of efficiency regardless of their age—a one year old chair does the same job as a 20 year old chair, and so on.” Charles R. Hulten (1990; 124)[244].

Thus the basic assumptions of this model are that the period  $t$  efficiencies and hence cross sectional rental prices  $f_n^t$  are all equal to say  $f^t$  for ages  $n$  that are less than  $L$  periods old and for older ages

<sup>\*43</sup> Canning (1929; 204)[44] criticized this strategy as follows: “The interminable argument that has been carried on by the text writers and others about the relative merits of the many formulas for measuring depreciation has failed, not only to produce the real merits of the several methods, but, more significantly, it has failed to produce a rational set of criteria of excellence whereby to test the aptness of any formula for any sub-class of fixed assets.”

<sup>\*44</sup> For an actual empirical example using Canadian data, see Diewert (2001)[121] or (2004a)[125]. Diewert constructed machinery and equipment and structures capital stocks for Canada. Average lives for these two asset classes varies greatly across countries. One source of information about asset lives is the OECD (1993)[323] where average service lives for various asset classes were reported for 14 or so OECD countries. For machinery and equipment (excluding vehicles) used in manufacturing activities, the average life ranged from 11 years for Japan to 26 years for the United Kingdom. For vehicles, the average service lives ranged from 2 years for passenger cars in Sweden to 14 years in Iceland and for road freight vehicles, the average life ranged from 3 years in Sweden to 14 years in Iceland. For buildings, the average service lives ranged from 15 years (for petroleum and gas buildings in the US) to 80 years for railway buildings in Sweden. Faced with this wide range of possible lives, Diewert followed the example of Angus Madison (1993)[298] and assumed an average service life of 14 years for machinery and equipment and 39 years for nonresidential structures.

of the asset, the efficiencies fall to zero. Thus we have:

$$f_n^t = \begin{cases} f^t & \text{for } n = 0, 1, 2, \dots, L-1; \\ 0 & \text{for } n = L, L+1, L+2, \dots \end{cases} \quad (8.31)$$

Now substitute (8.31) into the first equation in (8.5) and get the following formula<sup>\*45</sup> for the rental price  $f^t$  in terms of the price of a new asset at the beginning of year  $t$ ,  $P_0^t$ :

$$f^t = P_0^t/[1 + (\gamma^t) + (\gamma^t)^2 + \dots + (\gamma^t)^{L-1}] \quad (8.32)$$

where the period  $t$  discount factor  $\gamma^t$  is defined in terms of the period  $t$  nominal interest rate  $r^t$  and the period  $t$  expected asset inflation rate  $i^t$  as follows:

$$\gamma^t \equiv (1 + i^t)/(1 + r^t). \quad (8.33)$$

Now that the period  $t$  rental price  $f^t$  for an unretired asset has been determined, substitute equations (8.31) into equations (8.5) and determine the sequence of period  $t$  asset prices by age,  $P_n^t$ :

$$P_n^t = \begin{cases} f^t[1 + (\gamma^t) + (\gamma^t)^2 + \dots + (\gamma^t)^{L-1-n}] & \text{for } n = 0, 1, 2, \dots, L-1; \\ 0 & \text{for } n = L, L+1, L+2, \dots \end{cases} \quad (8.34)$$

Finally, use equations (8.8) to determine the end of period  $t$  rental prices,  $u_n^t$ , in terms of the corresponding beginning of period  $t$  rental prices,  $f_n^t$ :

$$u_n^t = (1 + r^t)f_n^t; \quad n = 0, 1, 2, \dots \quad (8.35)$$

Given the asset prices by age defined by (8.34), we can use equations (8.12) above to determine the corresponding cross sectional depreciation rates  $\delta_n^t$ .

We turn now to our second model of depreciation and efficiency.

## 8.8 The Declining Balance or Geometric Depreciation Model

The declining balance method of depreciation dates back to Matheson (1910; 55)[306] at least.<sup>\*46</sup> In terms of the algebra presented in section 8.4 above, the method is very simple: all of the cross sectional vintage depreciation rates  $\delta_n^t$  defined by (8.12) are assumed to be equal to the same rate  $\delta$ , where  $\delta$  a positive number less than one; i.e., we have for all time periods  $t$ :

$$\delta_n^t = \delta; \quad n = 0, 1, 2, \dots \quad (8.36)$$

Substitution of (8.36) into (8.14) leads to the following formula for the sequence of period  $t$  vintage user costs:

$$\begin{aligned} u_n^t &= (1 - \delta)^{n-1}[(1 + r^t) - (1 + i^t)(1 - \delta)]P_0^t; & n = 0, 1, 2, \dots \\ &= (1 - \delta)^{n-1}u_0^t; & n = 1, 2, \dots \end{aligned} \quad (8.37)$$

<sup>\*45</sup> This formula simplifies to  $P_0^t[1 - (\gamma^t)^L]/[1 - \gamma^t]$  provided that  $\gamma^t$  is less than 1 in magnitude. This last restriction does not always hold empirically, since for some years,  $i^t$  could exceed  $r^t$ . However, (8.32) is still valid under these conditions.

<sup>\*46</sup> Matheson (1910; 91)[306] used the term “diminishing value” to describe the method. Hotelling (1925; 350)[241] used the term “the reducing balance method” while Canning (1929; 276)[44] used the term the “declining balance formula”.

The second set of equations in (8.37) says *that all of the vintage user costs are proportional to the user cost for a new asset*. This proportionality means that we do not have to use an index number formula to aggregate over vintages to form a capital services aggregate. To see this, using (8.37), the period  $t$  services aggregate  $S^t$  defined earlier by (8.26) can be rewritten as follows:

$$\begin{aligned} S^t &\equiv u_0^t K_0^t + u_1^t K_1^t + u_2^t K_2^t + \dots \\ &= u_0^t [K_0^t + (1 - \delta)K_1^t + (1 - \delta)^2 K_2^t + \dots] \\ &= u_0^t K_A^t \end{aligned} \tag{8.38}$$

where the *period  $t$  capital aggregate*  $K_A^t$  is defined as

$$K_A^t \equiv K_0^t + (1 - \delta)K_1^t + (1 - \delta)^2 K_2^t + \dots \tag{8.39}$$

If the depreciation rate  $\delta$  and the vintage capital stocks are known, then  $K_A^t$  can readily be calculated using (8.39). Then using the last line of (8.38), we see that the value of capital services (over all vintages),  $S^t$ , decomposes into the price term  $u_0^t$  times the quantity term  $K_A^t$ . Hence, it is not necessary to use an index number formula to aggregate over vintages using this depreciation model.

A similar simplification occurs when calculating the wealth stock using this depreciation model. Substitution of (8.36) into (8.13) leads to the following formula for the sequence of period  $t$  vintage asset prices:

$$P_n^t = (1 - \delta)^{n-1} P_0^t; \quad n = 1, 2, \dots \tag{8.40}$$

Equations (8.40) say *that all of the vintage asset prices are proportional to the price of a new asset*. This proportionality means that again, we do not have to use an index number formula to aggregate over vintages to form a capital stock aggregate. To see this, using (8.40), the period  $t$  wealth aggregate  $W^t$  defined earlier by (8.25) can be rewritten as follows:

$$\begin{aligned} W^t &\equiv P_0^t K_0^t + P_1^t K_1^t + P_2^t K_2^t + \dots \\ &= P_0^t [K_0^t + (1 - \delta)K_1^t + (1 - \delta)^2 K_2^t + \dots] \\ &= P_0^t K_A^t \end{aligned} \tag{8.41}$$

where  $K_A^t$  was defined by (8.39). Thus  $K_A^t$  can serve as both a capital stock aggregate or a flow of services aggregate, which is a major advantage of this model.<sup>\*47</sup>

There is a further simplification of the model which is useful in applications. If we compare equation (8.39) for period  $t + 1$  and period  $t$ , we see that the following formula results using equations (8.39):

$$K_A^{t+1} \equiv K_0^{t+1} + (1 - \delta)K_A^t. \tag{8.42}$$

Thus the period  $t + 1$  aggregate capital stock,  $K_A^{t+1}$ , is equal to the investment in new assets that took place in period  $t$ , which is  $K_0^{t+1}$ , plus  $1 - \delta$  times the period  $t$  aggregate capital stock,  $K_A^t$ . This means that given a starting value for the capital stock, we can readily update it just using the depreciation rate  $\delta$  and the new investment in the asset during the prior period.

We now need to address the problem of determining the depreciation rate  $\delta$  for a particular asset class. Matheson was perhaps the first engineer to address this problem. On the basis of his experience, he simply postulated some approximate rates that could be applied:

<sup>\*47</sup> This advantage of the model has been pointed out by Jorgenson (1989)[255] (1996b)[257] and his coworkers. Its early application dates back to Jorgenson and Griliches (1967)[258] and Christensen and Jorgenson (1969)[52] (1973)[54].

“In most [brick or stone] factories an average of 3 per cent for buildings will generally be found appropriate, if due attention is paid to repairs. Such a rate will bring down a value of £1000 to £400 in thirty years.” Ewing Matheson (1910; 69)[306].

“Buildings of wood or iron would require a higher rate, ranging from 5 to 10 per cent, according to the design and solidity of the buildings, the climate, the care and the regularity of the painting, and according also, to the usage they are subjected to.” Ewing Matheson (1910; 69)[306].

“Contractors’ locomotives working on imperfect railroads soon wear out, and a rate of 20 per cent is generally required, bringing down the value of an engine costing £1000 to £328 in five years.” Ewing Matheson (1910; 86)[306].

“In engineering factories, where the work is of a moderate kind which does not strain the machines severely, and where the hours of working do not average more than fifty per week, 5 per cent written off each year from the diminishing value will generally suffice for the wear-and-tear of machinery, cranes and fixed plant of all kinds, if steam engines and boilers be excluded.” Ewing Matheson (1910; 82)[306].

“The high speed of the new turbo generators introduced since 1900, and their very exact fitting, render them liable to certain risks from variations in temperature and other causes. Several changes in regard to speed and methods of blading have occurred since their first introduction and if these generators are taken separately, only after some longer experience has been acquired can it be said that a depreciation rate of 10 per cent on the diminishing value will be too much for maintaining a book-figure appropriate to their condition. Such a rate will reduce £1000 to £349 in ten years.” Ewing Matheson (1910; 91)[306].

How did Matheson arrive at his estimated depreciation rates? He gave some general guidance as follows:

“The main factors in arriving at a fair rate of depreciation are:

1. The Original value.
2. The probable working Life.
3. The Ultimate value when worn out or superceded.

Therefore, in deciding upon an appropriate rate of depreciation which will in a term of years provide for the estimated loss, it is not the original value or cost which has to be so provided for, but that cost less the ultimate or scrap value.” Ewing Matheson (1910; 76)[306].

The algebra corresponding to Matheson’s method for determining  $\delta$  was explicitly described by the accountant Canning (1929; 276)[44]. Let the initial value of the asset be  $V_0$  and let its scrap value  $n$  years later be  $V_n$ . Then  $V_0, V_n$  and the depreciation rate  $\delta$  are related by the following equation:

$$V_n = (1 - \delta)^n V_0. \quad (8.43)$$

Canning goes on to explain that  $1 - \delta$  may be determined by solving the following equation:

$$\log(1 - \delta) = [\log V_n - \log V_0]/n. \quad (8.44)$$

It is clear that Matheson used this framework to determine depreciation rates even though he did not lay out formally the above straightforward algebra.

However, Canning had a very valid criticism of the above method:

“This method can be summarily rejected for a reason quite independent of the deficiencies of formulas 1 and 2 above [(8.43) and (8.44) above]. Overwhelming weight is given to  $V_n$  in determining book values. ... Thus the least important constant in reality is given the greatest effect in the formula.” John B. Canning (1929; 276)[44].

Thus Canning pointed out that the scrap value,  $V_n$ , which is not determined very accurately from an a priori point of view, is the tail that is wagging the dog; i.e., this poorly determined value plays a crucial role in the determination of the depreciation rate.<sup>\*48</sup>

An effective response to Canning's criticism of the declining balance method of depreciation did not emerge until relatively recently when Hall (1971)[208], Beidelman (1973)[32] (1976)[33] and Hulten and Wykoff (1981a)[246] (1981b)[247] used an entire array of used asset prices at point in time in order to determine the geometric depreciation rate which best matched up with the data.<sup>\*49</sup> Another theoretical possibility would be to use information on vintage rental prices in order to deduce the depreciation rate.<sup>\*50</sup> Hulten and Wykoff summarize their experience in estimating depreciation rates from used asset prices by concluding that the assumption of geometric or declining balance depreciation described their data relatively well:

“We have used the approach to study the depreciation patterns of a variety of fixed business assets in the United States (e.g., machine tools, construction equipment, autos and trucks, office equipment, office buildings, factories, warehouses, and other buildings). The straight line and concave patterns [i.e., one hoss shay patterns] are strongly rejected ; geometric is also rejected, but the estimated patterns are extremely close to (though steeper than) the geometric form, even for structures. Although it is rejected statistically, the geometric pattern is far closer than either of the other two candidates. This leads us to accept the geometric pattern as a reasonable approximation for broad groups of assets, and to extend our results to assets for which no resale markets exist by imputing depreciation rates based on an assumption relating the rate of geometric decline to the useful lives of assets.” Charles C. Hulten and Frank C. Wykoff (1996; 16)[248].

This brings us to our next problem: how can assumptions about asset lives in years be converted into geometric depreciation rates? In particular, how should we convert an estimated asset life of 39 years for structures and 14 years for machinery and equipment into comparable geometric rates?

One possible method for converting an average asset life,  $L$  periods say, into a comparable geometric depreciation rate is to argue as follows. Suppose that the straight line model of depreciation is the correct one (see Problem 1 below for a description of this model) and the asset under consideration has a useful life of  $L$  periods. Suppose further that investment in this type of asset is constant over time at one unit per period and asset prices are constant over time. Under these conditions, the long run equilibrium capital stock for this asset would be<sup>\*51</sup>:

$$1 + [(L - 1)/L] + [(L - 2)/L] + \cdots + [2/L] + [1/L] = L(L + 1)/2L = (L + 1)/2. \quad (8.45)$$

Under the same conditions, the long run equilibrium geometric depreciation capital stock would be

<sup>\*48</sup> “There may be cases in which the formula fits the facts, but ... the chance of its being a formula of close fit is remote indeed. Its chief usefulness seems to be to furnish drill in the use of logarithms for students in accounting.” John B. Canning (1929; 277)[44].

<sup>\*49</sup> Jorgenson (1996a)[256] has a nice review of most of the empirical studies of depreciation. It should be noted that Beidelman (1973)[32] (1976)[33] and Hulten and Wykoff (1981a)[246] (1996; 22)[248] showed that equation (8.44) must be adjusted to correct for the early retirement of assets. The accountant Schmalenbach (1959; 91)[351] (the first German edition was published in 1919) also noticed this problem: “The mistake should not be made, however, of drawing conclusions about useful life from those veteran machines which are to be seen in most businesses. Those which one sees are but the rare survivors; the many dead have long lain buried. This can be the source of serious errors.”

<sup>\*50</sup> This possibility is mentioned by Hulten and Wykoff (1996; 15)[248]: “In other words, if there were active rental markets for capital services as there are for labor services, the observed prices could be used to estimate the marginal products. And the rest of the framework would follow from these estimates. But, again, there is bad news: most capital is owner utilized, like much of the stock of single family houses. This means that owners of capital, in effect, rent it to themselves, leaving no data track for the analyst to observe.”

<sup>\*51</sup> The linear depreciation model implies that the vintage asset prices are proportional. Hence Hicks' Aggregation Theorem will imply that the capital aggregate will be the simple sum on the right hand side of (8.45).

equal to the following sum:

$$1 + (1 - \delta) + (1 - \delta)^2 + \dots = 1/[1 - (1 - \delta)] = 1/\delta. \quad (8.46)$$

Now find the depreciation rate  $\delta$  which will make the two capital stocks equal; i.e., equate (8.45) to (8.46) and solve for  $\delta$ . The resulting  $\delta$  is:

$$\delta = 2/(L + 1). \quad (8.47)$$

Obviously, there are a number of problematical assumptions that were made in order to derive the depreciation rate  $\delta$  that corresponds to the length of life  $L$ <sup>\*52</sup> but (8.47) gives us at least a definite method of conversion from one model to the other.

If we assume that the average length of life for nonresidential construction is  $L$  equal to 39 years, applying the conversion formula (8.47) implies that the corresponding  $\delta$  equals .05; similarly, an assumed life of 14 years for machinery and equipment translates into a  $\delta$  equal to a 13. 1/3% geometric depreciation rate for this asset class.

There is one remaining problem associated with the geometric depreciation model: how do we obtain a starting value for a geometric capital stock?

One method for dealing with this problem is due to Kohli<sup>\*53</sup> and it works as follows. Suppose we have collected investment data on a particular asset class for a number of periods, starting at period 0. Let the period  $t$  (real) investment be  $I^t$  and suppose that one plus the rate of growth of investment going from period  $t - 1$  to  $t$  is

$$1 + g^t \equiv I^t/I^{t-1} \quad \text{for } t = 1, 2, \dots, T. \quad (8.48)$$

Calculate the geometric average of these sample rates of growth  $g$  as follows:

$$1 + g \equiv \left\{ \prod_{t=1}^T (1 + g^t) \right\}^{1/T}. \quad (8.49)$$

If investments in this asset class had been growing in the past at the average sample rate of growth  $g$  and the geometric depreciation rate is  $\delta$ , then the geometric capital stock at the start of period 0,  $K^0$ , would be equal to:

$$\begin{aligned} K^0 &= I^0(1 + g)^{-1} \{ 1 + [(1 - \delta)/(1 + g)] + [(1 - \delta)/(1 + g)]^2 + [(1 - \delta)/(1 + g)]^3 + \dots \} \\ &= I^0(1 + g)^{-1} / (1 - [(1 - \delta)/(1 + g)]) \\ &= I^0 / [(1 + g) - (1 - \delta)] \\ &= I^0 / [g + \delta]. \end{aligned} \quad (8.50)$$

This completes our discussion of the geometric depreciation model for capital. Given its simplicity, it is our recommended model.

<sup>\*52</sup> The two assumptions that are the least justified are: (1) the assumption that the straight line depreciation model is the correct model to do the conversion and (2) the assumption that investment has been constant back to minus infinity. Hulten and Wykoff (1996; 16)[248] made the following suggestions for converting an  $L$  into a  $\delta$ : "Information is available on the average service life,  $L$ , from several sources. The rate of depreciation for non-marketed assets can be estimated using a two step procedure based on the 'declining balance' formula  $\delta = X/L$ . Under the 'double declining balance' formula,  $X = 2$ . The value of  $X$  can be estimated using the formula  $X = \delta L$  for those assets for which these estimates are available. In the Hulten-Wykoff studies, the average value for of  $X$  for producer's durable equipment was found to be 1.65 (later revised to 1.86). For nonresidential structures,  $X$  was found to be 0.91. Once  $X$  is fixed,  $\delta$  follows for other assets whose average service life is available."

<sup>\*53</sup> This method for obtaining a starting value for the geometric capital stock is due to Griliches (1980; 427)[204] and Kohli (1982)[273]; see also Fox and Kohli (1998)[190].

## 8.9 The Straight Line Method of Depreciation

The *straight line method of depreciation* is very simple in a world without price change: one simply makes an estimate of the most probable length of life for a new asset,  $L$  periods say, and then the original purchase price  $P_0^t$  is divided by  $L$  to yield as estimate of period by period depreciation for the next  $L$  periods. In a way, this is the simplest possible model of depreciation, just as the one hoss shay model was the simplest possible model of efficiency decline. We now set out the equations which describe the straight line model of depreciation in the general case when the anticipated asset inflation rate  $i^t$  is nonzero. Assuming that the asset has a life of  $L$  periods and that the cross sectional amounts of depreciation  $D_n^t \equiv P_n^t - P_{n+1}^t$  defined by (8.10) above are all equal for the assets in use, then it can be seen that the beginning of period  $t$  vintage asset prices  $P_n^t$  will decline linearly for  $L$  periods and then remain at zero; i.e., the  $P_n^t$  will satisfy the following restrictions:

$$P_n^t = \begin{cases} P_0^t[L - n]/L & n = 0, 1, 2, \dots, L \\ 0 & n = L + 1, L + 2, \dots \end{cases} \quad (8.51)$$

**Problem 1** Recall definition (8.12) above, which defined the cross sectional depreciation rate for an asset that is  $n$  periods old at the beginning of period  $t$ ,  $\delta_n^t$ .

(a) Using (8.51) and the  $n$ th equation in (8.13), show that:

$$(1 - \delta_0^t)(1 - \delta_1^t) \cdots (1 - \delta_{n-1}^t) = P_n^t/P_0^t = 1 - (n/L) \quad \text{for } n = 1, 2, \dots, L. \quad (i)$$

(b) Using (i) for  $n$  and  $n + 1$ , show that

$$(1 - \delta_n^t) = [L - (n + 1)]/[L - n] \quad n = 0, 1, 2, \dots, L - 1. \quad (ii)$$

(c) Now substitute (i) and (ii) into the general user cost formula (8.14) in order to obtain a formula for the *period  $t$  end of the period straight line user costs*,  $u_n^t$ , for  $n = 0, 1, 2, \dots, L - 1$ .<sup>\*54</sup>

*Comments:* Equations (8.51) give us the sequence of vintage asset prices that are required to calculate the wealth capital stock while your answer to part (c) gives us the vintage user costs that are required to calculate capital services for the asset.

(d) Suppose that the nominal interest rate  $r^t$  and the nominal asset inflation rate  $i^t$  are both zero. Calculate the sequence of asset prices  $P_n^t$  and user costs for  $n = 0, 1, 2, \dots, L$  under these conditions.

**Problem 2** Consider the one hoss shay model of depreciation.

(a) Suppose that the nominal interest rate  $r^t$  and the nominal asset inflation rate  $i^t$  are both zero. Calculate the sequence of asset prices  $P_n^t$  and user costs for  $n = 0, 1, 2, \dots, L$  for the one hoss shay model under these conditions.

(b) Compare the answers you obtained in part (a) of this question with the answer you obtained in part (d) of problem 1. Are there any relationships between your answers?

We conclude this section with some information on the early history of the straight line method for depreciation.

The accountant Canning summarized the straight line depreciation model as follows:

*“Straight Line Formula ...* In general, only two primary estimates are required to be made, viz., scrap value at the end of  $n$  periods and the numerical value of  $n$ . ... Obviously the number of

<sup>\*54</sup> The user costs for  $n = L, L + 1, L + 2, \dots$  are all zero.

periods of contemplated use of an asset can seldom be intelligently estimated without reference to the anticipated conditions of use. If the formula is to be respectable at all, the value of  $n$  must be the most probable number of periods that will yield the most economical use." John B. Canning (1929; 265-266)[44].

The following quotations indicate that the use of straight line depreciation dates back to the 1800's at least:

"Sometimes an equal installment is written off every year from the original value of the plant; sometimes each machine or item of plant is considered separately; but it is more usual to write off a percentage, not of the original value, but from the balance of the plant account of the preceding year." Ewing Matheson (1910; 55)[306].

"In some instances the amount charged to revenue account for depreciation is a fixed sum, or an arbitrary percentage on the book value." Emile Garcke and John Manger Fells (1893; 98)[197].

The last two quotations indicate that the declining balance or geometric depreciation model (to be considered in the next section) also dates back to the 1800's as a popular method for calculating depreciation.

## 8.10 The Linear Efficiency Decline Model

Recall that our first class of models (the one horse models) assumed that the efficiency (or cross section user cost) of the asset remained constant over the useful life of the asset. In our third class of models (the straight line depreciation models), we assumed that the cross sectional depreciation of the asset declined at a linear rate. In our second class of models (the geometric depreciation models), we assumed that cross section depreciation declined at a geometric rate. Comparing the third class with the second class of models, it can be seen that geometric depreciation is more *accelerated* than straight line depreciation; i.e., depreciation is relatively large for new vintages compared to older ones. In this section, we will consider another class of models that gives rise to an accelerated pattern of depreciation: the class of models that exhibit *a linear decline in efficiency*.

It is relatively easy to develop the mathematics of this model. Let  $f_0^t$  be the period  $t$  rental price for an asset that is new at the beginning of period  $t$ . If the useful life of the asset is  $L$  years and the efficiency decline is linear, then the sequence of period  $t$  cross sectional user costs  $f_n^t$  is defined as follows:

$$f_n^t \equiv \begin{cases} f_0^t[L - n]/L; & n = 0, 1, 2, \dots, L - 1; \\ 0 & n = L, L + 1, L + 2, \dots \end{cases} \quad (8.52)$$

**Problem 3** (a) Substitute (8.52) into the first equation in (8.5) and obtain a formula for the rental price  $f_0^t$  in terms of the price of a new asset at the beginning of year  $t$ ,  $P_0^t$ .

(b) Substitute the formula for  $f_0^t$  that you obtained in part (a) above into (8.5) and obtain the sequence of period  $t$  vintage asset prices,  $P_n^t$ , for  $n = 0, 1, 2, \dots, L - 1$  (the  $P_n^t$  will be 0 for  $n = L, L + 1, L + 2, \dots$ ).

**Problem 4** Assume that the nominal interest rate  $r^t$  and the nominal asset inflation rate  $i^t$  are both zero.

(a) Using your answer in part 3 (b) above, calculate:

$$D_n^t \equiv P_n^t - P_{n+1}^t \quad \text{for } n = 0, 1, 2, \dots, L. \quad (i)$$

**Comment:** Formula (i) should show that when  $r^t = i^t = 0$ , *depreciation declines at a linear rate* for the linear efficiency decline model. When depreciation declines at a linear rate, the resulting

formula for depreciation is called *the sum of the year digits formula*.<sup>\*55</sup> Thus just as the one hoss shay and straight line depreciation models coincide when  $r^t = i^t = 0$ , so too do the linear efficiency decline and sum of the digits depreciation models coincide.

## 8.11 Appendix 1: A Theoretical Treatment of Inventory Change

A theoretical framework is needed to measure the contribution of the change inventory stock over a period to production. It is also necessary to work out the user cost of the beginning of the period stock of inventories. A framework to answer these questions is outlined, taken from Diewert and Smith (1994)[150].

First consider the theory for a single inventory stock item. Consider a firm that perhaps produces a noninventory output during period  $t$ ,  $Y^t$ , uses a noninventory input  $X^t$ , sells the amount  $S^t$  of an inventory item during period  $t$  and makes purchases of the inventory item during period  $t$  in the amount  $B^t$ . Suppose that the average prices during period  $t$  of  $Y^t$ ,  $X^t$ ,  $S^t$  and  $B^t$  are  $P_Y^t$ ,  $P_X^t$ ,  $P_S^t$  and  $P_B^t$  respectively. Then neglecting balance sheet items, the firm's period  $t$  cash flow is:<sup>\*56</sup>

$$CF^t \equiv P_Y^t Y^t - P_X^t X^t + P_S^t S^t - P_B^t B^t. \quad (A1)$$

Let the firm's beginning of period  $t$  stock of inventory be  $K^{t-1}$  and let its end of period stock of inventory be  $K^t$ . These inventory stocks are valued at the balance sheet prices prevailing at the beginning and end of period  $t$ ,  $P_K^{t-1}$  and  $P_K^t$  respectively. Note that all 4 prices involving inventory items,  $P_S^t$ ,  $P_B^t$ ,  $P_K^{t-1}$  and  $P_K^t$  can be different.

The firm's period  $t$  economic income or net profit is defined as its cash flow plus the value of its end of period  $t$  stock of inventory items less  $(1 + r^t)$  times the value of its beginning of period  $t$  stock of inventory items:<sup>\*57</sup>

$$EI^t \equiv CF^t + P_K^t K^t - (1 + r^t) P_K^{t-1} K^{t-1} \quad (A2)$$

where  $r^t$  is the nominal cost of capital that the firm faces at the beginning of period  $t$ . Thus in definition (A2), it is assumed that the firm has to borrow financial capital or raise equity capital at the cost  $r^t$  in order to finance its initial holdings of inventory items. This cost could be real (in the case of a firm whose initial capital is funded by bonds) or it could be an opportunity cost (in the case of a firm entirely funded by equity capital).

The end of period stock of inventory is related to the beginning of the period stock by the following equation:

$$K^t = K^{t-1} + B^t - S^t - U^t \quad (A3)$$

where  $U^t$  denotes inventory items that are lost, spoiled, damaged or are used internally by the firm. In the case of livestock inventories, there is a natural growth rate of inventories over the period so equation (A3) is replaced by:

$$K^t = K^{t-1} + B^t - S^t + G^t \quad (A4)$$

<sup>\*55</sup> Canning (1929; 277)[44] describes the method in some detail so it was already in common use by that time.

<sup>\*56</sup> Note that this framework is flexible enough to allow the firm to either purchase or produce internally inventory items. Note also that firm purchases of inventory items from other domestic firms would appear in the national accounts as intermediate input purchases and purchases from foreign suppliers would appear as imports. On the other hand, sales of inventory items by the firm to domestic producers, households or foreigners would appear in the national accounts as gross outputs, final household consumption or exports respectively.

<sup>\*57</sup> In this definition of economic income, we are assuming that current flow variables are realized at the end of the accounting period and we are "discounting" the beginning of the period values of the inventory items in place at the beginning of the period to the end of the period. The resulting user costs will be end of the period user costs.

where  $G^t$  denotes the natural growth of the stock over period  $t$ .<sup>\*58</sup>

Define the *change in inventory stocks* over period  $t$  as:

$$\Delta K^t \equiv K^t - K^{t-1}. \quad (\text{A5})$$

Using (A5), both (A3) and (A4) can be written as:

$$K^t = K^{t-1} + \Delta K^t. \quad (\text{A6})$$

Now substitute (A6) into the definition of economic income (A2) and the following expression is obtained:

$$\begin{aligned} EI^t &\equiv CF^t + P_K^t [K^{t-1} + \Delta K^t] - (1 + r^t) P_K^{t-1} K^{t-1} \\ &= CF^t + P_K^t \Delta K^t - [r^t P_K^{t-1} - (P_K^t - P_K^{t-1})] K^{t-1}. \end{aligned} \quad (\text{A7})$$

Thus *economic income is equal to cash flow plus the value of the change in inventory (valued at end of period balance sheet prices) minus the user cost of inventories times the starting stocks of inventories* where this *period  $t$  user cost* is defined as

$$P_U^t \equiv r^t P_K^{t-1} - (P_K^t - P_K^{t-1}). \quad (\text{A8})$$

Note that the above algebra works for both livestock and ordinary inventory items.

There can be two versions of the user cost:

- An ex post version where the *actual* end of period balance sheet price of inventories is used or
- An ex ante version where at the beginning of period  $t$ , we estimate a *predicted value* for the end of period balance sheet price.

For the production accounts in the SNA, the ex ante version is the appropriate version, which means the national income accountant has some leeway in forming estimates of the end of period balance sheet price for the inventory item. Looking at (A7), it is important to note that the change in inventories that occurred over period  $t$ ,  $\Delta K^t$ , should be valued at the end of period  $t$  price for the inventory item,  $P_K^t$ .<sup>\*59</sup>

If the firm is using or selling many inventory items, say  $J$  items, then equation (A7) becomes:

$$EI^t \equiv CF^t + \sum_{j=1}^J P_{K_j}^t \Delta K_j^t - \sum_{j=1}^J [r^t P_{K_j}^{t-1} - (P_{K_j}^t - P_{K_j}^{t-1})] K_j^{t-1} \quad (\text{A9})$$

where the notation is obvious. The terms involving the value of the change in inventories over the period are the following ones:

$$\sum_{j=1}^J P_{K_j}^t \Delta K_j^t = \sum_{j=1}^J P_{K_j}^t [K_j^t - K_j^{t-1}] \quad (\text{A10})$$

$$= \sum_{j=1}^J P_{K_j}^t K_j^t - \sum_{j=1}^J P_{K_j}^t K_j^{t-1}. \quad (\text{A11})$$

Looking at (A10), it would appear that normal index number theory could be applied to the sum of terms in the value aggregate on the right hand side, with prices defined as the end of period  $t$

<sup>\*58</sup> If the firm is constructing inventory items either for direct sale or as an intermediate step in its production processes, then these produced additions to the stock would be included in the term  $G^t$ .

<sup>\*59</sup> However, the current SNA methodology requires that inventory change over the production period be evaluated at the average prices of the period. This requirement could be accommodated in our framework by replacing the end of period price of the inventory item,  $P_{K_j}^t$ , by an appropriate average inventory price for period  $t$ . If this is done, and if the actual end of period price of the inventory item is used for balance sheet purposes, then a reconciliation entry will be required in the Revaluation Accounts.

balance sheet prices  $P_{K_j}^t$  and corresponding quantities defined as the inventory changes  $K_j^t - K_j^{t-1}$  over period  $t$ . However, this value aggregate is not necessarily of one sign over time: it could be positive, negative or zero. *Normal index number theory breaks down for value aggregates that can be either positive or negative over time.*<sup>\*60</sup> Thus index number theory should not be applied to the value aggregate on the right hand side of (A10). Instead, it is recommended that index number theory be applied *separately* to the two value aggregates on the right hand side of (A11).<sup>\*61</sup> Thus  $\sum_{j=1}^J P_{K_j}^t K_j^t$  should be decomposed (using normal index number theory) into  $P_{KE}^t K_E^t$  where  $P_{KE}^t$  is the scalar end of period  $t$  aggregate price of inventories and  $K_E^t$  is the corresponding end of period  $t$  aggregate stock and  $\sum_{j=1}^J P_{K_j}^t K_j^{t-1}$  should be decomposed into  $P_{KB}^t K_B^t$  where  $P_{KB}^t$  is the scalar beginning of period  $t$  aggregate price of inventories and  $K_B^t$  is the corresponding beginning of period  $t$  aggregate stock. Then in place of the current single aggregate for inventory change that is reported in the current System of National Accounts, it is recommended that *inventory change be treated in a manner that is symmetric to the treatment of aggregate exports and imports in the accounts*; i.e., the end of period aggregates  $P_{KE}^t$  and  $K_E^t$  (the counterparts to the aggregate price of exports and the aggregate quantity of exports) and the beginning of period aggregates  $P_{KB}^t$  and  $K_B^t$  would be reported separately just as exports and imports are reported separately in the current SNA.

There is another treatment of inventory change that could be used by statistical agencies that is much more straightforward. The definition of economic income, (A2) above, can be rewritten as follows:

$$EI^t \equiv CF^t + P_K^t K^t - P_K^{t-1} K^{t-1} - r^t P_K^{t-1} K^{t-1}. \quad (\text{A12})$$

Using (A12), *the value of inventory change for period  $t$*  is simply defined as the end of period  $t$  value of the stock,  $VK^t$ , less the beginning of period  $t$  value of the stock,  $VK^{t-1}$ :

$$VK^t - VK^{t-1} = P_K^t K^t - P_K^{t-1} K^{t-1}. \quad (\text{A13})$$

Using this decomposition of economic income, the user cost value aggregate is defined as the last term on the right hand side of (A12) and so *the new user cost of inventories* is:

$$P_U^{t*} \equiv r^t P_K^{t-1}. \quad (\text{A14})$$

The new user cost of inventories,  $P_U^{t*}$  defined by (A14), can be compared to the initial user cost of inventories,  $P_U^t$  defined by (A8), and the new value of inventory change defined by (A13) can be compared to the earlier expression for the value of inventory change defined by (A11). Both the old and the new decomposition of economic income are theoretically valid. However, note that a nominal interest rate  $r^t$  appears in (A14) whereas a type of real interest rate appeared in (A8). Hence for a country experiencing high inflation, the new user cost of inventories will be higher than the old user cost and similarly, the new value of inventory change defined by (A13) will be higher than the old value of inventory change defined by (A11).<sup>\*62</sup> Thus nominal GDP will tend to be higher using the new decomposition compared to the initial one and it will be substantially higher under conditions of high inflation.

<sup>\*60</sup> To see why this breakdown occurs, consider a situation where the value aggregate just happens to be zero in the base period. Laspeyres price and quantity indexes will be undefined under these circumstances and nonsensical numbers will be obtained if the value aggregate is very close to zero in the base period. However, if the Laspeyres, Paasche or Fisher formula is used in forming a larger aggregate that is bounded well away from zero, then the right hand side of (A10) can be used when forming this larger aggregate and the same results will be obtained as using the right hand side of (A11) in forming the larger aggregate.

<sup>\*61</sup> This solution to the aggregation problem was suggested by Diewert (2004b; 36)[126].

<sup>\*62</sup> If the initial decomposition of economic income is used, then the beginning of the period inventory stocks are valued at the higher end of period prices but since this value aggregate is given a minus sign, this will reduce nominal GDP.

There are advantages and disadvantages of using the second decomposition of economic income compared to the first:

- The *main advantage* of the second decomposition is that it is much more straightforward and will be easier to explain to users. Also, it is much easier to reconcile quarterly changes in inventories to annual changes using the second decomposition.
- The *main disadvantage* of the second decomposition is that the resulting user cost of inventories is *different* from the user cost formula for reproducible capital and so an awkward asymmetry would be introduced into the SNA if a user cost approach to reproducible capital were introduced.\*<sup>63</sup>

Both decompositions of economic income involve a difference in two value aggregates where the sign of the difference cannot be bounded away from zero. Hence for both decompositions, it is recommended that the beginning and end of period values be separately deflated and shown as two items in the real accounts in a manner that is analogous to the present treatment of exports less imports.

## 8.12 Appendix 2: The Underlying Model of Production

In this Appendix, the motivation for the model of production that was defined by equations (A1) and (A2) in the previous Appendix is explained. This model of is based on a well established model of production that is used both by economists and thoughtful accountants as the following two quotations will show:

“We must look at the production process during a period of time, with a beginning and an end. It starts, at the commencement of the Period, with an Initial Capital Stock; to this there is applied a Flow Input of labour, and from it there emerges a Flow Output called Consumption; then there is a Closing Stock of Capital left over at the end. If Inputs are the things that are put in, the Outputs are the things that are got out, and the production of the Period is considered in isolation, then the Initial Capital Stock is an Input. A Stock Input to the Flow Input of labour; and further (what is less well recognized in the tradition, but is equally clear when we are strict with translation), the Closing Capital Stock is an Output, a Stock Output to match the Flow Output of Consumption Goods. Both input and output have stock and flow components; capital appears both as input and as output” John R. Hicks (1961; 23)[223].

“The business firm can be viewed as a receptacle into which factors of production, or inputs, flow and out of which outputs flow...The total of the inputs with which the firm can work within the time period specified includes those inherited from the previous period and those acquired during the current period. The total of the outputs of the business firm in the same period includes the amounts of outputs currently sold and the amounts of inputs which are bequeathed to the firm in its succeeding period of activity.” Edgar O. Edwards and Philip W. Bell (1961; 71-72)[168].

Hicks and Edwards and Bell obviously had the same model of production in mind: in each accounting period, the business unit combines the capital stocks and goods in process that it has inherited from the previous period with “flow” inputs purchased in the current period (such as labour, materials, services and additional durable inputs) to produce current period “flow” outputs as well as end of the period depreciated capital stock components which are regarded as outputs from the perspective

\*<sup>63</sup> The ex ante user cost for a reproducible capital asset contains an anticipated asset inflation rate in it similar to (A8), which offsets the nominal interest rate term. The ex ante user cost concept should be close to an actual rental or leasing price for the asset since it based on the same considerations that an owner would consider in setting a rental price. Hence, it seems desirable to have the user cost of inventories aligned with the user cost of reproducible capital.

of the current period (but will be regarded as inputs from the perspective of the next period).<sup>\*64</sup> The model could be viewed as an Austrian model of production in honour of the Austrian economist Böhm-Bawerk (1891)[40] who viewed production as an activity which used raw materials and labour to further process partly finished goods into finally demanded goods.

Now relate this theoretical model of production to equations (A1) and (A2) in the previous Appendix. All of the “flow” inputs that are purchased during the period and all of the “flow” outputs that are sold during the period are the inputs and outputs that appear in the definition of cash flow in definition (A1). These are the flow inputs and outputs that are very familiar to national income accountants. But this is not the end of the story: the firm inherits an endowment of assets at the beginning of the production period and at the end of the period, the firm will have the net profit or loss that has occurred due to its sales of outputs and its purchases of inputs during the period. As well, *it will have a stock of assets that it can use when it starts production in the following period.* Hence it seems clear that just focusing on the flow transactions that occur within the production period will not give a complete picture of the firm’s productive activities. National income accountants are aware of this when they make allowance for “work in progress”; i.e., production that takes place during the period but without any visible sales because it takes multiple periods to produce a saleable unit. Hence, to get a complete picture of the firm’s production over the course of a period, it is necessary to add the value of the closing stock of assets less the beginning of the period stock of assets to the cash flow that accrued to the firm from its sales and purchases of market goods and services during the accounting period. Using the notation explained in the previous appendix, this leads to the following definition of the firm’s period  $t$  *gross income* or *gross profit*, defined as its cash flow plus the value of its end of period  $t$  stock of inventory items less the value of its beginning of period  $t$  stock of inventory items:

$$GI^t \equiv CF^t + P_K^t K^t - P_K^{t-1} K^{t-1}. \quad (\text{A15})$$

The gross income or profit approach does not explicitly recognize interest as a cost of production. However, in order to induce investors in the firm to hold the starting stocks of capital items for productive purposes (instead of immediately selling them), it is necessary to pay interest. Thus it is necessary to subtract interest times the beginning of the period value of the capital stock from gross income in order to get the economic income or net profit  $EI^t$  defined earlier by (A2), which is repeated here for convenience:

$$EI^t \equiv CF^t + P_K^t K^t - (1 + r^t) P_K^{t-1} K^{t-1}. \quad (\text{A16})$$

There are two versions of economic income that could be considered for national income accounting purposes:

- An *ex post* version that uses the *actual* end of period  $t$  price as the price  $P_K^t$  in (A16) or
- An *ex ante* version that uses an *anticipated* end of period  $t$  price as the price  $P_K^t$  in (A16).

Diewert (1980; 476)[89] and Hill and Hill (2003)[237] endorsed the ex ante version for most purposes, since it will tend to be smoother than the ex post version and it will generally be closer to a rental or leasing price for the asset.

However, there are several practical measurement issues that will make it difficult to implement the ex ante version of net income:<sup>\*65</sup>

<sup>\*64</sup> For more on this model of production and additional references to the literature, see the Appendices in Diewert (1977)[83] (1980)[89].

<sup>\*65</sup> For noninventory assets, there will be difficulties in determining appropriate depreciation rates as well as the difficulties listed below.

- There may be difficulties in estimating the beginning of the period values of the various stocks held by firms since by definition, these stocks are being held (and not sold immediately) and so there are no unambiguous market prices to value these stocks.
- There may be difficulties in determining the right opportunity cost of financial capital  $r^t$ .
- It will be difficult to provide reproducible estimates of the anticipated end of period prices for the capital stocks being held by firms.

### 8.13 References

- Anthony, R.N. (1973), "Accounting for the Cost of Equity", *Harvard Business Review* 51, 88-102.
- Babbage, C. (1835), *On the Economy of Machinery and Manufactures*, Fourth Edition, London: Charles Knight.
- Baxter, W.T. (1971), *Depreciation*, London: Sweet and Maxwell.
- Baxter, W.T. (1984), *Inflation Accounting*, Oxford: Philip Allen Publishers.
- Beidelman, C. (1973), *Valuation of Used Capital Assets*, Sarasota Florida: American Accounting Association.
- Beidelman, C.R. (1976), "Economic Depreciation in a Capital Goods Industry", *National Tax Journal* 29, 379-390.
- Böhm-Bawerk, E. V. (1891), *The Positive Theory of Capital*, W. Smart (translator of the original German book published in 1888), New York: G.E. Stechert.
- Canning, J.B. (1929), *The Economics of Accountancy*, New York: The Ronald Press Co.
- Christensen, L.R. and D.W. Jorgenson (1969), "The Measurement of U.S. Real Capital Input, 1929-1967", *Review of Income and Wealth* 15, 293-320.
- Christensen, L.R. and D.W. Jorgenson (1973), "Measuring the Performance of the Private Sector of the U.S. Economy, 1929-1969", pp. 233-351 in *Measuring Economic and Social Performance*, M. Moss (ed.), New York: Columbia University Press.
- Church, A.H. (1901), "The Proper Distribution of Establishment Charges, Parts I, II, and III", *The Engineering Magazine* 21, 508-517; 725-734; 904-912.
- Clark, D. (1940), *The Conditions of Economic Progress*, London: Macmillan.
- Crandell, W.T. (1935), "Income and its Measurement", *The Accounting Review* 10, 380-400.
- Daniels, M.B. (1933), "The Valuation of Fixed Assets", *The Accounting Review* 8, 302-316.
- Davies, G.R. (1924), "The Problem of a Standard Index Number Formula", *Journal of the American Statistical Association* 27, 180-188.
- Diewert, W.E. (1974), "Intertemporal Consumer Theory and the Demand for Durables", *Econometrica* 42, 497-516.
- Diewert, W.E. (1976), "Exact and Superlative Index Numbers", *Journal of Econometrics* 4, 114-145.
- Diewert, W.E. (1977), "Walras' Theory of Capital Formation and the Existence of a Temporary Equilibrium", pp. 73-126 in *Equilibrium and Disequilibrium in Economic Theory*, E. Schwödiauer (ed.), Reidel Publishing Co.
- Diewert, W.E. (1978), "Superlative Index Numbers and Consistency in Aggregation", *Econometrica* 46, 883-900.
- Diewert, W.E. (1980), "Aggregation Problems in the Measurement of Capital", pp. 433-528 in *The Measurement of Capital*, D. Usher (ed.), Chicago: The University of Chicago Press.
- Diewert, W.E. (1992a), "The Measurement of Productivity", *Bulletin of Economic Research* 44:3, 163-198.

- Diewert, W.E. (1992b), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *The Journal of Productivity Analysis* 3, 211-248.
- Diewert, W.E. (1996), "Seasonal Commodities, High Inflation and Index Number Theory". Discussion Paper No. 96-06, Department of Economics, University of British Columbia, Vancouver, Canada, V6T 1Z1, January, available on the web at: <http://web.arts.ubc.ca/econ/diewert/Disc.htm>
- Diewert, W.E. (1998), "High Inflation, Seasonal Commodities and Annual Index Numbers", *Macroeconomic Dynamics* 2, 456-471.
- Diewert, W.E. (1999), "Index Number Approaches to Seasonal Adjustment", *Macroeconomic Dynamics* 3, 48-68.
- Diewert, W.E. (2001), "Measuring the Price and Quantity of Capital Services Under Alternative Assumptions", Discussion Paper 01-24, Department of Economics, University of British Columbia, Vancouver, Canada, June.
- Diewert, W.E. (2004a), "Measuring Capital", Discussion Paper 04-10, Department of Economics, University of British Columbia, Vancouver, Canada, July.
- Diewert, W.E. (2004b), "Index Number Problems in the Measurement of Real Net Exports and Real Net Changes in Inventories", paper presented at the Bureau of Economic Analysis, Washington D.C., September 21.
- Diewert, W.E. (2005a), "Issues in the Measurement of Capital Services, Depreciation, Asset Price Changes and Interest Rates", pp. 479-542 in *Measuring Capital in the New Economy*, C. Corrado, J. Haltiwanger and D. Sichel (eds.), Chicago: University of Chicago Press.
- Diewert, W.E. (2005b), "On Measuring Inventory Change in Current and Constant Dollars", Discussion Paper 05-12, Department of Economics, University of British Columbia, Vancouver, Canada, August.
- Diewert, W.E. and K.J. Fox (1999), "Can Measurement Error Explain the Productivity Paradox?", *Canadian Journal of Economics* 32, 251-280.
- Diewert, W.E. and D.A. Lawrence (2000), "Progress in Measuring the Price and Quantity of Capital", pp. 273-326 in *Econometrics and the Cost of Capital: Essays in Honor of Dale W. Jorgenson*, L.J. Lau (ed.), Cambridge MA: The MIT Press.
- Diewert, W.E. and D. Lawrence (2006), *Measuring the Contributions of Productivity and Terms of Trade to Australia's Economic Welfare*, Report by Meyrick and Associates to the Productivity Commission, Canberra, Australia.
- Diewert, W.E. H Mizobuchi, K Nomura (2005), "On Measuring Japan's Productivity, 1955-2003", Discussion Paper 05-05, Department of Economics, University of British Columbia, Vancouver, Canada.
- Diewert, W.E. and A.M. Smith (1994), "Productivity Measurement for a Distribution Firm", *The Journal of Productivity Analysis* 5, 335-347.
- Doms, M.E. (1996), "Estimating Capital Efficiency Schedules within Production Functions", *Economic Inquiry* 34, 78-92.
- Edwards, E.O. and P.W. Bell (1961), *The Theory and Measurement of Business Income*, Berkeley, California: University of California Press.
- Epstein, L.G. (1977), *Essays in the Economics of Uncertainty*, unpublished Ph. D thesis, Vancouver: The University of British Columbia.
- Eurostat (1993), *System of National Accounts 1993*, Brussels, Washington, Paris, New York and Washington: Eurostat, IMF, OECD, UN and World Bank.
- Fisher, I. (1896), *Appreciation and Interest*, New York: Macmillan.
- Fisher, I. (1897), "The Role of Capital in Economic Theory", *The Economic Journal* 7, 341-367.

- Fisher, I. (1908), "Are Savings Income?", *Publications of the American Economic Association*, Third Series 9, 21-47.
- Fisher, I. (1922), *The Making of Index Numbers*, London: Macmillan and Company.
- Fox, Kevin J. and Ulrich Kohli (1998) "GDP Growth, Terms-of-trade Effects, and Total Factor Productivity", *Journal of International Trade and Economic Development* 7, 87-110.
- Garcke, E. and J.M. Fells (1893), *Factory Accounts: Their Principles and Practice*, Fourth Edition, (First Edition 1887), London: Crosby, Lockwood and Son.
- Gilman, S. (1939), *Accounting Concepts of Profit*, New York: The Rolland Press Co.
- Griliches, Z. (1963), "Capital Stock in Investment Functions: Some Problems of Concept and Measurement", pp. 115-137 in *Measurement in Economics*, C. Christ and others (eds.), Stanford California: Stanford University Press; reprinted as pp. 123-143 in *Technology, Education and Productivity*, Z. Griliches (ed.), (1988), Oxford: Basil Blackwell.
- Griliches, Z. (1980), "Returns to Research and Development Expenditures in the Private Sector", pp. 419-454 in *New Developments in Productivity Measurement*, J.W. Kendrick and B. Vaccara (eds.), Studies in Income and Wealth, Volume 44, Chicago: University of Chicago Press.
- Hadar, E. and S. Peleg (1998), "Efforts to Present Useful National Accounts under High Inflation", paper presented at the 25th General Conference of The International Association for Research in Income and Wealth, Cambridge, UK, August 23-29.
- Hall, R.E. (1971), "the Measurement of Quality Change from Vintage Price Data", pp. 240-271 in *Price Indexes and Quality Change*, Z. Griliches (ed.), Cambridge Massachusetts: Harvard University Press.
- Harper, M.J., E.R. Berndt and D.O. Wood (1989), "Rates of Return and Capital Aggregation Using Alternative Rental Prices", pp. 331-372 in *Technology and Capital Formation*, D.W. Jorgenson and R. Landau (eds.), Cambridge MA: The MIT Press.
- Hicks, J.R. (1939), *Value and Capital*, Oxford: The Clarendon Press.
- Hicks, J.R. (1946), *Value and Capital*, Second Edition, Oxford: Clarendon Press.
- Hicks, J.R. (1961), "The Measurement of Capital in Relation to the Measurement of Other Economic Aggregates", pp. 18-31 in *The Theory of Capital*, F.A. Lutz and D.C. Hague (eds.), London: Macmillan.
- Hicks, J. (1973), *Capital and Time*, Oxford: The Clarendon Press.
- Hill, P. (1996), *Inflation Accounting: A Manual on National Accounting under Conditions of High Inflation*, Paris: OECD.
- Hill, P. (1999); "Capital Stocks, Capital Services and Depreciation"; paper presented at the third meeting of the Canberra Group on Capital Stock Statistics, Washington, D.C..
- Hill, P. (2000); "Economic Depreciation and the SNA"; paper presented at the 26th Conference of the International Association for Research on Income and Wealth; Cracow, Poland.
- Hill, R.J. and T.P. Hill (2003), "Expectations, Capital Gains and Income", *Economic Inquiry* 41, 607-619.
- Hotelling, H. (1925), "A General Mathematical Theory of Depreciation", *Journal of the American Statistical Association* 20, 340-353.
- Hulten, C.R. (1990), "The Measurement of Capital", pp. 119-152 in *Fifty Years of Economic Measurement*, E.R. Berndt and J.E. Triplett (eds.), Studies in Income and Wealth, Volume 54, The National Bureau of Economic Research, Chicago: The University of Chicago Press.
- Hulten, C.R. (1996), "Capital and Wealth in the Revised SNA", pp. 149-181 in *The New System of National Accounts*, J.W. Kendrick (ed.), New York: Kluwer Academic Publishers.

- Hulten, C.R. and F.C. Wykoff (1981a), "The Estimation of Economic Depreciation using Vintage Asset Prices", *Journal of Econometrics* 15, 367-396.
- Hulten, C.R. and F.C. Wykoff (1981b), "The Measurement of Economic Depreciation", pp. 81-125 in *Depreciation, Inflation and the Taxation of Income from Capital*, C.R. Hulten (ed.), Washington D.C.: The Urban Institute Press.
- Hulten, C.R. and F.C. Wykoff (1996), "Issues in the Measurement of Economic Depreciation: Introductory Remarks", *Economic Inquiry* 34, 10-23.
- Jorgenson, D.W. (1963), "Capital Theory and Investment Behaviour", *American Economic Review* 53:2, 247-259.
- Jorgenson, D.W. (1989), "Capital as a Factor of Production", pp. 1-35 in *Technology and Capital Formation*, D.W. Jorgenson and R. Landau (eds.), Cambridge MA: The MIT Press.
- Jorgenson, D.W. (1996a), "Empirical Studies of Depreciation", *Economic Inquiry* 34, 24-42.
- Jorgenson, D.W. (1996b), *Investment: Volume 2; Tax Policy and the Cost of Capital*, Cambridge, Massachusetts: The MIT Press.
- Jorgenson, D.W. and Z. Griliches (1967), "The Explanation of Productivity Change", *The Review of Economic Studies* 34, 249-283.
- Jorgenson, D.W. and Z. Griliches (1972), "Issues in Growth Accounting: A Reply to Edward F. Denison", *Survey of Current Business* 52:4, Part II (May), 65-94.
- Kohli, Ulrich (1982) "Production Theory, Technological Change, and the Demand for Imports: Switzerland, 1948-1974", *European Economic Review* 18, 369-386.
- Lachman, L.M. (1941), "On the Measurement of Capital", *Economica* 8, 361-377.
- Littleton, A.C. (1933), "Socialized Accounts", *The Accounting Review* 8, 267-271.
- Maddison, A. (1993), "Standardized Estimates of Fixed Capital Stock: A Six Country Comparison", *Innovazione e materie prime*, April, 1-29.
- Matheson, E. (1910), *Depreciation of Factories, Mines and Industrial Undertakings and their Valuations*, Fourth Edition, (First Edition 1884), London: Spon.
- Middleditch, L. (1918), "Should Accounts Reflect the Changing Value of the Dollar?", *The Journal of Accountancy* 25. 114-120.
- OECD (1993), *Methods Used by OECD Countries to Measure Stocks of Fixed Capital. National Accounts: Sources and Methods No. 2*, Paris: Organisation of Economic Co-operation and Development.
- Oliner, S.D. (1996), "New Evidence on the Retirement and Depreciation of Machine Tools", *Economic Inquiry* 34, 57-77.
- Schmalenbach, E. (1959), *Dynamic Accounting*, translated from the German 12th edition of 1955 (first edition 1919) by G.W. Murphy and K.S. Most, London: Gee and Company.
- Schreyer, P. (2001), *OECD Productivity Manual: A Guide to the Measurement of Industry-Level and Aggregate Productivity Growth*, Paris: OECD.
- Solomons, D. (1961), "Economic and Accounting Concepts of Income", *The Accounting Review* 36, 374-383.
- Solomons, D. (1968), "The Historical Development of Costing", pp. 3-49 in *Studies in Cost Analysis*, D. Solomons (ed.), Homewood Illinois: Richard D. Irwin.
- Triplett, J.E. (1996), "Depreciation in Production Analysis and in Income and Wealth Accounts: Resolution of an Old Debate", *Economic Inquiry* 34, 93-115.
- Tweedie, D. and G. Whittington (1984), *The Debate on Inflation Accounting*, London: Cambridge University Press.

Vanoli, A. (1998), "Interest and Inflation Accounting", paper presented at the 25th General Conference of The International Association for Research in Income and Wealth, Cambridge, UK, August 23-29.

Walsh, C.M. (1901), *The Measurement of General Exchange Value*, New York: Macmillan and Company.

Walsh, C.M. (1921), *The Problem of Estimation*, London: P.S. King and Son.

Walras, L. (1954), *Elements of Pure Economics*, a translation by W. Jaffé of the Edition Définitive (1926) of the *Eléments d'économie pure*, first edition published in 1874, Homewood, Illinois: Richard D. Irwin.

Whittington, G. (1980), "Pioneers of Income Measurement and Price Level Accounting: A Review Article", *Accounting and Business Research* 10 (Spring), 232-240.

Whittington, G. (1992), "Inflation Accounting", pp. 400-402 in *The New Palgrave Dictionary of Money and Finance*, Volume 2, P. Newman, M. Milgate and J. Eatwell (eds.), London: Macmillan.

## Chapter 9

# The Measurement of Income and the Determinants of Income Growth

### 9.1 Introduction

In this chapter, we will consider how to measure income. This would seem to be a very straightforward subject but as we shall see, it is far from being simple, even when we assume that there is only a single homogeneous reproducible capital good. We will also study the determinants of income growth; in particular, we will provide a formal production theoretic framework that will enable us to determine the relative importance to income growth of output price changes (including changes in the terms of trade), capital and labour growth and productivity growth.\*<sup>1</sup>

Virtually all economic discussions about the economic strength of a country use Gross Domestic or Gross National Product as “the” measure of output. But gross product measures do not account for the capital that is used up during the production period; i.e., the gross measures neglect depreciation. Thus in section 9.2, we consider some possible reasons why gross measures seem to be more popular than net measures.

Even though it may be difficult empirically to estimate depreciation and hence to estimate net output as opposed to gross output, we nevertheless conclude that for welfare purposes, the net measure is to be preferred. Net measures of output are also known as *income* measures. In section 9.3, we study in some detail Samuelson’s (1961)[345] discussion on alternative income concepts and how they might be implemented empirically. In particular, Samuelson (1961; 46)[345] gives a nice geometric interpretation of Hicks’ (1939; 174)[219] Income Number 3.

In section 9.4, we digress temporarily and generalize Samuelson’s (1961; 45-46)[345] index number method for measuring “income” change; i.e., we cover the pure theory of the output quantity index that was developed by Samuelson and Swamy (1974)[348], Sato (1976)[349] and Diewert (1983)[97].

In section 9.5, we note that Samuelson’s measures of income do not capture all of the complexities of the concept. Samuelson worked with a net investment framework but net investment is equal to capital at the end of the period less capital at the beginning of the period. Unfortunately, prices at the beginning of the period are not necessarily equal to prices at the end of the period. Thus Hicks noted that there was a “kind of index number problem” in comparing capital stocks at the beginning and end of the period:\*<sup>2</sup>

“At once we run into the difficulty that if Net Investment is interpreted as the difference between the value of the Capital Stock at the beginning and end of the year, the transformation

---

\*<sup>1</sup> This chapter draws heavily on Diewert (2006b)[132] and Diewert and Lawrence (2006)[145].

\*<sup>2</sup> We studied this problem in the previous chapter but we revisit it again in the present chapter.

would not be possible. It is only in the special case when the prices of all sorts of capital instruments are the same (if their condition is the same) at the end of the year as at the beginning, that we should be able to measure the money value of Real Net Investment by the increase in the Money value of the Capital stock. In all probability these prices will have changed during the year, so that we have a kind of index number problem, parallel to the index number problem of comparing real income in different years. The characteristics of that other problem are generally appreciated; what is not so generally appreciated is the fact that before we can begin to compare real income in different years, we have to solve a similar problem within the single year—we have to reduce the Capital stock at the beginning and end of the years into comparable real terms.” J.R. Hicks (1942; 175-176)[221].

In section 9.5, we look at various possible alternatives for making the capital stocks at the beginning and end of the year comparable to each other in real terms.

In section 9.6, we return to the accounting problems associated with the profit maximization problem of a production unit, using the Hicks (1961; 23)[223] and Edwards and Bell (1961; 71-72)[168] Austrian production function framework studied in Appendix 2 of chapter 8. In this section, we show how the traditional gross rentals user cost formula can be decomposed into three terms—one reflecting the reward for “waiting”, the second one reflecting anticipated asset price changes and the last term reflecting depreciation—and then we show how the depreciation term can be transferred from the list of inputs and regarded as a negative output, which leads to an income concept that was studied in section 9.5. This transfer is equivalent to treating depreciation as an intermediate input. Another income concept studied in section 9.5 emerges if we also regard the anticipated price change term as an intermediate input.

In section 9.7 we present various approximations to the theoretical target income concept—approximations that can be implemented empirically. Sections 9.6 and 9.7 also touches on the obsolescence and depreciation controversy that dates back to Hayek (1941)[218] and Pigou (1941)[329]. Section 9.8 summarizes our discussion on income concepts.

The final sections in the chapter develop a production theory framework that tries to explain the various factors behind the growth in real income that the market sector of an economy can generate. The main factors that explain real income growth are:

- Changes in the prices of the outputs that the market sector produces and changes in the prices of the intermediate inputs that it uses. These price changes include changes in the economy’s terms of trade.
- Changes in the amounts of primary inputs that the market sector uses.
- Changes in the productivity of the economy.

Section 9.9 develops the production theory framework in general terms while section 9.10 uses the assumption that the technology of the market sector can be described by a translog variable profit function. In the latter case, an exact decomposition of real income growth generated by the market sector into explanatory factors can be obtained.\*<sup>3</sup>

Section 9.11 extends the analysis of section 9.10 to deal with the contribution of changes in the terms of trade to real income growth.

## 9.2 Measuring National Product: Gross versus Net

Real Gross Domestic Product, per capita real GDP and labour productivity (real GDP divided by hours worked in the economy) are routinely used to compare “welfare” levels between countries (and between time periods in the same country). Gross Domestic Product is the familiar  $C+G+I+X-M$

---

\*<sup>3</sup> This decomposition is due to Diewert and Lawrence (2006)[145].

or, in a closed economy with no government, it is simply  $C + I$ , consumption plus gross investment that takes place during an accounting period. However, economists have argued for a long time that GDP is not the “right” measure of output for welfare purposes; rather NDP (Net National Product), equal to consumption plus net investment accruing to nationals, is a much better measure, where net investment equals gross investment less depreciation.\*<sup>4</sup> Why has GDP remained so much more popular than NDP, given that NDP seems to be the better measure for “welfare” comparison purposes?\*<sup>5</sup>

Samuelson (1961)[345] had a good discussion of the arguments that have been put forth to justify the use of GDP over NDP:

Within the framework of a purely theoretical model such as this one, I believe that we should certainly prefer net national product, NNP, to gross national product, GNP, if we were forced to choose between them. This is somewhat the reverse of the position taken by many official statisticians, and so let me dispose of three arguments used to favour the gross concept.” Paul A. Samuelson (1961; 33)[345].

The first argument that Samuelson considered was that our estimates of depreciation are so inaccurate that it is better to measure GDP or GNP rather than NDP or NNP. Samuelson was able to dispose of this argument in his context of a purely theoretical model as follows:

“Within our simple model, we know precisely what depreciation is and so for our present purpose this argument can be provisionally ruled out of order.” Paul A. Samuelson (1961; 33)[345].

However, in our practical measurement context, we cannot dismiss this argument so easily and we have to concede that the fact that our empirical estimates of depreciation are so shaky, is indeed an argument to focus on measuring GDP rather than NDP.\*<sup>6</sup>

The second argument that Samuelson considered was the argument that GNP reflects the productive potential of the economy:

“Second, there is the argument that GNP gives a better measure than does NNP of the maximum consumption sprint that an economy could make by consuming its capital in time of future war or emergency.” Paul A. Samuelson (1961; 33)[345].

Samuelson (1961; 34)[345] is able to dismiss this argument by noting that NNP is not the maximum short run production that could be squeezed out of an economy: by running down capital to an extraordinary degree, we could increase present period output to a level well beyond current GNP.

The third argument that Samuelson considered had to do with the difficulties involved in determining obsolescence:

“A third argument favouring a gross rather than net product figure proceeds as follows: new

---

\*<sup>4</sup> See Marshall (1890)[304] and Pigou (1924; 46)[327] (1935; 240-241)[328] (1941; 271)[329] for example. A more recent paper that argues for the net product framework is Diewert and Fox (2005)[139].

\*<sup>5</sup> In the present chapter, we will assume that the economy is closed so that the distinction between domestic product and national product (e.g., NDP versus NNP) vanishes. Hence our focus is on justifying either a gross product or a net product concept.

\*<sup>6</sup> Hicks (1973; 155)[227] conceded that this argument for GDP or GNP has some validity: “There are items, of which Depreciation and Stock Appreciation are the most important, which do not reflect actual transactions, but are estimates of the changes in the value of assets which have not yet been sold. These are estimates in a different sense from that previously mentioned. They are not estimates of a statistician’s true figure, which happens to be unavailable; there is no true figure to which they correspond. They are estimates relative to a purpose; for different purposes they may be made in different ways. This is of course the basic reason why it has become customary to express the National Accounts in terms of Gross National Product (before deduction of Depreciation) so as to clear them of contamination with the ‘arbitrary’ depreciation item; though it should be noticed that even with GNP another arbitrary element remains, in stock accumulation.”

capital is progressively of better quality than old, so that net product calculated by the subtraction of all depreciation and obsolescence does not yield an ideal measure ‘based on the principle of keeping intact the physical productivity of the capital goods in some kind of welfare sense.’” Paul A. Samuelson (1961; 35)[345].

Again Samuelson dismisses this argument in the context of his theoretical model (where all is known) but in the practical measurement context, we have to concede that this argument has some validity, just as did the first argument.

From our point of view, the problem with the gross concept is that it gives us a measure of output that is not sustainable. By deducting even an imperfect measure of depreciation (and obsolescence) from gross investment, we will come closer to a measure of output that could be consumed in the present period without impairing production possibilities in future periods. Hence, for welfare purposes, measures of net product seem to be much preferred to gross measures, even if our estimates of depreciation and obsolescence are imperfect.\*7

In the following section, we will look at some alternative definitions of net product. Given a specific definition for net product and given an accounting system that distributes the value of outputs produced to inputs utilized, each definition of net product gives rise to a corresponding definition of “income”. In the economic literature, most of the discussion of alternative measures of net output has occurred in the context of alternative “income” measures and so in the following section, we will follow the literature and discuss alternative “income” measures rather than alternative measures of “net product”.

### 9.3 Measuring Income: Hicks versus Samuelson

Samuelson (1961; 45-46)[345] constructed a nice diagram which illustrated alternative income concepts in a very simple model where the economy produces only two goods: consumption  $C$  and a durable capital input  $K$ . *Net investment* during period  $t$  is defined as  $\Delta K^t \equiv K^t - K^{t-1}$ , the end of the period capital stock,  $K^t$ , less the beginning of the period capital stock,  $K^{t-1}$ . In Figure 9.1 below, let the economy’s period 2 production possibilities set for producing combinations of consumption  $C$  and net investment  $\Delta K$  be represented by the curve HGBE\*8 and let the economy’s period 1 production possibilities set for producing consumption and nonnegative net investment be represented by the curve FAD. Assume that the actual period 2 production point is represented by the point B and the actual period 1 production point is represented by the point A.

Samuelson used the definition of income that was due to Marshall (1890)[304] and Haig (1921)[207], who (roughly speaking) defined income as consumption plus the consumption equivalent of the increase in net wealth over the period:

\*7 This point of view is also expressed in the *System of National Accounts 1993*: “As value added is intended to measure the additional value created by a process of production, it ought to be measured net, since consumption of fixed capital is a cost of production. However, as explained later, consumption of fixed capital can be difficult to measure in practice and it may not always be possible to make a satisfactory estimate of its value and hence of net value added.” Eurostat (1993; 121)[174]. “The consumption of fixed capital is one of the most important elements in the System. ... Moreover, consumption of fixed capital does not represent the aggregate value of a set of transactions. It is an imputed value whose economic significance is different from entries in the accounts based only on market transactions. For these reasons, the major balancing items in national accounts have always tended to be recorded both gross and net of consumption of fixed capital. This tradition is continued in the System where provision is also made for balancing items from value added through to saving to be recorded both ways. In general, the gross figure is obviously the easier to estimate and may, therefore, be more reliable, but the net figure is usually the one that is conceptually more appropriate and relevant for analytical purposes.” Eurostat (1993; 150)[174].

\*8 The point H on the period 2 production possibilities set would represent a consumption net investment point where the end of the period capital stock is less than the beginning of the period stock so that consumption is increased at the cost of running down the capital stock. The period 1 production possibilities set could similarly be extended to the left of the point F.

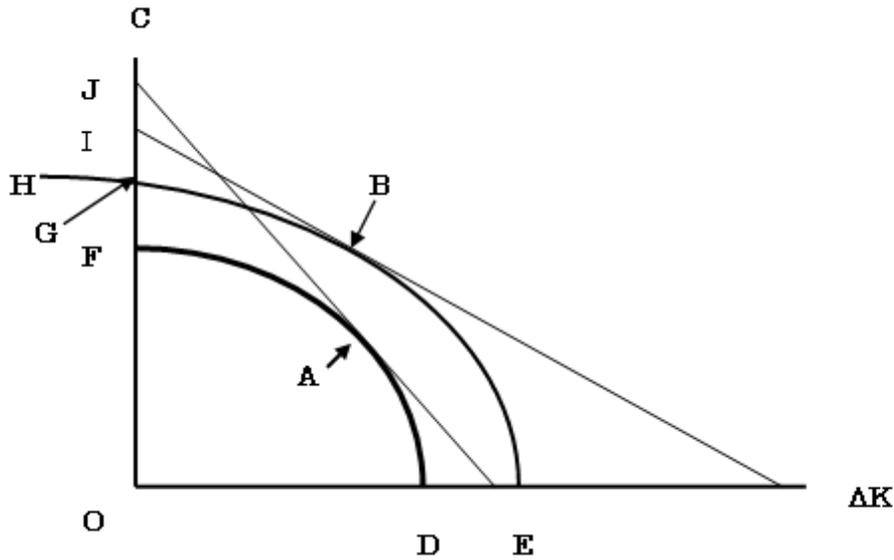


Fig. 9.1 Alternative Income Concepts

“The Haig-Marshall definition of income can be defended by one who admits that consumption is the ultimate end of economic activity. In our simple model, the Haig-Marshall definition measures the economy’s *current power to consume* if it wishes to do so.” Paul A. Samuelson (1961; 45)[345].

Samuelson went on to describe a number of methods by which the Haig-Marshall definition of income or net product could be implemented. Three of his suggested methods will be of particular interest to us.

**Method 1: The Market Prices Method**

If producers are maximizing the value of consumption plus net investment subject to available labour and initial capital resources in each period, then in each period, there will be a market revenue line that is tangent to the production possibilities set. Thus the revenue line BI is tangent to the period 2 set and the line JA is tangent to the period 1 production possibilities set. Each of these revenue lines can be used to convert the period’s net investment into consumption equivalents at the market prices prevailing in each period. Thus in period 1, the consumption equivalent of the observed production point A is the point J while in period 2, the consumption equivalent of the observed production point B is the point I and so using this method, Haig-Marshall income is higher in period 1 than in 2, since J is above I.\*<sup>9</sup>

**Method 2: Samuelson’s Index Number Method**

Let the point A be the period 1 consumption, net investment point  $C^1, I^1$  with corresponding market prices  $P_C^1, P_I^1$  and let the point B be the period 2 consumption, net investment point  $C^2, I^2$  with corresponding market prices  $P_C^2, P_I^2$ . Samuelson suggested computing the Laspeyres and Paasche

\*<sup>9</sup> Some statisticians would, I think, tend to measure incomes by the vertical intercepts of the tangent lines through A and B. On their definition, A would involve more income than B.” Paul A. Samuelson (1961; 45)[345].

quantity indexes for net output,  $Q_L$  and  $Q_P$ :

$$Q_L \equiv [P_C^1 C^2 + P_I^1 I^2] / [P_C^1 C^1 + P_I^1 I^1]; \quad (9.1)$$

$$Q_P \equiv [P_C^2 C^2 + P_I^2 I^2] / [P_C^2 C^1 + P_I^2 I^1]. \quad (9.2)$$

If  $Q_L$  and  $Q_P$  are both greater than one, then Samuelson would say that income in period 2 is greater than in period 1; if  $Q_L$  and  $Q_P$  are both less than one, then Samuelson would say that income in period 2 is less than in period 1; if  $Q_L$  and  $Q_P$  are both equal to one, then Samuelson would say that income in period 1 is equal to period 2 income. If  $Q_L$  and  $Q_P$  are such that one is less than one and the other greater than one, then Samuelson would term the situation inconclusive.\*<sup>10</sup>

We will indicate in the following section how Samuelson's analysis can be generalized to deal with the indeterminate case, using modern index number theory. Note that this method for measuring the growth in real income boils down to a method which somehow summarizes the shift in the production possibilities frontier going from period 1 to period 2.

### Method 3: Hicksian Income

Hicks (1939)[219] made a number of definitions of income. The one that Samuelson chose to model is Hicks' income Number 3:\*<sup>11</sup>

“Income No. 3 must be defined as the maximum amount of money which the individual can spend this week, and still expect to be able to spend the same amount *in real terms* in each ensuing week.” J.R. Hicks (1939; 174)[219].

Referring back to Figure 9.1 above, Samuelson (1961; 46)[345] interpreted Hicksian income in period 1 as the point F (which is where the period 1 production frontier intersects the consumption axis so that net investment would be 0 at this point) and Hicksian income in period 2 as the point G (which is where the period 2 production frontier intersects the consumption axis so that net investment would be 0 at this point). However, Samuelson also noted that this definition of income is less useful to the economic statistician than the above two definitions because the economic statistician will not be able to determine where the production frontier will intersect the consumption axis:

“Others (e.g. Hicks of the earlier footnote) want to measure income by comparing the vertical intercepts of the curved production possibility schedules passing respectively through A and B. This is certainly one attractive interpretation of the spirit behind Haig and Marshall. The practical statistician might despair of so defining income: using market prices and quantities, he could conceivably apply any of the other definitions; but this one would be non-observable to him.” Paul A. Samuelson (1961; 46)[345].

All three of the above definitions of income have some appeal. At this stage, we will not commit to any single definition since we have not yet explored the full complexities of the income concept.\*<sup>12</sup>

\*<sup>10</sup> “Neither Haig nor Marshall have told us exactly how they would evaluate and compare A and B in Fig. 3. Certainly some economic statisticians would interpret them as follows: Money national income is meaningless; you must deflate the money figures and reduce things to constant dollars. To deflate, apply the price ratios of B to the A situation and compare with B; alternatively, apply the price ratios of A to the B situation and compare with A. If both tests give the same answer—and in Fig. 3 they will, because B lies outside A on straight lines parallel to the tangent at either A or at B—then you can be sure that one situation has ‘more income’ than the other. If these Laspeyres and Paasche tests disagree, reserve judgment or split the difference depending upon your temperament.” Paul A. Samuelson (1961; 45-46)[345].

\*<sup>11</sup> This is the “best” Hicksian definition in my opinion but it has some ambiguity associated with it: how exactly do we interpret the word “real”?

\*<sup>12</sup> Samuelson's model did not have the added complexities of the Edwards and Bell (1961; 71-72)[168] and Hicks (1961; 23)[223] Austrian production model that distinguished the beginning of the period and end of the period capital stocks as separate inputs and outputs. Also Samuelson had only a single consumption good and a single

We conclude this section with another astute observation made by Samuelson:<sup>\*13</sup>

“Our dilemma is now well depicted. The simplest economic model involves two current variables, consumption and investment. A measure of national income is one variable. How can we fully summarize a doublet of numbers by a single number?” Paul A. Samuelson (1961; 47)[345].

## 9.4 The Theory of the Output Index

In this section, we will have another look at Samuelson’s index number method for measuring income growth; i.e., his second income or net product concept studied in the previous section. We consider a more general model where there are  $M$  consumption goods and net investment goods and  $N$  primary inputs. We also consider more general indexes than the Laspeyres and Paasche output quantity indexes considered by Samuelson.

We assume that the market sector of the economy produces quantities of  $M$  (net) outputs,  $\mathbf{y} \equiv [y_1, \dots, y_M]$ , which are sold at the positive producer prices  $\mathbf{P} \equiv [P_1, \dots, P_M]$ . We further assume that the market sector of the economy uses positive quantities of  $N$  primary inputs,  $\mathbf{x} \equiv [x_1, \dots, x_N]$  which are purchased at the positive primary input prices  $\mathbf{W} \equiv [W_1, \dots, W_N]$ . In period  $t$ , we assume that there is a feasible set of output vectors  $\mathbf{y}$  that can be produced by the market sector if the vector of primary inputs  $\mathbf{x}$  is utilized by the market sector of the economy; denote this period  $t$  production possibilities set by  $S^t$ . We assume that  $S^t$  is a closed convex cone that exhibits a free disposal property.<sup>\*14</sup>

Given a vector of output prices  $\mathbf{P}$  and a vector of available primary inputs  $\mathbf{x}$ , we define *the period  $t$  market sector net product function*,  $g^t(\mathbf{P}, \mathbf{x})$ , as follows:<sup>\*15</sup>

$$g^t(\mathbf{P}, \mathbf{x}) \equiv \max_{\mathbf{y}} \{ \mathbf{P} \cdot \mathbf{y} : (\mathbf{y}, \mathbf{x}) \text{ belongs to } S^t \}; \quad t = 0, 1, 2, \dots \quad (9.3)$$

Thus market sector NDP depends on  $t$  (which represents the period  $t$  technology set  $S^t$ ), on the vector of output prices  $\mathbf{P}$  that the market sector faces and on  $\mathbf{x}$ , the vector of primary inputs that is available to the market sector.

If  $\mathbf{P}^t$  is the period  $t$  output price vector and  $\mathbf{x}^t$  is the vector of inputs used by the market sector during period  $t$  and if the NDP function is differentiable with respect to the components of  $\mathbf{P}$  at the point  $\mathbf{P}^t, \mathbf{x}^t$ , then the period  $t$  vector of market sector outputs  $\mathbf{y}^t$  will be equal to the vector of first order partial derivatives of  $g^t(\mathbf{P}^t, \mathbf{x}^t)$  with respect to the components of  $\mathbf{P}$ ; i.e., we will have

---

capital input in his model and we need to also consider the problems involved in aggregating over consumption and capital stock components.

<sup>\*13</sup> Samuelson’s quotation succinctly states the index number problem! For additional material on output indexes, see Balk (1998)[20].

<sup>\*14</sup> For more explanation of the meaning of these properties, see Chapter 3 or Diewert (1973)[74] (1974; 134)[76] or Woodland (1982)[405] or Kohli (1978)[272] (1991)[275]. The assumption that  $S^t$  is a cone means that the technology is subject to constant returns to scale. This is an important assumption since it implies that the value of outputs should equal the value of inputs in equilibrium. In empirical work, this property can be imposed upon the data by using an ex post rate of return in the user costs of capital, which forces the value of inputs to equal the value of outputs for each period. The function  $g^t$  is known as the *NDP function* or the *net national product function* in the international trade literature (see Kohli (1978)[272] (1991)[275], Woodland (1982)[405] and Feenstra (2004; 76)[176]. It was introduced into the economics literature by Samuelson (1953)[342]. Alternative terms for this function include: (i) the *gross profit function*; see Gorman (1968)[201]; (ii) the *restricted profit function*; see Lau (1976)[286] and McFadden (1978)[309]; and (iii) the *variable profit function*; see Diewert (1973)[74] (1974)[76].

<sup>\*15</sup> The function  $g^t(\mathbf{P}, \mathbf{x})$  will be linearly homogeneous and concave in the components of  $\mathbf{P}$  and linearly homogeneous and concave in the components of  $\mathbf{x}$ ; see Chapter 3 or Diewert (1973)[74] (1974; 136)[76]. Notation:  $\mathbf{P} \cdot \mathbf{y} \equiv \sum_{m=1}^M P_m y_m$ .

the following equations for each period  $t$ :<sup>\*16</sup>

$$\mathbf{y}^t = \nabla_P g^t(\mathbf{P}^t, \mathbf{x}^t); \quad t = 1, 2. \quad (9.4)$$

Thus the period  $t$  market sector (net) supply vector  $\mathbf{y}^t$  can be obtained by differentiating the period  $t$  market sector NDP function with respect to the components of the period  $t$  output price vector  $\mathbf{P}^t$ .

If the NDP function is differentiable with respect to the components of  $\mathbf{x}$  at the point  $\mathbf{P}^t, \mathbf{x}^t$ , then the period  $t$  vector of input prices  $\mathbf{W}^t$  will be equal to the vector of first order partial derivatives of  $g^t(\mathbf{P}^t, \mathbf{x}^t)$  with respect to the components of  $\mathbf{x}$ ; i.e., we will have the following equations for each period  $t$ :<sup>\*17</sup>

$$\mathbf{W}^t = \nabla_x g^t(\mathbf{P}^t, \mathbf{x}^t); \quad t = 1, 2. \quad (9.5)$$

Thus the period  $t$  market sector input prices  $\mathbf{W}^t$  paid to primary inputs can be obtained by differentiating the period  $t$  market sector NDP function with respect to the components of the period  $t$  input quantity vector  $\mathbf{x}^t$ .

The constant returns to scale assumption on the technology sets  $S^t$  implies that the value of outputs will equal the value of inputs in period  $t$ ; i.e., we have the following relationships:

$$g^t(\mathbf{P}^t, \mathbf{x}^t) = \mathbf{P}^t \cdot \mathbf{y}^t = \mathbf{W}^t \cdot \mathbf{x}^t; \quad t = 1, 2. \quad (9.6)$$

With the above preliminaries out of the way, we can now consider a definition of a family of output indexes which will capture the idea behind Samuelson's second definition of income or net output in the previous section. Diewert (1983; 1063)[97] defined a *family of output indexes* between periods 1 and 2 for each reference output price vector  $\mathbf{P}$  as follows:<sup>\*18</sup>

$$Q(\mathbf{P}, \mathbf{x}^1, \mathbf{x}^2) \equiv g^2(\mathbf{P}, \mathbf{x}^2)/g^1(\mathbf{P}, \mathbf{x}^1). \quad (9.7)$$

Note that the above definition combines the effects of technical progress and of input growth. A *family of technical progress indexes* between periods 1 and 2 can be defined as follows for each reference input vector  $\mathbf{x}$  and each reference output price vector  $\mathbf{P}$ :<sup>\*19</sup>

$$\tau(\mathbf{P}, \mathbf{x}) \equiv g^2(\mathbf{P}, \mathbf{x})/g^1(\mathbf{P}, \mathbf{x}). \quad (9.8)$$

Thus in definition (9.8), the market sector of the economy is asked to produce the maximum output possible given the same reference vector of primary inputs  $\mathbf{x}$  and given that producers face the same reference net output price vector  $\mathbf{P}$  but in the numerator of (9.8), producers have access to the technology of period 2 whereas in the denominator of (9.8), they only have access to the technology of period 1. Hence, if  $\tau(\mathbf{P}, \mathbf{x})$  is greater than 1, there has been *technical progress* going from period 1 to 2.

A *family of input growth indexes*  $\gamma(\mathbf{P}, t, \mathbf{x}^1, \mathbf{x}^2)$  between periods 1 and 2 can be defined for each reference net output price vector  $\mathbf{P}$  and each technology indexed by the time period  $t$  as follows:<sup>\*20</sup>

$$\gamma(\mathbf{P}, t, \mathbf{x}^1, \mathbf{x}^2) \equiv g^t(\mathbf{P}, \mathbf{x}^2)/g^t(\mathbf{P}, \mathbf{x}^1). \quad (9.9)$$

<sup>\*16</sup> These relationships are due to Hotelling (1932; 594)[242]. Note that  $\nabla_P g^t(\mathbf{P}^t, \mathbf{x}^t) \equiv [\partial g^t(\mathbf{P}^t, \mathbf{x}^t)/\partial P_1, \dots, \partial g^t(\mathbf{P}^t, \mathbf{x}^t)/\partial P_M]$ .

<sup>\*17</sup> These relationships are due to Samuelson (1953)[342] and Diewert (1974; 140)[76]. Note that  $\nabla_x g^t(\mathbf{P}^t, \mathbf{x}^t) \equiv [\partial g^t(\mathbf{P}^t, \mathbf{x}^t)/\partial x_1, \dots, \partial g^t(\mathbf{P}^t, \mathbf{x}^t)/\partial x_N]$ .

<sup>\*18</sup> Diewert generalized the definitions used by Samuelson and Swamy (1974)[348] and Sato (1976; 438)[349]. Samuelson and Swamy assumed only one input and no technical change while Sato had many inputs and outputs in his model but no technical change. These authors recognized the analogy of the output quantity index with Allen's (1949)[6] definition of a quantity index in the consumer context.

<sup>\*19</sup> Definition (9.8) may be found in Diewert (1983; 1063)[97], Diewert and Morrison (1986; 662)[146] and Kohli (1990)[274].

<sup>\*20</sup> Definition (9.9) can also be found in Diewert (1983; 1063)[97].

Thus using the period  $t$  technology and the reference net output price vector  $\mathbf{P}$ , we say that there has been positive input growth going from the period 1 input quantity vector  $\mathbf{x}^1$  to the observed period 2 input quantity vector  $\mathbf{x}^2$  if  $g^t(\mathbf{P}, \mathbf{x}^2) > g^t(\mathbf{P}, \mathbf{x}^1)$  or equivalently, if  $\gamma(\mathbf{P}, t, \mathbf{x}^1, \mathbf{x}^2) > 1$ .

**Problem 1** Show that the output quantity index defined by (9.7) has the following decompositions:

$$Q(\mathbf{P}, \mathbf{x}^1, \mathbf{x}^2) = \tau(\mathbf{P}, \mathbf{x}^2)\gamma(\mathbf{P}, 1, \mathbf{x}^1, \mathbf{x}^2); \quad (\text{a})$$

$$Q(\mathbf{P}, \mathbf{x}^1, \mathbf{x}^2) = \tau(\mathbf{P}, \mathbf{x}^1)\gamma(\mathbf{P}, 2, \mathbf{x}^1, \mathbf{x}^2). \quad (\text{b})$$

Thus the output quantity index between periods 1 and 2 does combine the effects of technical progress and input growth between periods 1 and 2.

**Problem 2** We now specialize definition (9.7) to the case where the reference net output price vector is chosen to be the period 1 price vector  $\mathbf{P}^1$ , which leads to the following *Laspeyres type theoretical output quantity index*:

$$Q(\mathbf{P}^1, \mathbf{x}^1, \mathbf{x}^2) \equiv g^2(\mathbf{P}^1, \mathbf{x}^2)/g^1(\mathbf{P}^1, \mathbf{x}^1). \quad (\text{a})$$

If we choose  $\mathbf{P}$  to be the period 2 price vector  $\mathbf{P}^2$ , we obtain the following *Paasche type theoretical output quantity index*:

$$Q(\mathbf{P}^2, \mathbf{x}^1, \mathbf{x}^2) \equiv g^2(\mathbf{P}^2, \mathbf{x}^2)/g^1(\mathbf{P}^2, \mathbf{x}^1). \quad (\text{b})$$

Under assumptions (9.6) above, show that the theoretical output quantity indexes defined by (a) and (b) above satisfy the following inequalities:

$$Q(\mathbf{P}^1, \mathbf{x}^1, \mathbf{x}^2) \geq \mathbf{P}^1 \cdot \mathbf{y}^2 / \mathbf{P}^1 \cdot \mathbf{y}^1 \equiv Q_L(\mathbf{P}^1, \mathbf{P}^2, \mathbf{y}^1, \mathbf{y}^2); \quad (\text{c})$$

$$Q(\mathbf{P}^2, \mathbf{x}^1, \mathbf{x}^2) \leq \mathbf{P}^2 \cdot \mathbf{y}^2 / \mathbf{P}^2 \cdot \mathbf{y}^1 \equiv Q_P(\mathbf{P}^1, \mathbf{P}^2, \mathbf{y}^1, \mathbf{y}^2) \quad (\text{d})$$

where  $Q_L(\mathbf{P}^1, \mathbf{P}^2, \mathbf{y}^1, \mathbf{y}^2)$  and  $Q_P(\mathbf{P}^1, \mathbf{P}^2, \mathbf{y}^1, \mathbf{y}^2)$  are the observable Laspeyres and Paasche net output quantity indexes.

**Problem 3** Under what conditions will the inequalities (c) and (d) in problem 2 above hold as equalities?

**Problem 4** Is the constant returns to scale assumption required to derive the results in problems 1 and 2 above?

**Problem 5** Illustrate the two inequalities in problem 2 above using Figure 9.1; i.e., specialize  $M$  to the case  $M = 2$ , and then modify Figure 9.1 to illustrate the two inequalities in problem 2.

## 9.5 Maintaining Capital Again: the Physical versus Real Financial Perspectives

Recalling the material in section 8.2 of chapter 8 (on aggregation problems within the period; i.e., the beginning, end and middle of the period decomposition of the period) and Appendix 2 of chapter 8 (on the Austrian production function concept), we see that Samuelson's  $C + I$  framework for discussing alternative income concepts is not quite adequate to illustrate all of the problems involved in defining income concepts.

Recall that when using Samuelson's second income concept, nominal income in period 1 was defined as  $P_C^1 C^1 + P_I^1 I^1$  where  $I^1$  was defined to be net investment in period 1. Net investment can be redefined in terms of the difference between the beginning and end of period 1 capital stocks,  $K^0$

and  $K^1$ , so that  $I^1$  equals  $K^1 - K^0$ . If we substitute this definition of net into Samuelson's definition of period 1 nominal income, we obtain the following definition for *period 1 nominal income*:

$$\text{Income 1} \equiv P_C^1 C^1 + P_I^1 I^1 = P_C^1 C^1 + P_I^1 (K^1 - K^0) = P_C^1 C^1 + P_I^1 K^1 - P_I^1 K^0. \quad (9.10)$$

Note that in the above definition, the beginning and end of period capital stocks are valued at the same price,  $P_I^1$ . But this same price concept does not quite fit in with our Austrian one period production function framework where the beginning of the period capital stock should be valued at the beginning of the period opportunity cost of capital,  $P_K^0$  say, and the end of the period capital stock should be valued at the end of the period expected opportunity cost of capital,  $P_K^1$ .<sup>\*21</sup> Replacing  $P_I^1$  in (9.10) by  $P_K^1$  (for  $K^1$ ) and by  $P_K^0$  (for  $K^0$ ) leads to the following estimate for period 1 nominal income:

$$\text{Income 2} \equiv P_C^1 C^1 + P_K^1 K^1 - P_K^0 K^0. \quad (9.11)$$

But Income 2 is expressed in heterogeneous units:  $P_C^1$  reflects the average level of prices of the consumption good in period 1 whereas  $P_K^1$  reflects the price of capital at the *end* of period 1 while  $P_K^0$  reflects the price of capital at the *beginning* of period 1. The problem is that there could be a considerable amount of price change going from the beginning to the end of period 1. Hence we need to adjust the beginning of the period price of capital,  $P_K^0$ , into a comparable end of period price that eliminates the effects of inflation over the duration of period 1. There are two possible price indexes that we could use: a (capital) *specific price index*  $1 + i^1$  or a *general price index*  $1 + \rho^1$  that is based on the movement of consumer prices from the beginning of period 1 to the end of period 1; i.e., define  $i^1$  and  $\rho^1$  as follows:

$$1 + i^0 \equiv P_K^1 / P_K^0; \quad (9.12)$$

$$1 + \rho^0 \equiv P_{CE}^1 / P_{CE}^0 \quad (9.13)$$

where  $P_{CE}^1$  is the level of consumer prices at the *end* of period 1 and  $P_{CE}^0$  is the level of consumer prices at the *beginning* of period 1 or the *end* of period 0.

Now insert either  $1 + i^0$  or  $1 + \rho^0$  in front of the term  $P_K^0 K^0$  in (9.11) and we obtain the following *two income concepts that measure income from the perspective of the level of prices prevailing at the end of period 1*:

$$\text{Income 3} \equiv P_C^1 C^1 + P_K^1 K^1 - (1 + \rho^0) P_K^0 K^0; \quad (9.14)$$

$$\begin{aligned} \text{Income 4} &\equiv P_C^1 C^1 + P_K^1 K^1 - (1 + i^0) P_K^0 K^0 \\ &= P_C^1 C^1 + P_K^1 K^1 - P_K^1 K^0 \quad \text{using (9.12)} \\ &= \text{Income 1} \quad \text{using (9.10) if } P_K^1 = P_I^1. \end{aligned} \quad (9.15)$$

Thus if the end of period 1 price of capital  $P_K^1$  is equal to the period 1 investment price  $P_I^1$ , then Income 4 coincides with Income 1; i.e., if  $P_K^1 = P_I^1$ , then Income 4 ends up equaling Samuelson's Income 1. However, in general,  $P_K^1$  (then end of period 1 price of capital) will not be equal to  $P_I^1$  (the average price of capital during period 1) but in a low inflation environment, the differences between Income 1 and 4 will usually be small.

The first line in (9.15) shows that Income 4 can be interpreted as a type of *specific price level adjusted income* and (9.14) shows that Income 3 is a type of *general price level adjusted accounting*

<sup>\*21</sup> For now, we will assume that expectations are realized in order to save on notational complexity. We will return to the problem of modeling expectations later in the chapter.

*income*.<sup>\*22</sup> The idea behind the Income 3 measure defined by (9.14) is this: at the end of the period, the investors who provided financial capital to the firm have access to the end of period 1 value of the firm's capital stock,  $P_K^1 K^1$ , which they could turn into consumption equivalents if they wanted to do this. However, at the beginning of period 1, they provided financial capital to the firm in the amount  $P_K^0 K^0$ . This amount of money could be turned into consumption at the beginning of period 1 and this amount of consumption represents the opportunity cost of their investment at the beginning of the period. To measure the benefit of the investment ( $P_K^1 K^1$ ) against the cost of the investment ( $P_K^0 K^0$ ) in comparable amounts of consumption gained versus sacrificed, we need to discount  $P_K^1 K^1$  by the Consumer Price Index inflation rate over period 1,  $1 + \rho^0$ , or alternatively, escalate  $P_K^0 K^0$  by  $1 + \rho^0$ . We follow accounting conventions and escalate the beginning of the period sacrifice value to make it comparable to the end of period benefit value and so the net consumption benefit of the investment is  $P_K^1 K^1 - (1 + \rho^0)P_K^0 K^0$ . If this amount is zero or positive, then *the investor's real financial capital has been kept intact* by the firm's actions over period 1.<sup>\*23</sup>

Turning now to an interpretation of Income 4 defined by (9.15), we again start with the investor's end of period benefit of their investment in the firm,  $P_K^1 K^1$ , which again could be turned into consumption equivalents at the end of period 1. However, instead of converting the beginning of the period investment in the firm,  $P_K^0 K^0$ , into consumption forgone, we simply convert the beginning of the period price of the capital stock,  $P_K^0$ , into the corresponding end of the period price of the capital stock,  $(1 + i^0)P_K^0$ , which is equal to  $P_K^1$ . Thus instead of attempting to maintain the investor's real financial capital intact, *we now attempt to maintain the firm's physical stock of capital in use intact* (at end of period prices). This type of accounting adjustment is called the *specific price level method* for constructing current values for an asset held by a business unit. The method was suggested by Daines (1929; 101)[62], Sweeney (1934; 110)[365] and many other accountants.<sup>\*24</sup>

Since Income 1 does not fit into the Hicks and Edwards and Bell one period production function framework where beginning of the period capital is regarded as an input and end of the period capital is regarded as an output, we will not consider Income 1 any further. Moreover, we also have rejected Income 2 since it does not adjust for general inflation over the course of period 1. Hence we are left with Incomes 3 and 4 and the question is: how do we choose between Income 3 and Income 4? We will address this question in the following section.

**Problem 6** Refer back to Samuelson's Method 2 in section 9.3. Use Income 3 in place of Samuelson's income measure and construct the Laspeyres and Paasche measures of income growth going from period 1 to 2. Are there any potential problems due to the fact that not all components of Income 3 have positive signs?

## 9.6 Measuring Business Income: the End of the Period Perspective

In order to see if one of the income concepts explained in the previous sections of this chapter can emerge as being the "right" concept, we will return to the one period profit maximization problem

<sup>\*22</sup> These types of balance sheet adjustments for inflation over an accounting period are discussed in the inflation accounting literature; e.g., see Middleditch (1918)[311], Sweeney (1934)[365] (1935)[366] (1964)[367], Edwards and Bell (1961)[168], Baxter (1975)[28], Sterling (1975)[364], Whittington (1980)[400], Carsberg (1982)[47] and Tweedie and Whittington (1984)[377].

<sup>\*23</sup> Keeping financial capital intact does not include interest payments that typically must be made to investors in order for them to postpone consumption. We will see how interest gets into the picture later.

<sup>\*24</sup> "Inasmuch as the price level is not stable for any great length of time, and since this calculation is contemplated for each fiscal period, the only feasible procedure for a company with thousands of assets is the use of price index numbers." Albert L. Bell (1953; 49)[34]. "Where no market exists for new fixed assets of the type used by the firm, two means of measuring current costs are available: (1) appraisal, and (2) the use of price index numbers for like fixed assets to adjust the original cost base to the level which would now have to be paid to purchase the asset in question." Edgar O. Edwards and Philip W. Bell (1961; 186)[168].

of the market sector of the economy using the Austrian one period production function framework explained in Appendix 2 of chapter 8.\*<sup>25</sup>

Using the notation introduced in the previous section and adding labour  $L$  as an input (with price  $W$ ) and letting the market sector of the economy face the beginning of period 1 nominal interest rate  $r^0$ , the period 1 Austrian profit maximization problem can be defined as follows:

$$\max_{C^1, L^1, K^1} \{(1 + r^0)^{-1}(P_C^1 C^1 - W^1 L^1 + P_K^1 K^1) - P_K^0 K^0 : (C^1, L^1, K^0, K^1) \in S^1\} \quad (9.16)$$

where  $S^1$  is the period 1 Austrian production possibilities set. Note that we have treated the price  $P_C^1$  of period 1 consumption and the price of period 1 labour  $W^1$  as end of period 1 prices and hence the corresponding value flows are discounted to their beginning of period 1 equivalents using the beginning of period 1 nominal interest rate  $r^0$ . From a practical measurement perspective, it is more useful to work with end of the period equivalents and so if we multiply the objective function in (9.16) through by  $(1 + r^0)$ , we obtain the following *period 1 (end of period perspective) profit maximization problem*:

$$\max_{C^1, L^1, K^1} \{P_C^1 C^1 - W^1 L^1 + P_K^1 K^1 - (1 + r^0)P_K^0 K^0 : (C^1, L^1, K^0, K^1) \in S^1\}. \quad (9.17)$$

Recall equation (9.12) above,  $1 + i^0 \equiv P_K^1/P_K^0$ , which defined the period 1 asset specific inflation rate  $i^0$ , and equation (9.13) above,  $1 + \rho^0 \equiv P_{CE}^1/P_{CE}^0$ , which defined the period 1 general inflation rate. The period 1 general inflation rate,  $\rho^0$ , can be used to define the beginning of period 1 *real interest rate*  $r^{0*}$  and the period 1 *real rate of asset price inflation*  $i^{0*}$  as follows:

$$1 + r^{0*} \equiv (1 + r^0)/(1 + \rho^0). \quad (9.18)$$

$$1 + i^{0*} \equiv (1 + i^0)/(1 + \rho^0). \quad (9.19)$$

Now substitute (9.18) into the objective function in (9.17) and we obtain the following expression for period 1 *pure profits*:

$$\begin{aligned} & P_C^1 C^1 - W^1 L^1 + P_K^1 K^1 - (1 + r^0)P_K^0 K^0 \\ &= P_C^1 C^1 - W^1 L^1 + P_K^1 K^1 - (1 + r^{0*})(1 + \rho^0)P_K^0 K^0 \\ &= P_C^1 C^1 + P_K^1 K^1 - (1 + \rho^0)P_K^0 K^0 - \{W^1 L^1 + r^{0*}(1 + \rho^0)P_K^0 K^0\} \\ &= \text{Income 3} - \{W^1 L^1 + U^1 K^0\} \end{aligned} \quad (9.20)$$

where Income 3 was defined by (9.14) in the previous section and the *period 1 waiting services user cost of the initial capital stock*\*<sup>26</sup> is defined as

$$U^1 \equiv r^{0*}(1 + \rho^0)P_K^0. \quad (9.21)$$

With a constant returns to scale technology, competitive pricing on the part of market sector producers and correct expectations, pure profits will be zero and hence (9.20) set equal to zero will give

\*<sup>25</sup> Recall that this framework is based on Hicks (1961; 23)[223] and Edwards and Bell (1961; 71-72)[168]. Their work is related to the earlier work of Böhm-Bawerk (1891)[40], von Neumann (1937)[385], Hicks (1946; 230)[222] and Malinvaud (1953)[299] and the later work of Diewert (1977)[84] (1980; 472-474)[90].

\*<sup>26</sup> Rymes (1968)[338] (1983)[339] stressed waiting services as a primary input.

us the following equations:<sup>\*27</sup>

$$\begin{aligned} \text{Income 3} &= P_C^1 C^1 + P_K^1 K^1 - (1 + \rho^0) P_K^0 K^0 \\ &= W^1 L^1 + U^1 K^0 \end{aligned} \quad (9.22)$$

where  $W^1 L^1$  represents period 1 payments to labour and  $U^1 K^0$  represents interest payments to holders of the initial capital stock in terms of end of period 1 dollars. Note that all prices in (9.22) are expressed in end of period 1 equivalents.

What is the significance of equation (9.22)? This equation seems to suggest that Income 3 is the “right” concept of net output for period 1!<sup>\*28</sup> However, we shall see later see that Income 4 is also consistent with the Austrian production function framework.

At this point, the reader may be slightly confused and may well ask: what happened to our usual user cost formula? The user cost  $U^1$  defined by (9.21) does not look very familiar and so there might be a suspicion that something might be wrong with the above algebra. In order to address this issue, we will specialize the Austrian model to the usual production function model, defined as follows:

$$C^1 = F(I_G^1, L^1, K^0); \quad K^1 = (1 - \delta)K^0 + I_G^1 \quad (9.23)$$

where  $I_G^1$  is gross investment in period 1,  $C^1$  is period 1 consumption output,  $L^1$  is period 1 labour input,  $K^0$  is the start of period 1 capital stock,  $K^1$  is the end of period 1 finishing capital stock,  $0 < \delta < 1$  is the constant (geometric) physical depreciation rate and  $F$  is the production function, which is decreasing in  $I_G$  and increasing in  $L$  and  $K$ .

If we substitute (9.23) into the objective function in (9.17) and solve the resulting period 1 profit maximization problem, we find that the optimal objective function can be written as follows:

$$\begin{aligned} &P_C^1 C^1 - W^1 L^1 + P_K^1 K^1 - (1 + r^0) P_K^0 K^0 \quad (9.24) \\ &= P_C^1 C^1 - W^1 L^1 + P_K^1 [(1 - \delta)K^0 + I_G^1] - (1 + r^{0*})(1 + \rho^0) P_K^0 K^0 \quad \text{using (9.18)} \\ &= P_C^1 C^1 + P_K^1 I_G^1 - W^1 L^1 - (1 + r^{0*})(1 + \rho^0) P_K^0 K^0 + (1 - \delta) P_K^1 K^0 \\ &= P_C^1 C^1 + P_K^1 I_G^1 - W^1 L^1 - (1 + r^{0*})(1 + \rho^0) P_K^0 K^0 + (1 - \delta)(1 + i^0) P_K^0 K^0 \quad \text{using (9.12)} \\ &= P_C^1 C^1 + P_K^1 I_G^1 - W^1 L^1 - (1 + r^{0*})(1 + \rho^0) P_K^0 K^0 + (1 - \delta)(1 + \rho^0)(1 + i^{0*}) P_K^0 K^0 \\ &\quad \text{using (9.19)} \\ &= P_C^1 C^1 + P_K^1 I_G^1 - W^1 L^1 - [(1 + r^{0*})(1 + \rho^0) - (1 - \delta)(1 + \rho^0)(1 + i^{0*})] P_K^0 K^0. \end{aligned} \quad (9.25)$$

The term in square brackets in (9.25) times  $P_K^0$  represents the usual (end of period) gross rental user cost of capital  $u^1$ ; i.e., we have<sup>\*29</sup>

$$u^1 \equiv [(1 + r^{0*})(1 + \rho^0) - (1 - \delta)(1 + \rho^0)(1 + i^{0*})] P_K^0 = [(1 + r^0) - (1 - \delta)(1 + i^0)] P_K^0. \quad (9.26)$$

<sup>\*27</sup> To form a net investment aggregate in this framework, we aggregate over the value difference  $P_K^1 K^1 - (1 + \rho^0) P_K^0 K^0$  using normal index number theory provided that this value aggregate is bounded away from 0 over all periods; i.e., we use normal index number theory, with  $K^1$  a positive quantity with the corresponding price  $P_K^1$  and  $-K^0$  as a negative quantity with price  $(1 + \rho^0) P_K^0$ . If the value aggregate approaches or passes through 0 during any period, then we cannot form a net investment aggregate for this capital component; i.e., we would have to combine the value aggregate  $P_K^1 K^1 - (1 + \rho^0) P_K^0 K^0$  with an additional substantially positive value aggregate.

<sup>\*28</sup> Note that our Income 3 follows the adjustments to cash flows recommended by the accountant Sterling: “It follows that the appropriate procedure is to (1) adjust the present statement to current values and (2) adjust the previous statement by a price index. It is important to recognize that *both* adjustments are necessary and that neither is a substitute for the other. Confusion on this point is widespread.” Robert R. Sterling (1975; 51)[364]. Sterling (1975; 50)[364] termed his income concept *Price Level Adjusted Current Value Income*. Unfortunately, Sterling’s income concept has not been widely endorsed in accounting circles (but it should be).

<sup>\*29</sup> See section 8.3 of chapter 8.

Thus if pure profits are zero for the period 1 data, expression (9.25) set equal to 0 translates into the following usual gross output equals labour payments plus gross payments to the starting stock of capital:

$$P_C^1 C^1 + P_K^1 I_G^1 = W^1 L^1 + u^1 K^0. \quad (9.27)$$

We now show that  $u^1$  can be expressed as the sum of three terms where the terms are defined as follows:

$$U^1 \equiv r^{0*}(1 + \rho^0)P_K^0; \quad (9.28)$$

$$D^1 \equiv \delta(1 + i^{0*})(1 + \rho^0)P_K^0 = \delta P_K^1; \quad (9.29)$$

$$R^1 \equiv -i^{0*}(1 + \rho^0)P_K^0. \quad (9.30)$$

**Problem 7** Show that the usual gross rentals user cost formula  $u^1$  defined above by (9.26) can be written as the sum of the three terms defined by (9.28)-(9.30); i.e., show that

$$u^1 = U^1 + D^1 + R^1. \quad (9.31)$$

We now need to provide economic interpretations for the three terms on the right hand side of (9.31). It can be seen that  $U^1$  defined by (9.28) is the same definition as the  $U^1$  defined by (9.21) and we interpreted this  $U^1$  as a *real waiting services user cost* for the initial beginning of the period capital stock  $K^0$ . Obviously,  $\delta P_K^0 K^0$  can be interpreted as the amount of wear and tear depreciation that the initial capital stock will undergo during the period. However, this amount of depreciation is expressed in the beginning of the period price of the capital stock,  $P_K^0$ . In keeping with our conventions, we convert this beginning of the period price into its consumption equivalent price at the end of the period by multiplying by  $1 + \rho^0$ . Thus  $D^1 K^0$  is equal to  $\delta(1 + i^{0*})(1 + \rho^0)P_K^0 K^0 = \delta P_K^1 K^0$  and can be interpreted as the *real value of wear and tear depreciation*, expressed in end of period consumption equivalents. Finally,  $R^1$  is a *real revaluation term*; if the real asset inflation rate  $i^{0*}$  is negative, then  $R^1 K^0$  can be interpreted as an obsolescence charge; i.e., the rate of nominal asset price inflation  $i^0$  is less than the general nominal inflation rate  $\rho^0$  and so an extra charge for the use of the asset in period 0 must be made in addition to normal wear and tear depreciation.<sup>\*30</sup> On the other hand, if the real asset inflation rate  $i^{0*}$  is positive, then  $R^1 K^0 = -i^{0*}(1 + \rho^0)P_K^0 K^0 < 0$  can be interpreted as an offset to the wear and tear depreciation charge  $D^1 K^0$ . This offset is due to the fact that the firm “transports” capital from a time period where it is less valuable in real terms (the beginning of period 0) to a time period where the capital is more highly valued (the end of period 0).

The (real) decomposition of the user cost of capital  $u^1$  defined by (9.31) and the three definitions (9.28)-(9.30) seems a bit awkward compared to the following more straightforward (nominal) decomposition of the user cost:

$$\begin{aligned} u^1 &\equiv [(1 + r^0) - (1 - \delta)(1 + i^0)]P_K^0 \\ &= [r^0 - i^0 + \delta(1 + i^0)]P_K^0. \end{aligned} \quad (9.32)$$

The nominal waiting services part of  $u^1$  is obviously  $r^0 P_K^0$ , the nominal revaluation term is  $-i^0 P_K^0$  and the nominal wear and tear depreciation term is  $\delta(1 + i^0)P_K^0$ . The problem with the user cost decomposition given by (9.32) is that *it does not readily integrate with the two main income concepts that we defined earlier*.

We now show how the decomposition of the gross user cost  $u^1$  defined by (9.28)-(9.31) is related to Income 4 and Income 3 defined earlier. We do this first for Income 3. Substitute (9.31) into (9.27)

<sup>\*30</sup> The sum of the wear and tear depreciation term and the revaluation term is called real time series depreciation by Diewert (2005a)[128] (2006b)[132] and it formalizes a definition due to Hill (2000)[230].

and we obtain the following equation:

$$P_C^1 C^1 + P_K^1 I_G^1 = W^1 L^1 + [U^1 + D^1 + R^1] K^0. \quad (9.33)$$

Now subtract  $[D^1 + R^1] K^0$  from both sides of (9.33) and we obtain the following equation:

$$\begin{aligned} P_C^1 C^1 + P_K^1 I_G^1 - D^1 K^0 - R^1 K^0 &= W^1 L^1 + U^1 K^0 \\ &= \text{Income 3} \quad \text{using (9.22)}. \end{aligned} \quad (9.34)$$

Equations (9.34) are simply our earlier equations (9.22) when we substitute (9.23) and other equations into (9.22). Net investment in this model can be viewed as an aggregate of gross investment less real wear and tear depreciation less real revaluations.<sup>\*31</sup>

As noted earlier, the net output that is generated by the left hand side of (9.34) can be interpreted as an income concept that maintains real financial capital. Thus at this point, we might tentatively conclude that working with the usual discounted profits model leads to *a preference for a maintenance of real financial capital income concept over a maintenance of a specific inflation adjusted income concept*, at least at the theoretical level. However, this tentative conclusion is not correct:<sup>\*32</sup> it turns out that we can manipulate the Austrian discounted profits model in a way that will justify the maintenance of real physical capital as opposed to real financial capital. We show this in the following paragraphs.

We establish our result for the general Austrian capital model; i.e., the model that was defined before we specialized the model in equations (9.23). We first use the definition of  $R^1$  to establish the following identity:

$$\begin{aligned} (1 + \rho^0) P_K^0 K^0 - (1 + i^0) P_K^0 K^0 &= (1 + \rho^0) P_K^0 K^0 - (1 + i^{0*}) (1 + \rho^0) P_K^0 K^0 \quad \text{using (9.19)} \\ &= -i^{0*} (1 + \rho^0) P_K^0 K^0 \\ &= R^1 K^0 \quad \text{using (9.30)}. \end{aligned} \quad (9.35)$$

Using definitions (9.14) and (9.15) for Incomes 3 and 1 respectively, we see that if we add the left hand side of (9.35) to Income 3, we will obtain Income 4. Hence adding the right hand side of (9.35) to Income 3 will give us Income 4. Thus we have:

$$\begin{aligned} \text{Income 4} &\equiv P_C^1 C^1 + P_K^1 K^1 - (1 + i^0) P_K^0 K^0 \\ &= \text{Income 3} + R^1 K^0 \\ &= W^1 L^1 + U^1 K^0 + R^1 K^0 \quad \text{using (9.22)}. \end{aligned} \quad (9.36)$$

*Thus if we adopt a physical maintenance of capital point of view to measure income, the matching user cost for the beginning of the period stock of capital  $K^0$  is  $U^1 + R^1$ , the sum of the real waiting services and revaluation terms.*

If we now specialize the general Austrian model to the geometric depreciation model and substitute (9.12) and (9.23) into the first equation in (9.36), we obtain the following expression for Income 4:

$$\begin{aligned} \text{Income 4} &\equiv P_C^1 C^1 + P_K^1 K^1 - (1 + i^0) P_K^0 K^0 \\ &= P_C^1 C^1 + P_K^1 [I_G^1 + (1 - \delta) K^0] - P_K^1 K^0 \quad \text{using (9.12) and (9.23)} \\ &= P_C^1 C^1 + P_K^1 [I_G^1 - \delta K^0] \end{aligned} \quad (9.37)$$

<sup>\*31</sup> Normal index number theory can be used to aggregate the three terms, provided that gross investment is always larger than the sum of depreciation and revaluation; i.e., treat all three prices as positive, the first quantity as positive and the next two quantities as negative numbers in the index number formula.

<sup>\*32</sup> I owe this point to Paul Schreyer.

where  $I_G^1 - \delta K^0$  is gross investment less wear and tear depreciation so that this difference can clearly be interpreted as net investment, with corresponding price equal to the end of period price of a new investment good,  $P_K^1$ .

Thus the Austrian production framework is consistent with both income concepts: the financial maintenance of capital concept (Income 3) and the physical maintenance of capital concept (Income 1 or 4). In fact, (9.27) shows that the Austrian production framework is also consistent with an “income” concept that is equal to gross product. Table 9.1 below summarizes the definitions of the different income concepts in the case of geometric depreciation and gives the user cost concept that matches up with the corresponding income concept.

Table 9.1 Alternative Income Concepts and the Corresponding User Costs of Capital

Income Concept	Corresponding Net Output Definition	Corresponding User Cost Value
Gross Output	$P_C^1 C^1 + P_K^1 I_G^1$	$U^1 K^0 + D^1 K^0 + R^1 K^0 = u^1 K^0$
Income 4	$P_C^1 C^1 + P_K^1 I_G^1 - D^1 K^0$	$U^1 K^0 + R^1 K^0$
Income 3	$P_C^1 C^1 + P_K^1 I_G^1 - D^1 K^0 - R^1 K^0$	$U^1 K^0$

Looking at Table 9.1, it can be seen that the usual gross output (or GDP) definition of “income” matches up with the usual gross rentals user cost of capital,  $u^1$ . The Income 4 definition, which corresponds to a physical maintenance of capital income concept, takes the physical depreciation term  $D^1 K^0$  out of the gross rentals user cost and treats it as a negative contribution to output. Finally, the Income 3 definition, which corresponds to a maintenance of real financial capital concept, takes both the physical depreciation term  $D^1 K^0$  and the revaluation or obsolescence term  $R^1 K^0$  out of the gross rentals user cost and treats them both as a negative contributions to output. Thus the Austrian model of production is consistent with all three income concepts.

Typically, the price of capital goods will decline relative to the price of consumption goods and services (or will increase at a lower rate) so that the real asset inflation rate,  $i^{0*}$  defined by (9.19), will usually be negative. Under this hypothesis,  $R^1$  will be positive<sup>\*33</sup> and we will have the following inequalities between the three income concepts:

$$\text{Gross Income} > \text{Income 4} > \text{Income 3.} \quad (9.38)$$

We conclude this section with a brief discussion on which income concept is “best” from the perspective of describing household consumption possibilities over time. The gross income concept clearly overstates long run consumption for the consumer and so this concept can be dismissed. It is clear that Income 2 is a very defective measure of sustainable consumption prospects, since this measure can be made very large if there is a substantial amount of inflation between the beginning and end of the period. All of the other income measures are invariant to general inflation between the beginning and end of the accounting period.<sup>\*34</sup> However, choosing between the physical and real financial maintenance perspectives is more problematical: reasonable economists could differ on this choice. The merits of the two perspectives were discussed by Pigou and Hayek over 60 years ago. Pigou (1941; 273-274)[329] favored the maintenance of physical capital approach (Incomes 1 or 4) while Hayek (1941; 276-277)[218] favored the maintenance of real financial capital approach (Income 3). Our preference is for Income 3, following in the footsteps of Hayek and Hill (2000; 6)[230], who felt that Income 1 or 4 would generally overstate the real value of consumption in any period due to its neglect of (foreseen) obsolescence (due to expected real price decreases in the asset). Conversely, if real price increases in the asset are foreseen, then the revaluation term can be regarded as a positive contribution to the net revenues produced by the production unit under consideration; i.e., the unit

<sup>\*33</sup> Under these conditions, we can say that the capital good is experiencing a form of obsolescence.

<sup>\*34</sup> We regard this invariance property as a fundamental property that any sensible income measure should satisfy.

“transports” the asset from a time when it is less valued (in real terms) to a time when it is more highly valued. However, it is certainly possible to argue in favour of the physical maintenance of capital concept of income.

To summarize: there are two ways that can be used to justify the Haig Marshall Pigou Samuelson Income 1 or 4 versus the Hayek Sterling Hill Income 3 (or Price level Adjusted Current Value Income to use Sterling’s (1975; 50)[364] terminology):

- One can look at income from the output side perspective and think about the relative merits of preserving physical capital versus real financial capital or
- One can look at income from the primary input side perspective and ask whether the (anticipated) revaluation term is a source of primary income or not.

As Hicks (1939; 184)[219] said in his Income chapter: “What a tricky business this all is!”

## 9.7 Approximations to the Income Concept

We now relax the perfect foresight assumptions that were made in the previous section. This means that  $i^0$  and  $\rho^0$  defined by (9.12) and (9.13) are not known variables; rather  $\rho^0$  is the anticipated (Consumer Price Index) nominal inflation rate for consumption goods and services and  $i^0$  is the anticipated specific asset inflation rate, where the anticipations are formed at the start of period 1. This means that the period 1 real interest rate  $r^{0*}$  and real asset inflation rate  $i^{0*}$  defined by (9.18) and (9.19) are also anticipated variables. Thus we must now address the question as to how exactly these anticipated variables will be estimated in empirical applications of the income concept.\*<sup>35</sup>

We will consider three alternative methods for approximating these anticipated variables but the reader will be able to construct many additional approximations, depending on the purpose at hand.

### Approximation Method 1:

The two assumptions that are made in order to implement this first method are the following ones:

- Approximate general inflation adjusted beginning of the period price of capital,  $(1 + \rho^0)P_K^0$  by the period 1 average price for the corresponding investment good  $P_I^1$  and
- Set the anticipated specific asset real inflation rate  $i^{0*}$  equal to zero.

Thus we make the following assumptions:

$$(1 + \rho^0)P_K^0 = P_I^1; \quad (9.39)$$

$$i^{0*} = 0. \quad (9.40)$$

With the above two assumptions, we find that the waiting user cost of capital  $U^1$  defined by (9.21) or (9.28), the gross rentals user cost of capital  $u^1$  defined by (9.26) and the wear and tear depreciation term  $D^1$  defined by (9.29) simplify as follows:

$$U^1 \equiv r^{0*}(1 + \rho^0)P_K^0 = r^{0*}P_I^1; \quad (9.41)$$

$$u^1 \equiv [(1 + r^{0*})(1 + \rho^0) - (1 - \delta)(1 + \rho^0)(1 + i^{0*})]P_K^0 = [r^{0*} + \delta]P_I^1; \quad (9.42)$$

$$D^1 \equiv \delta(1 + i^{0*})(1 + \rho^0)P_K^0 = \delta P_K^1 = \delta P_I^1. \quad (9.43)$$

The only remaining approximation issue is the approximation for the anticipated period 1 real interest rate  $r^{0*}$ . There are three obvious possible choices for  $r^{0*}$ :

\*<sup>35</sup> The approximations that we suggest below can be used to implement either Gross Income, Income 1 or Income 3, depending on the user’s choice of income concept.

- Calculate the balancing real rate of return that will make the profits of the production unit under consideration equal to zero;<sup>\*36</sup>
- Smooth past balancing real rates of return and use the smoothed rate as the predicted rate; or
- Simply pick a plausible constant real rate of return, such as 3 or 4 percent (for annual data).

This method of approximation should be appealing to national income accountants.

#### Approximation Method 2:

One problem with the previous method is that it neglects *foreseen obsolescence* that is due to anticipated (real) asset price decline.<sup>\*37</sup> For example, for the past 40 years, the real price of computers in constant quality units has steadily declined and these declines are very likely to continue. Hence, the  $i^{0*}$  term in our definition of the real depreciation term,  $D^1$  defined by (9.29) and in the real revaluation term  $R^1$  defined by (9.30), should be a negative number if the asset under consideration is computers (or any related asset that has a substantial computer chip component). Thus again make assumption (9.39) above but now for assets that are expected to decline in price, estimate  $i^{0*}$  by smoothed past real declines in the asset price (expressed in constant quality units). With these assumptions, we find that the real waiting user cost of capital  $U^1$  defined by (9.21) or (9.28), the gross rentals user cost of capital  $u^1$  defined by (9.26) and  $R^1$  and  $D^1$  simplify as follows:

$$U^1 \equiv r^{0*}(1 + \rho^0)P_K^0 = r^{0*}P_I^1; \quad (9.44)$$

$$u^1 \equiv [(1 + r^{0*})(1 + \rho^0) - (1 - \delta)(1 + \rho^0)(1 + i^{0*})]P_K^0 = [r^{0*} + \delta - (1 - \delta)i^{0*}]P_I^1; \quad (9.45)$$

$$D^1 \equiv \delta(1 + i^{0*})(1 + \rho^0)P_K^0 = \delta(1 + i^{0*})P_I^1; \quad (9.46)$$

$$R^1 \equiv -i^{0*}(1 + \rho^0)P_K^0 = -i^{0*}P_I^1 \quad (9.47)$$

Examining (9.45)-(9.47), it can be seen that if  $i^{0*}$  is negative, then the gross rental user cost  $u^1$  and the real revaluation term  $R^1$  term become *larger* compared to the corresponding values when  $i^{0*}$  is set equal to 0, while the real wear and tear depreciation term  $D^1$  becomes smaller. Once an estimate for  $i^{0*}$  has been obtained,  $r^{0*}$  can be estimated using any of the three methods outlined under Approximation Method 1 above.

However some assets have a long history of real price appreciation (e.g., urban land) and so the question is: for those assets which we expect to appreciate in real terms (i.e.,  $i^{0*}$  is expected to be positive instead of negative), should we insert these positive expected values into the user cost terms defined by (9.45)-(9.47)?

Symmetry suggests that the answer to the above question is yes.<sup>\*38</sup> However, note that if  $i^{0*}$  is large and positive enough, then it could lead to the gross rental price  $u^1$  defined by (9.45) being *negative*. It is not plausible that expected gross rentals be negative, since under these conditions, we would expect the corresponding asset price to be immediately be bid up to eliminate this negative expected rental price. Alternatively, if we regard the gross rental user cost as an approximation to an actual market rental rate for the asset, then since usually rental rates are positive, it is not plausible to approximate this market rental rate by a negative number. Thus some caution is called for when simply inserting a smoothed value of past real rates of asset price appreciation  $i^{0*}$  into the formulae (9.45)-(9.47): we do not want to insert a value of  $i^{0*}$  that is so large that it makes  $u^1$  negative.

<sup>\*36</sup> Substitute  $u^1$  defined by (9.44) into equation (9.28) and solve the resulting equation for the balancing real rate  $r^{0*}$ .

<sup>\*37</sup> Not all foreseen obsolescence is due to expected future real price declines; i.e., some models of obsolescence imply contracting asset lives.

<sup>\*38</sup> See the discussion on the basic forms of productive activity in section 4 of Diewert (2006a)[131] where it was concluded that anticipated capital gains were productive.

Hence we want  $i^{0*}$  to satisfy the following inequality:

$$i^{0*} < (1 - \delta)^{-1}[r^{0*} + \delta]. \quad (9.48)$$

In practical applications of this method, we suggest that for most assets, the assumption that the corresponding anticipated real inflation  $i^{0*}$  is zero is appropriate. Only in exceptional cases where we are fairly certain that producers are anticipating real capital gains or losses should we insert a nonzero  $i^{0*}$  into formulae (9.45)-(9.47).

### Approximation Method 3:

Approximation Methods 1 and 2 explained above are suitable for applications of the income concept at the national economy or industry levels where current prices for each class of asset can be obtained.<sup>\*39</sup> However, these methods are not usually suitable for applications at the level of the individual firm or enterprise, because current objective and replicable prices for each asset used by the enterprise will generally not be available. Hence the question arises: how can we approximate the income concept at the firm level?

The simplest and most useful assumptions in this context are the following ones:

$$(1 + \rho^0) = P_C^1/P_C^0; \quad (9.49)$$

$$i^{0*} = 0. \quad (9.50)$$

Thus we set  $1 + \rho^0$  equal to the ex post amount of Consumer Price Index inflation that occurred from the beginning to the end of the accounting period and assume that specific real asset price inflation is 0.

With these assumptions, we find that the user cost components defined in equations (9.44)-(9.47) simplify as follows:

$$U^1 \equiv r^{0*}(1 + \rho^0)P_K^0 = r^{0*}(1 + \rho^0)P_K^0; \quad (9.51)$$

$$u^1 \equiv [(1 + r^{0*})(1 + \rho^0) - (1 - \delta)(1 + \rho^0)(1 + i^{0*})]P_K^0 = [r^{0*} + \delta](1 + \rho^0)P_K^0; \quad (9.52)$$

$$D^1 \equiv \delta(1 + i^{0*})(1 + \rho^0)P_K^0 = \delta(1 + \rho^0)P_K^0; \quad (9.53)$$

$$R^1 \equiv -i^{0*}(1 + \rho^0)P_K^0 = 0. \quad (9.54)$$

The net effect of the above assumptions is that we can basically use historical cost accounting, except that historical cost depreciation allowances should be escalated each accounting period by the amount of CPI inflation that occurred over the period; i.e., our present approximations lead to *purchasing power adjusted historical cost accounting*, see section 3 in Diewert (2005b)[130].

The accounting profession is unlikely to embrace the above very simple and straightforward accounting adjustments to historical cost depreciation but it is important that the tax authorities recognize the importance of indexing depreciation allowances for general inflation. Using the notation developed in section 9.6 above and taking Income 3 as our desired income concept, *taxable income* should be defined as follows:

$$\text{Taxable income} \equiv P_C^1 C^1 + P_K^1 I_G^1 - D^1 K^0 - R^1 K^0 \quad (9.55)$$

where  $D^1$  and  $R^1$  are defined by (9.29) and (9.30).<sup>\*40</sup> With this definition of taxable income, the real return to capital will be taxed and the ‘unjust’ taxation of inflation inflated ‘profits’ will be prevented.

<sup>\*39</sup> Typically, we will have to rely on national statistical agency index numbers for estimates of current asset prices.

<sup>\*40</sup> If we wanted the business income tax to fall on pure profits or rents (rather than on the gross return to capital), we would also subtract  $U^1 K^0$  equal to  $r^{0*}(1 + \rho^0)P_K^0 K^0$  from the right hand side of (9.55) where  $r^{0*}$  would be a ‘normal’ real rate of return to capital that would be chosen by the tax authorities. The resulting system of business income taxation would lead to minimal deadweight loss.

For additional material on the role of expectations in income measures, the reader is referred to Hill and Hill (2003)[237].

## 9.8 Choosing an Income Concept: A Summary

Table 9.1 in section 9.6 presented 3 “income” or output concepts:

- Gross output;
- Income 1 or 4 or “wear and tear” adjusted net product<sup>\*41</sup> and
- Income 3 or “wear and tear” and “anticipated revaluation” adjusted net product.<sup>\*42</sup>

Table 9.1 also indicated that the “traditional” user cost of capital (which approximates a market rental rate for the services of a capital input for the accounting period),  $u^1$ , consists of three additive terms; i.e., we have:

$$u^1 = U^1 + D^1 + R^1 \quad (9.56)$$

where  $U^1$  is the reward for waiting term (interest rate term),  $D^1$  is the cross sectional depreciation term (or wear and tear depreciation term) and  $R^1$  is the anticipated revaluation term, which can be interpreted as an obsolescence charge if the asset is anticipated to fall in price over the accounting period. The Gross output income concept corresponds to the traditional user cost term  $u^1$ . This income measure can be used as an approximate indicator of short run production potential. However, it is not suitable for use as an indicator of sustainable consumption. In order to obtain indicators of sustainable consumption, we turn to Incomes 1 or 4 and 3.

To obtain Income 4, we simply take the wear and tear component of the traditional user cost,  $D^1$ , times the beginning of period corresponding capital stock,  $K^0$ , out of the primary input category and treat it as a negative offset to the period’s gross investment. The resulting Income 4 can be interpreted to be consistent with the position of Pigou (1941)[329], who argued against including any kind of revaluation effect in an income concept. This position can also be interpreted as a *maintenance of physical capital approach* to income measurement. In terms of the Hicks (1961)[223] and Edwards and Bell (1961)[168] Austrian production model, capital at the beginning and end of the period ( $K^0$  and  $K^1$  respectively) are both valued at the end of period stock price for a unit of capital,  $P_K^1$ , and the contribution of capital accumulation to period income is simply the difference between the end of period value of the capital stock and the beginning of the period value (valued at end of period prices),  $P_K^1 K^1 - P_K^1 K^0$ .<sup>\*43</sup> This difference between end and beginning of period values for the capital stock can be converted into consumption equivalents and then can be added to actual period 1 consumption in order to obtain Income 1. This income concept is certainly defensible.

To obtain Income 3, we subtract both wear and tear depreciation from gross output,  $D^1 K^0$ , as well as the revaluation term,  $R^1 K^0$ , and treat both of these terms as negative offsets to the period’s gross investment. The resulting Income 3 can be interpreted to be consistent with the position of Hayek (1941)[218], Sterling (1975)[364] and Hill (2000)[230]. This position can also be interpreted as a *maintenance of real financial capital approach* to income measurement. In terms of the Hicks (1961)[223] and Edwards and Bell (1961)[168] Austrian production model, capital stocks at the beginning of the period and end of the period are valued at the prices prevailing at the beginning and the end of the period,<sup>\*44</sup>  $P_K^0$  and  $P_K^1$  respectively, and then these beginning and end of period values of the capital stock are converted into consumption equivalents and then differenced. Thus the

<sup>\*41</sup> We can associate this income concept with Marshall (1890)[304], Haig (1921)[207], Pigou (1941)[329] and Samuelson (1961)[345].

<sup>\*42</sup> We can associate this income concept with Hayek (1941)[218], Sterling (1975)[364] and Hill (2000)[230].

<sup>\*43</sup> Using Samuelson’s (1961)[345] Figure 9.1 above, this income can be interpreted as the distance OJ along the consumption axis.

<sup>\*44</sup> Strictly speaking, the end of period price is an expected end of period price.

end of the period value of the capital stock is  $P_K^1 K^1$  and this value can be converted into consumption equivalents at the consumption prices prevailing at the end of the period. The beginning of the period value of the capital stock is  $P_K^0 K^0$  but to convert this value into consumption equivalents at end of period prices, we must multiply this value by  $(1 + \rho^0)$ , which is one plus the rate of consumer price inflation over the period. This price level adjusted difference between end and beginning of period values for the capital stock,  $P_K^1 K^1 - (1 + \rho^0)P_K^0 K^0$ , can be converted into consumption equivalents and then can be added to actual period 1 consumption in order to obtain Income 3. Thus the difference between Income 1 and Income 3 can be viewed as follows: Income 1 uses the end of period stock price of capital to value both the beginning and end of period capital stocks and then converts the resulting difference in values into consumption equivalents at the prices prevailing at the end of the period whereas Income 3 values beginning and end of period capital stocks at the stock prices prevailing at the beginning and end of the period and *directly* converts these values into consumption equivalents and then adds the difference in these consumption equivalents to actual consumption. Thus Income 3 also seems to be a defensible concept.\*<sup>45</sup>

In order to highlight the difference between Incomes 4 and 3, use definitions (9.10), (9.12) and (9.14) in order to compute their difference:

$$\begin{aligned} \text{Income 4} - \text{Income 3} &= P_C^1 C^1 + P_I^1 K^1 - P_I^1 K^0 - [P_C^1 C^1 + P_K^1 K^1 - (1 + \rho^0)P_K^0 K^0] \\ &= (\rho^0 - i^0)P_K^0 K^0. \end{aligned} \quad (9.57)$$

If  $\rho^0$  (the general consumer price inflation rate) is greater than  $i^0$  (the asset inflation rate) over the course of the period, then there is a negative real revaluation effect (so that obsolescence effects dominate). In this case, Income 3 is less than Income 4, reflecting the fact that capital stocks have become less valuable (in terms of consumption equivalents) over the course of the period. If  $\rho^0$  is less than  $i^0$  over the course of the period, then the real revaluation effect is positive (so that capital stocks have become more valuable over the period). In this case, Income 3 exceeds Income 4, reflecting the fact that capital stocks have become more valuable over the course of the period and this real increase in value contributes to an increase in the period's income which is not reflected in Income 4.

To summarize: both Income 3 and Income 4 both have reasonable justifications. Choosing between them is not a straightforward matter.\*<sup>46</sup>

In the following sections of this chapter, we develop a formal economic model of production that can help to explain the factors behind growth in real income in a market economy.

## 9.9 Productivity and Real Income Growth: A Theoretical Framework

Recall the notation and assumptions made in section 9.4 above. The same assumptions will be made in the present section. Recall that in section 9.4, we assumed that there was a period  $t$  market sector technology set  $S^t$  that exhibited constant returns to scale, the period  $t$  net output and input

\*<sup>45</sup> Income 1 is much easier to justify to national income accountants because it relies on the standard production function model. On the other hand, Income 3 relies on the Austrian model of production as developed by Hicks (1961)[223] and Edwards and Bell (1961)[168] and this production model is not very familiar. This Austrian model of production has its roots in the work of Böhm-Bawerk (1891)[40], von Neumann (1937)[385] and Malinvaud (1953)[299] but these authors did not develop the user cost implications of the model. On the user cost implications of the Austrian model, see Hicks (1973; 27-35)[227] and Diewert (1977;108-111)[84] (1980; 472-474)[90].

\*<sup>46</sup> However, we lean towards Income 3 over income 1 for two reasons: (i) It seems to us that (expected) obsolescence charges are entirely similar to normal depreciation charges and Income 3 reflects this similarity and (ii) it seems to us that waiting services ( $U^1 K^0$ ) along with labour services and land rents are natural primary inputs whereas depreciation and revaluation services ( $D^1 K^0$  and  $R^1 K^0$  respectively) are more naturally regarded as a kind of intertemporal intermediate input charge (or benefit if  $R^1$  is negative).

quantity vectors were  $\mathbf{y}^t$  and  $\mathbf{x}^t$  respectively and the corresponding period  $t$  price vectors were  $\mathbf{P}^t$  and  $\mathbf{W}^t$ . We assume that the components of  $\mathbf{y}$  are the components of  $C + G + I + X - M - D - R$ , which are the usual components of market sector GDP less wear and tear depreciation  $D$  and less the revaluation term  $R$ . The components of  $\mathbf{x}$  consist of different types of labour services supplied to the market sector by households and the various types of (waiting) capital services used by the market sector.

The constant returns to scale assumption on the technology sets  $S^t$  implies that the value of outputs will equal the value of inputs in period  $t$ ; i.e., we have the following relationships:

$$g^t(\mathbf{P}^t, \mathbf{x}^t) = \mathbf{P}^t \cdot \mathbf{y}^t = \mathbf{W}^t \cdot \mathbf{x}^t; \quad t = 0, 1, 2, \dots \quad (9.58)$$

The above material will be useful in what follows but of course, our focus is not on the outputs produced by the market sector; instead our focus is on the income generated by the market sector or more precisely, on *the real income generated by the market sector*. However, since market sector net output is distributed to the factors of production used by the market sector, nominal market sector NDP will be equal to nominal market sector income; i.e., from (9.58), we have  $g^t(\mathbf{P}^t, \mathbf{x}^t) = \mathbf{P}^t \cdot \mathbf{y}^t = \mathbf{W}^t \cdot \mathbf{x}^t$ . As an approximate welfare measure that can be associated with market sector production,<sup>\*47</sup> we will choose to measure the *real income generated by the market sector in period  $t$* ,  $\rho^t$ , in terms of the number of consumption bundles that the nominal income could purchase in period  $t$ ; i.e., define  $\rho^t$  as follows:<sup>\*48</sup>

$$\begin{aligned} \rho^t &\equiv \mathbf{W}^t \cdot \mathbf{x}^t / P_C^t; & t = 0, 1, 2, \dots \\ &= \mathbf{w}^t \cdot \mathbf{x}^t \\ &= \mathbf{p}^t \cdot \mathbf{y}^t \\ &= g^t(\mathbf{p}^t, \mathbf{x}^t) \end{aligned} \quad (9.59)$$

where  $P_C^t > 0$  is the *period  $t$  consumption expenditures deflator* and the market sector period  $t$  *real output price*  $\mathbf{p}^t$  and *real input price*  $\mathbf{w}^t$  vectors are defined as the corresponding nominal price vectors deflated by the consumption expenditures price index; i.e., we have the following definitions:<sup>\*49</sup>

$$\mathbf{p}^t \equiv \mathbf{P}^t / P_C^t; \quad \mathbf{w}^t \equiv \mathbf{W}^t / P_C^t; \quad t = 0, 1, 2, \dots \quad (9.60)$$

The first and last equality in (9.59) imply that period  $t$  real income,  $\rho^t$ , is equal to the period  $t$  NDP function, evaluated at the period  $t$  real output price vector  $\mathbf{p}^t$  and the period  $t$  input vector  $\mathbf{x}^t$ ,  $g^t(\mathbf{p}^t, \mathbf{x}^t)$ . Thus *the growth in real income over time can be explained by three main factors:  $t$  (Technical Progress or Total Factor Productivity growth), growth in real output prices and the growth of primary inputs*. We will shortly give formal definitions for these three growth factors.

Using the linear homogeneity properties of the GDP functions  $g^t(\mathbf{P}, \mathbf{x})$  in  $\mathbf{P}$  and  $\mathbf{x}$  separately, we can show that the following counterparts to the relations (9.4) and (9.5) hold using the deflated

<sup>\*47</sup> Since some of the primary inputs used by the market sector can be owned by foreigners, our measure of *domestic* welfare generated by the market production sector is only an approximate one. Moreover, our suggested welfare measure is not sensitive to the distribution of the income that is generated by the market sector.

<sup>\*48</sup> Note that our use of the symbol  $\rho$  in the present section is different from our use of the symbol in previous sections.

<sup>\*49</sup> Our approach is similar to the approach advocated by Kohli (2004b; 92)[278], except he essentially deflates nominal GDP by the domestic expenditures deflator rather than just the domestic (household) expenditures deflator; i.e., he deflates by the deflator for  $C + G + I$ , whereas we suggest deflating by the deflator for  $C$ . Another difference in his approach compared to the present approach is that we restrict our analysis to the market sector GDP, whereas Kohli deflates all of GDP (probably due to data limitations). Our treatment of the balance of trade surplus or deficit is also different.

prices  $\mathbf{p}$  and  $\mathbf{w}$ :<sup>\*50</sup>

$$\mathbf{y}^t = \nabla_{\mathbf{p}} g^t(\mathbf{p}^t, \mathbf{x}^t); \quad t = 0, 1, 2, \dots \quad (9.61)$$

$$\mathbf{w}^t = \nabla_{\mathbf{x}} g^t(\mathbf{p}^t, \mathbf{x}^t); \quad t = 0, 1, 2, \dots \quad (9.62)$$

Now we are ready to define a family of *period  $t$  productivity growth factors or technical progress shift factors*  $\tau(\mathbf{p}, \mathbf{x}, t)$ :<sup>\*51</sup>

$$\tau(\mathbf{p}, \mathbf{x}, t) \equiv g^t(\mathbf{p}, \mathbf{x})/g^{t-1}(\mathbf{p}, \mathbf{x}); \quad t = 1, 2, \dots \quad (9.63)$$

Thus  $\tau(\mathbf{p}, \mathbf{x}, t)$  measures the proportional change in the real income produced by the market sector at the reference real output prices  $\mathbf{p}$  and reference input quantities used by the market sector  $\mathbf{x}$  where the numerator in (9.63) uses the period  $t$  technology and the denominator in (9.63) uses the period  $t - 1$  technology. Thus each choice of reference vectors  $\mathbf{p}$  and  $\mathbf{x}$  will generate a possibly different measure of the shift in technology going from period  $t - 1$  to period  $t$ . Note that we are using the chain system to measure the shift in technology.

It is natural to choose special reference vectors for the measure of technical progress defined by (9.63): a *Laspeyres type measure*  $\tau_L^t$  that chooses the period  $t - 1$  reference vectors  $\mathbf{p}^{t-1}$  and  $\mathbf{x}^{t-1}$  and a *Paasche type measure*  $\tau_P^t$  that chooses the period  $t$  reference vectors  $\mathbf{p}^t$  and  $\mathbf{x}^t$ :

$$\tau_L^t \equiv \tau(\mathbf{p}^{t-1}, \mathbf{x}^{t-1}, t) = g^t(\mathbf{p}^{t-1}, \mathbf{x}^{t-1})/g^{t-1}(\mathbf{p}^{t-1}, \mathbf{x}^{t-1}); \quad t = 1, 2, \dots; \quad (9.64)$$

$$\tau_P^t \equiv \tau(\mathbf{p}^t, \mathbf{x}^t, t) = g^t(\mathbf{p}^t, \mathbf{x}^t)/g^{t-1}(\mathbf{p}^t, \mathbf{x}^t); \quad t = 1, 2, \dots \quad (9.65)$$

Since both measures of technical progress are equally valid, it is natural to average them to obtain an overall measure of technical change. If we want to treat the two measures in a symmetric manner and we want the measure to satisfy the time reversal property from index number theory<sup>\*52</sup> (so that the estimate going backwards is equal to the reciprocal of the estimate going forwards), then the geometric mean will be the best simple average to take.<sup>\*53</sup> Thus we define the geometric mean of (9.64) and (9.65) as follows:<sup>\*54</sup>

$$\tau^t \equiv [\tau_L^t \tau_P^t]^{1/2}; \quad t = 1, 2, \dots \quad (9.66)$$

At this point, it is not clear how we will obtain empirical estimates for the theoretical productivity growth indexes defined by (9.64)-(9.65). One obvious way would be to assume a functional form for the NDP function  $g^t(\mathbf{p}, \mathbf{x})$ , collect data on output and input prices and quantities for the market sector for a number of years (and for the consumption expenditures deflator), add error terms to equations (9.61) and (9.62) and use econometric techniques to estimate the unknown parameters in the assumed functional form. However, econometric techniques are generally not completely straightforward: different econometricians will make different stochastic specifications and will choose different functional forms.<sup>\*55</sup> Moreover, as the number of outputs and inputs grows, it will be impossible to estimate a flexible functional form. Thus we will suggest methods for implementing measures like (9.66) in this paper that are based on exact index number techniques.

<sup>\*50</sup> If producers in the market sector of the economy are solving the profit maximization problem that is associated with  $g^t(\mathbf{P}, \mathbf{x})$ , which uses the original output prices  $\mathbf{P}$ , then they will also solve the profit maximization problem that uses the normalized output prices  $\mathbf{p} \equiv \mathbf{P}/P_C$ ; i.e., they will also solve the problem defined by  $g^t(\mathbf{p}, \mathbf{x})$ .

<sup>\*51</sup> This measure of technical progress is due to Diewert (1983; 1063)[97] and Diewert and Morrison (1986; 662)[146].

<sup>\*52</sup> See Fisher (1922; 64)[187].

<sup>\*53</sup> See the discussion in Diewert (1997)[115] on choosing the “best” symmetric average of Laspeyres and Paasche indexes that will lead to the satisfaction of the time reversal test by the resulting average index.

<sup>\*54</sup> The specific theoretical productivity change indexes defined by (9.64)-(9.66) were first defined by Diewert and Morrison (1986; 662-663)[146]. See Diewert (1993)[107] for properties of symmetric means.

<sup>\*55</sup> “The estimation of GDP functions such as (19) can be controversial, however, since it raises issues such as estimation technique and stochastic specification. ... We therefore prefer to opt for a more straightforward index number approach.” Ulrich Kohli (2004a; 344)[277].

We turn now to the problem of defining theoretical indexes for the effects on real income due to changes in real output prices. Define a family of *period  $t$  real output price growth factors*  $\alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}, s)$ :<sup>\*56</sup>

$$\alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}, s) \equiv g^s(\mathbf{p}^t, \mathbf{x})/g^s(\mathbf{p}^{t-1}, \mathbf{x}); \quad s = 1, 2, \dots \quad (9.67)$$

Thus  $\alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}, s)$  measures the proportional change in the real income produced by the market sector that is induced by the change in real output prices going from period  $t-1$  to  $t$ , using the technology that is available during period  $s$  and using the reference input quantities  $\mathbf{x}$ . Thus each choice of the reference technology  $s$  and the reference input vector  $\mathbf{x}$  will generate a possibly different measure of the effect on real income of a change in real output prices going from period  $t-1$  to period  $t$ .

Again, it is natural to choose special reference vectors for the measures defined by (9.67): a *Laspeyres type measure*  $\alpha_L^t$  that chooses the period  $t-1$  reference technology and reference input vector  $\mathbf{x}^{t-1}$  and a *Paasche type measure*  $\alpha_P^t$  that chooses the period  $t$  reference technology and reference input vector  $\mathbf{x}^t$ :

$$\alpha_L^t \equiv \alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}^{t-1}, t-1) = g^{t-1}(\mathbf{p}^t, \mathbf{x}^{t-1})/g^{t-1}(\mathbf{p}^{t-1}, \mathbf{x}^{t-1}); \quad t = 1, 2, \dots; \quad (9.68)$$

$$\alpha_P^t \equiv \alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}^t, t) = g^t(\mathbf{p}^t, \mathbf{x}^t)/g^t(\mathbf{p}^{t-1}, \mathbf{x}^t); \quad t = 1, 2, \dots \quad (9.69)$$

Since both measures of real output price change are equally valid, it is natural to average them to obtain an overall measure of the effects on real income of the change in real output prices:<sup>\*57</sup>

$$\alpha^t \equiv [\alpha_L^t \alpha_P^t]^{1/2}; \quad t = 1, 2, \dots \quad (9.70)$$

Finally, we look at the problem of defining theoretical indexes for the effects on real income due to changes in real input quantities. Define a family of *period  $t$  real input quantity growth factors*  $\beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}, s)$ :<sup>\*58</sup>

$$\beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}, s) \equiv g^s(\mathbf{p}, \mathbf{x}^t)/g^s(\mathbf{p}, \mathbf{x}^{t-1}); \quad s = 1, 2, \dots \quad (9.71)$$

Thus  $\beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}, s)$  measures the proportional change in the real income produced by the market sector that is induced by the change in input quantities used by the market sector going from period  $t-1$  to  $t$ , using the technology that is available during period  $s$  and using the reference real output prices  $\mathbf{p}$ . Thus each choice of the reference technology  $s$  and the reference real output price vector  $\mathbf{p}$  will generate a possibly different measure of the effect on real income of a change in input quantities going from period  $t-1$  to period  $t$ .

Again, it is natural to choose special reference vectors for the measures defined by (9.71): a *Laspeyres type measure*  $\beta_L^t$  that chooses the period  $t-1$  reference technology and reference real output price vector  $\mathbf{p}^{t-1}$  and a *Paasche type measure*  $\beta_P^t$  that chooses the period  $t$  reference technology and reference real output price vector  $\mathbf{p}^t$ :

$$\beta_L^t \equiv \beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}^{t-1}, t-1) = g^{t-1}(\mathbf{p}^{t-1}, \mathbf{x}^t)/g^{t-1}(\mathbf{p}^{t-1}, \mathbf{x}^{t-1}); \quad t = 1, 2, \dots; \quad (9.72)$$

$$\beta_P^t \equiv \beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}^t, t) = g^t(\mathbf{p}^t, \mathbf{x}^t)/g^t(\mathbf{p}^t, \mathbf{x}^{t-1}); \quad t = 1, 2, \dots \quad (9.73)$$

<sup>\*56</sup> This measure of real output price change was essentially defined by Fisher and Shell (1972; 56-58)[181], Samuelson and Swamy (1974; 588-592)[348], Archibald (1977; 60-61)[10], Diewert (1980; 460-461)[90] (1983; 1055)[97] and Balk (1998; 83-89)[20]. Readers who are familiar with the theory of the true cost of living index will note that the real output price index defined by (9.67) is analogous to the Konüs (1924)[280] *true cost of living index* which is a ratio of cost functions, say  $C(u, \mathbf{p}^t)/C(u, \mathbf{p}^{t-1})$  where  $u$  is a reference utility level:  $g^s$  replaces  $C$  and the reference utility level  $u$  is replaced by the vector of reference variables  $\mathbf{x}$ .

<sup>\*57</sup> The indexes defined by (9.67)-(9.70) were defined by Diewert and Morrison (1986; 664)[146] in the nominal GDP function context.

<sup>\*58</sup> This type of index was defined as a true index of value added by Sato (1976; 438)[349] and as a real input index by Diewert (1980; 456)[90].

Since both measures of real input growth are equally valid, it is natural to average them to obtain an overall measure of the effects of input growth on real income:<sup>\*59</sup>

$$\beta^t \equiv [\beta_L^t \beta_P^t]^{1/2}; \quad t = 1, 2, \dots \quad (9.74)$$

Recall that market sector real income for period  $t$  was defined by (9.59) as  $\rho^t$  equal to nominal period  $t$  factor payments  $\mathbf{W}^t \cdot \mathbf{x}^t$  deflated by the household consumption price deflator  $P_C^t$ . It is convenient to define  $\gamma^t$  as the *period  $t$  chain rate of growth factor for real income*:

$$\gamma^t \equiv \rho^t / \rho^{t-1}; \quad t = 1, 2, \dots \quad (9.75)$$

It turns out that the definitions for  $\gamma^t$  and the technology, output price and input quantity growth factors  $\tau(\mathbf{p}, \mathbf{x}, t)$ ,  $\alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}, s)$ ,  $\beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}, s)$  defined by (9.63), (9.67) and (9.71) respectively satisfy some interesting identities, which we will now develop. We have:

$$\begin{aligned} \gamma^t &\equiv \rho^t / \rho^{t-1}; \quad t = 1, 2, \dots \\ &= g^t(\mathbf{p}^t, \mathbf{x}^t) / g^{t-1}(\mathbf{p}^{t-1}, \mathbf{x}^{t-1}) \quad \text{using definitions (9.59)} \\ &= [g^t(\mathbf{p}^t, \mathbf{x}^t) / g^{t-1}(\mathbf{p}^t, \mathbf{x}^t)] [g^{t-1}(\mathbf{p}^t, \mathbf{x}^t) / g^{t-1}(\mathbf{p}^{t-1}, \mathbf{x}^t)] [g^{t-1}(\mathbf{p}^{t-1}, \mathbf{x}^t) / g^{t-1}(\mathbf{p}^{t-1}, \mathbf{x}^{t-1})] \\ &= \tau_P^t \alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}^t, t-1) \beta_L^t \quad \text{using definitions (9.65), (9.67) and (9.72)}. \end{aligned} \quad (9.76)$$

In a similar fashion, we can establish the following companion identity:

$$\gamma^t \equiv \tau_L^t \alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}^{t-1}, t) \beta_P^t \quad \text{using definitions (9.64), (9.68) and (9.73)}. \quad (9.77)$$

Thus multiplying (9.76) and (9.77) together and taking positive square roots of both sides of the resulting identity and using definitions (9.66) and (9.74), we obtain the following identity:

$$\gamma^t \equiv \tau^t [\alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}^t, t-1) \alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}^{t-1}, t)]^{1/2} \beta^t; \quad t = 1, 2, \dots \quad (9.78)$$

In a similar fashion, we can derive the following alternative decomposition for  $\gamma^t$  into growth factors:

$$\gamma^t \equiv \tau^t \alpha^t [\beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}^t, t-1) \beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}^{t-1}, t)]^{1/2}; \quad t = 1, 2, \dots \quad (9.79)$$

It is quite likely that the real output price growth factor  $[\alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}^t, t-1) \alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}^{t-1}, t)]^{1/2}$  is fairly close to  $\alpha^t$  defined by (9.70) and it is quite likely that the input growth factor  $[\beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}^t, t-1) \beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}^{t-1}, t)]^{1/2}$  is quite close to  $\beta^t$  defined by (9.74); i.e., we have the following approximate equalities:

$$[\alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}^t, t-1) \alpha(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{x}^{t-1}, t)]^{1/2} \approx \alpha^t; \quad t = 1, 2, \dots; \quad (9.80)$$

$$[\beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}^t, t-1) \beta(\mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{p}^{t-1}, t)]^{1/2} \approx \beta^t; \quad t = 1, 2, \dots \quad (9.81)$$

Substituting (9.80) and (9.81) into (9.78) and (9.79) respectively leads to the following approximate decompositions for the growth of real income into explanatory factors:

$$\gamma^t \approx \tau^t \alpha^t \beta^t; \quad t = 1, 2, \dots \quad (9.82)$$

where  $\tau^t$  is a technology growth factor,  $\alpha^t$  is a growth in real output prices factor and  $\beta^t$  is a growth in primary inputs factor.

<sup>\*59</sup> The theoretical indexes defined by (9.71)-(9.74) were defined in Diewert and Morrison (1986; 665)[146] in the nominal GDP context.

Rather than look at explanatory factors for the growth in real market sector income, it is sometimes convenient to express the level of real income in period  $t$  in terms of an *index of the technology level* or of Total Factor Productivity in period  $t$ ,  $T^t$ , of the *level of real output prices* in period  $t$ ,  $A^t$ , and of the *level of primary input quantities* in period  $t$ ,  $B^t$ .<sup>\*60</sup> Thus we use the growth factors  $\tau^t$ ,  $\alpha^t$  and  $\beta^t$  as follows to define the levels  $T^t$ ,  $A^t$  and  $B^t$ :

$$T^0 \equiv 1; T^t \equiv T^{t-1}\tau^t; \quad t = 1, 2, \dots; \quad (9.83)$$

$$A^0 \equiv 1; A^t \equiv A^{t-1}\alpha^t; \quad t = 1, 2, \dots; \quad (9.84)$$

$$B^0 \equiv 1; B^t \equiv B^{t-1}\beta^t; \quad t = 1, 2, \dots \quad (9.85)$$

Using the approximate equalities (9.82) for the chain links that appear in (9.83)-(9.85), we can establish the following approximate relationship for the level of real income in period  $t$ ,  $\rho^t$ , and the period  $t$  levels for technology, real output prices and input quantities:

$$\rho^t/\rho^0 \approx T^t A^t B^t; \quad t = 0, 1, 2, \dots \quad (9.86)$$

In the following section, we note a set of assumptions on the technology sets that will ensure that the approximate real income growth decompositions (9.82) and (9.86) hold as exact equalities.

## 9.10 The Translog GDP Function Approach

We now follow the example of Diewert and Morrison (1986; 663)[146] and assume that the log of the period  $t$  (deflated) NDP function,  $g^t(\mathbf{p}, \mathbf{x})$ , has the following translog functional form:<sup>\*61</sup>

$$\begin{aligned} \ln g^t(\mathbf{p}, \mathbf{x}) \equiv & a_0^t + \sum_{m=1}^M a_m^t \ln p_m + (1/2) \sum_{m=1}^M \sum_{k=1}^M a_{mk} \ln p_m \ln p_k \\ & + \sum_{n=1}^N b_n^t \ln x_n + (1/2) \sum_{n=1}^N \sum_{j=1}^N b_{nj} \ln x_n \ln x_j + \sum_{m=1}^M \sum_{n=1}^N c_{mn} \ln p_m \ln x_n; \\ & t = 0, 1, 2, \dots \end{aligned} \quad (9.87)$$

Note that the coefficients for the quadratic terms are assumed to be constant over time. The coefficients must satisfy the following restrictions in order for  $g^t$  to satisfy the linear homogeneity

<sup>\*60</sup> This type of levels presentation of the data is quite instructive when presented in graphical form. It was suggested by Kohli (1990)[274] and used extensively by him; see Kohli (1991)[275], (2003)[276] (2004a)[277] (2004b)[278] and Fox and Kohli (1998)[190].

<sup>\*61</sup> This functional form was first suggested by Diewert (1974; 139)[76] as a generalization of the translog functional form introduced by Christensen, Jorgenson and Lau (1971)[53]. Diewert (1974; 139)[76] indicated that this functional form was flexible.

properties that we have assumed in section 9.4 above:<sup>\*62</sup>

$$\sum_{m=1}^M a_m^t = 1 \text{ for } t = 0, 1, 2, \dots; \quad (9.88)$$

$$\sum_{n=1}^N b_n^t = 1 \text{ for } t = 0, 1, 2, \dots; \quad (9.89)$$

$$a_{mk} = a_{km} \text{ for all } k, m; \quad (9.90)$$

$$b_{nj} = b_{jn} \text{ for all } n, j; \quad (9.91)$$

$$\sum_{k=1}^M a_{mk} = 0 \text{ for } m = 1, \dots, M; \quad (9.92)$$

$$\sum_{j=1}^N b_{nj} = 0 \text{ for } n = 1, \dots, N; \quad (9.93)$$

$$\sum_{n=1}^N c_{mn} = 0 \text{ for } m = 1, \dots, M; \quad (9.94)$$

$$\sum_{m=1}^M c_{mn} = 0 \text{ for } n = 1, \dots, N. \quad (9.95)$$

Recall the approximate decomposition of real income growth going from period  $t - 1$  to  $t$  given by (9.82) above,  $\gamma^t \approx \tau^t \alpha^t \beta^t$ . Diewert and Morrison (1986; 663)[146] showed that<sup>\*63</sup> if  $g^{t-1}$  and  $g^t$  are defined by (9.87)-(9.95) above and there is competitive profit maximizing behavior on the part of all market sector producers for all periods  $t$ , then (9.82) holds as an exact equality; i.e., we have<sup>\*64</sup>

$$\gamma^t = \tau^t \alpha^t \beta^t; \quad t = 1, 2, \dots \quad (9.96)$$

In addition, Diewert and Morrison (1986; 663-665)[146] showed that  $\tau^t$ ,  $\alpha^t$  and  $\beta^t$  could be calculated using empirically observable price and quantity data for periods  $t - 1$  and  $t$  as follows:

$$\begin{aligned} \ln \alpha^t &= \sum_{m=1}^M (1/2)[(p_m^{t-1} y_m^{t-1} / \mathbf{p}^{t-1} \cdot \mathbf{y}^{t-1}) + (p_m^t y_m^t / \mathbf{p}^t \cdot \mathbf{y}^t)] \ln(p_m^t / p_m^{t-1}) \\ &= \ln P_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t); \end{aligned} \quad (9.97)$$

$$\begin{aligned} \ln \beta^t &= \sum_{n=1}^N (1/2)[(w_n^{t-1} x_n^{t-1} / \mathbf{w}^{t-1} \cdot \mathbf{x}^{t-1}) + (w_n^t x_n^t / \mathbf{w}^t \cdot \mathbf{x}^t)] \ln(x_n^t / x_n^{t-1}) \\ &= \ln Q_T(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t); \end{aligned} \quad (9.98)$$

$$\tau^t = \gamma^t / \alpha^t \beta^t \quad (9.99)$$

where  $P_T(\mathbf{p}^{t-1}, \mathbf{p}^t, \mathbf{y}^{t-1}, \mathbf{y}^t)$  is the Törnqvist (1936)[373] and Törnqvist and Törnqvist (1937)[374] output price index and  $Q_T(\mathbf{w}^{t-1}, \mathbf{w}^t, \mathbf{x}^{t-1}, \mathbf{x}^t)$  is the Törnqvist input quantity index.

Since equations (9.96) now hold as exact identities under our present assumptions, equations (9.86), the cumulated counterparts to equations (9.82), will also hold as exact decompositions; i.e., under our present assumptions, we have

$$\rho^t / \rho^0 = T^t A^t B^t; \quad t = 1, 2, \dots \quad (9.100)$$

We will implement the real income decompositions (9.96) and (9.100) in our empirical projects.

<sup>\*62</sup> There are additional restrictions on the parameters which are necessary to ensure that  $g^t(\mathbf{p}, \mathbf{x})$  is convex in  $\mathbf{p}$  and concave in  $\mathbf{x}$ . Note that this functional form is a special case of the translog variable profit function defined in the previous chapter. In the present chapter, we assume price taking competitive behavior on the part of producers and constant returns to scale so there are no monopolistic markups.

<sup>\*63</sup> Diewert and Morrison established their proof using the nominal GDP function  $g^t(\mathbf{p}, \mathbf{x})$ . However, it is easy to rework their proof using the deflated GDP function  $g^t(\mathbf{p}, \mathbf{x})$  using the fact that  $g^t(\mathbf{p}, \mathbf{x}) = g^t(\mathbf{P}/P_C, \mathbf{x}) = g^t(\mathbf{P}, \mathbf{x})/P_C$  using the linear homogeneity property of  $g^t(\mathbf{P}, \mathbf{x})$  in  $\mathbf{P}$ .

<sup>\*64</sup> We essentially proved a more general version of this result in section 7.10 of chapter 7 and this more general result can be specialized to give us the exact decomposition (9.96).

## 9.11 The Translog GDP Function Approach and Changes in the Terms of Trade

For some purposes, it is convenient to decompose the aggregate period  $t$  contribution factor due to changes in all deflated output prices  $\alpha^t$  into separate effects for each change in each output price. Similarly, it can sometimes be useful to decompose the aggregate period  $t$  contribution factor due to changes in all market sector primary input quantities  $\beta^t$  into separate effects for each change in each input quantity. In this section, we indicate how this can be done, making the same assumptions on the technology that were made in the previous section.

We first model the effects of a change in a single (deflated) output price, say  $p_m$ , going from period  $t-1$  to  $t$ . Counterparts to the theoretical Laspeyres and Paasche type price indexes defined by (9.68) and (9.69) above for changes in all (deflated) output prices are the following *Laspeyres type measure*  $\alpha_{Lm}^t$  that chooses the period  $t-1$  reference technology and holds constant other output prices at their period  $t-1$  levels and holds inputs constant at their period  $t-1$  levels  $\mathbf{x}^{t-1}$  and a *Paasche type measure*  $\alpha_{Pm}^t$  that chooses the period  $t$  reference technology and reference input vector  $\mathbf{x}^t$  and holds constant other output prices at their period  $t$  levels:

$$\alpha_{Lm}^t \equiv g^{t-1}(p_1^{t-1}, \dots, p_{m-1}^{t-1}, p_m^t, p_{m+1}^{t-1}, \dots, p_M^{t-1}, \mathbf{x}^{t-1}) / g^{t-1}(\mathbf{p}^{t-1}, \mathbf{x}^{t-1});$$

$$m = 1, \dots, M; t = 1, 2, \dots; \quad (9.101)$$

$$\alpha_{Pm}^t \equiv g^t(\mathbf{p}^t, \mathbf{x}^t) / g^t(p_1^t, \dots, p_{m-1}^t, p_m^{t-1}, p_{m+1}^t, \dots, p_M^t, \mathbf{x}^t);$$

$$m = 1, \dots, M; t = 1, 2, \dots \quad (9.102)$$

Since both measures of real output price change are equally valid, it is natural to average them to obtain an *overall measure of the effects on real income of the change in the real price of output  $m$* :<sup>\*65</sup>

$$\alpha_m^t \equiv [\alpha_{Lm}^t \alpha_{Pm}^t]^{1/2}; \quad m = 1, \dots, M; t = 1, 2, \dots \quad (9.103)$$

Under the assumption that the deflated GDP functions  $g^t(\mathbf{p}, \mathbf{x})$  have the translog functional forms as defined by (9.87)-(9.95) in the previous section, the arguments of Diewert and Morrison (1986; 666)[146] can be adapted to give us the following result:

$$\ln \alpha_m^t = (1/2)[(p_m^{t-1} y_m^{t-1} / \mathbf{p}^{t-1} \cdot \mathbf{y}^{t-1}) + (p_m^t y_m^t / \mathbf{p}^t \cdot \mathbf{y}^t)] \ln(p_m^t / p_m^{t-1});$$

$$m = 1, \dots, M; t = 1, 2, \dots \quad (9.104)$$

Note that  $\ln \alpha_m^t$  is equal to the  $m$ th term in the summation of the terms on the right hand side of (9.97). This observation means that we have the following exact decomposition of the period  $t$  aggregate real output price contribution factor  $\alpha^t$  into a product of separate price contribution factors; i.e., we have under present assumptions:

$$\alpha^t = \alpha_1^t \alpha_2^t \cdots \alpha_M^t; \quad t = 1, 2, \dots \quad (9.105)$$

The above decomposition is useful for analyzing how real changes in the price of exports (i.e., a change in the price of exports relative to the price of domestic consumption) and in the price of imports impact on the real income generated by the market sector. In your empirical work, we let  $M$  equal three. The three net outputs are:

- Domestic sales less depreciation and revaluation ( $C + I + G - D - R$ );

<sup>\*65</sup> The indexes defined by (9.101)-(9.103) were defined by Diewert and Morrison (1986; 666)[146] in the nominal GDP function context.

- Exports ( $X$ ) and
- Imports ( $M$ ).

Since commodities 1 and 2 are outputs,  $y_1$  and  $y_2$  will be positive but since commodity 3 is an input into the market sector,  $y_3$  will be negative. Hence an increase in the real price of exports will *increase* real income but an increase in the real price of imports will *decrease* the real income generated by the market sector, as is evident by looking at the contribution terms defined by (9.104) for  $m = 2$  (where  $y_m^t > 0$ ) and for  $m = 3$  (where  $y_m^t < 0$ ).

As mentioned above, it is also useful to have a decomposition of the aggregate contribution of input growth to the growth of real income into separate contributions for each important class of primary input that is used by the market sector. We now model the effects of a change in a single input quantity, say  $x_n$ , going from period  $t-1$  to  $t$ . Counterparts to the theoretical Laspeyres and Paasche type quantity indexes defined by (9.72) and (9.73) above for changes in input  $n$  are the following *Laspeyres type measure*  $\beta_{Ln}^t$  that chooses the period  $t-1$  reference technology and holds constant other input quantities at their period  $t-1$  levels and holds real output prices at their period  $t-1$  levels  $\mathbf{p}^{t-1}$  and a *Paasche type measure*  $\beta_{Pn}^t$  that chooses the period  $t$  reference technology and reference real output price vector  $\mathbf{p}^t$  and holds constant other input quantities at their period  $t$  levels:

$$\beta_{Ln}^t \equiv g^{t-1}(\mathbf{p}^{t-1}, x_1^{t-1}, \dots, x_{n-1}^{t-1}, x_n^t, x_{n+1}^{t-1}, \dots, x_N^{t-1}) / g^{t-1}(\mathbf{p}^{t-1}, \mathbf{x}^{t-1});$$

$$n = 1, \dots, N; t = 1, 2, \dots; \quad (9.106)$$

$$\beta_{Pn}^t \equiv g^t(\mathbf{p}^t, \mathbf{x}^t) / g^t(\mathbf{p}^t, x_1^t, \dots, x_{n-1}^t, x_n^{t-1}, x_{n+1}^t, \dots, x_N^t);$$

$$n = 1, \dots, N; t = 1, 2, \dots \quad (9.107)$$

Since both measures of input change are equally valid, as usual, we average them to obtain *an overall measure of the effects on real income of the change in the quantity of input  $n$* :<sup>\*66</sup>

$$\beta_n^t \equiv [\beta_{Ln}^t \beta_{Pn}^t]^{1/2}; \quad n = 1, \dots, N; t = 1, 2, \dots \quad (9.108)$$

Under the assumption that the deflated GDP functions  $g^t(\mathbf{p}, \mathbf{x})$  have the translog functional forms as defined by (9.87)-(9.95) in the previous section, the arguments of Diewert and Morrison (1986; 667)[146] can be adapted to give us the following result:

$$\ln \beta_n^t = (1/2)[(w_n^{t-1} x_n^{t-1} / \mathbf{w}^{t-1} \cdot \mathbf{x}^{t-1}) + (w_n^t x_n^t / \mathbf{w}^t \cdot \mathbf{x}^t)] \ln(x_n^t / x_n^{t-1});$$

$$n = 1, \dots, N; t = 1, 2, \dots \quad (9.109)$$

Note that  $\ln \beta_n^t$  is equal to the  $n$ th term in the summation of the terms on the right hand side of (9.98). This observation means that we have the following exact decomposition of the period  $t$  aggregate input growth contribution factor  $\beta^t$  into a product of separate input quantity contribution factors; i.e., we have under present assumptions:

$$\beta^t = \beta_1^t \beta_2^t \cdots \beta_N^t; \quad t = 1, 2, \dots \quad (9.110)$$

For an empirical application of the methodology (to Australia) explained in the last 3 sections of this chapter, see Diewert and Lawrence (2006)[145].

<sup>\*66</sup> The indexes defined by (9.106)-(9.108) were defined by Diewert and Morrison (1986; 667)[146] in the nominal GDP function context.

**Problem 8** Let  $\mathbf{x}$  and  $\mathbf{y}$  be  $N$  and  $M$  dimensional vectors respectively and let  $f^1$  and  $f^2$  be two general quadratic functions defined as follows:

$$f^1(\mathbf{x}, \mathbf{y}) \equiv a_0^1 + \mathbf{a}^{1T} \mathbf{x} + \mathbf{b}^{1T} \mathbf{y} + (1/2)\mathbf{x}^T \mathbf{A}^1 \mathbf{x} + (1/2)\mathbf{y}^T \mathbf{B}^1 \mathbf{y} + \mathbf{x}^T \mathbf{C}^1 \mathbf{y}; \quad \mathbf{A}^{1T} = \mathbf{A}^1; \mathbf{B}^{1T} = \mathbf{B}^1; \quad (\text{i})$$

$$f^2(\mathbf{x}, \mathbf{y}) \equiv a_0^2 + \mathbf{a}^{2T} \mathbf{x} + \mathbf{b}^{2T} \mathbf{y} + (1/2)\mathbf{x}^T \mathbf{A}^2 \mathbf{x} + (1/2)\mathbf{y}^T \mathbf{B}^2 \mathbf{y} + \mathbf{x}^T \mathbf{C}^2 \mathbf{y}; \quad \mathbf{A}^{2T} = \mathbf{A}^2; \mathbf{B}^{2T} = \mathbf{B}^2 \quad (\text{ii})$$

where the  $a_0^i$  are scalar parameters, the  $\mathbf{a}^i$  and  $\mathbf{b}^i$  are parameter vectors and the  $\mathbf{A}^i, \mathbf{B}^i$  and  $\mathbf{C}^i$  are parameter matrices for  $i = 1, 2$ . Note that the  $\mathbf{A}^i$  and  $\mathbf{B}^i$  are symmetric matrices.

(a) If  $\mathbf{A}^1 = \mathbf{A}^2$ , show that the following equation holds for all  $\mathbf{x}^1, \mathbf{x}^2, \mathbf{y}^1$  and  $\mathbf{y}^2$ :

$$f^1(\mathbf{x}^2, \mathbf{y}^1) - f^1(\mathbf{x}^1, \mathbf{y}^1) + f^2(\mathbf{x}^2, \mathbf{y}^2) - f^2(\mathbf{x}^1, \mathbf{y}^2) = [\nabla_x f^1(\mathbf{x}^1, \mathbf{y}^1) + \nabla_x f^2(\mathbf{x}^2, \mathbf{y}^2)]^T [\mathbf{x}^2 - \mathbf{x}^1]. \quad (\text{iii})$$

(b) If  $\mathbf{B}^1 = \mathbf{B}^2$ , show that the following equation holds for all  $\mathbf{x}^1, \mathbf{x}^2, \mathbf{y}^1$  and  $\mathbf{y}^2$ :

$$f^1(\mathbf{x}^1, \mathbf{y}^2) - f^1(\mathbf{x}^1, \mathbf{y}^1) + f^2(\mathbf{x}^2, \mathbf{y}^2) - f^2(\mathbf{x}^2, \mathbf{y}^1) = [\nabla_y f^1(\mathbf{x}^1, \mathbf{y}^1) + \nabla_y f^2(\mathbf{x}^2, \mathbf{y}^2)]^T [\mathbf{y}^2 - \mathbf{y}^1]. \quad (\text{iv})$$

*Hint:* Straightforward substitution into both sides of (a) and (b) will establish the above identities. These identities are a generalization of Diewert's (1976; 118)[82] *quadratic identity*. Logarithmic versions of the above identities correspond to the *translog identity* which was established in the Appendix to Caves, Christensen and Diewert (1982; 1412-1413)[49].

**Problem 9** Prove (9.104).

**Problem 10** Prove (9.109).

## 9.12 References

- Allen, R.G.D. (1949), "The Economic Theory of Index Numbers", *Economica* 16, 197-203.
- Archibald, R.B. (1977), "On the Theory of Industrial Price Measurement: Output Price Indexes", *Annals of Economic and Social Measurement* 6, 57-72.
- Balk, B.M. (1998), *Industrial Price, Quantity and Productivity Indices*, Boston: Kluwer Academic Publishers.
- Baxter, W.T. (1975), *Accounting Values and Inflation*, London: McGraw-Hill.
- Bell, A.L. (1953), "Fixed Assets and Current Costs", *The Accounting Review* 28, 44-53.
- Böhm-Bawerk, E. V. (1891), *The Positive Theory of Capital*, W. Smart (translator of the original German book published in 1888), New York: G.E. Stechert.
- Carsberg, B. (1982), "The Case for Financial Capital Maintenance", pp. 59-74 in *Maintenance of Capital: Financial versus Physical*, R.R. Sterling and K.W. Lemke (eds.), Houston: Scholars Book Co.
- Caves, D.W., L.R. Christensen and W.E. Diewert (1982), "The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity", *Econometrica* 50, 1393-1414.
- Christensen, L.R., D.W. Jorgenson and L.J. Lau (1971), "Conjugate Duality and the Transcendental Logarithmic Production Function", *Econometrica* 39, 255-256.
- Daines, H.C. (1929), "The Changing Objectives of Accounting", *The Accounting Review* 4, 94-110.
- Diewert, W.E. (1973), "Functional Forms for Profit and Transformation Functions", *Journal of Economic Theory* 6, 284-316.
- Diewert, W.E., (1974), "Applications of Duality Theory," pp. 106-171 in M.D. Intriligator and D.A. Kendrick (ed.), *Frontiers of Quantitative Economics*, Vol. II, Amsterdam: North-Holland.

- Diewert, W.E. (1977), "Walras' Theory of Capital Formation and the Existence of a Temporary Equilibrium", pp. 73-126 in *Equilibrium and Disequilibrium in Economic Theory*, G. Schwödauer (ed.), Dordrecht: D. Reidel.
- Diewert, W.E. (1976), "Exact and Superlative Index Numbers", *Journal of Econometrics* 4, 114-145.
- Diewert, W.E. (1978), "Superlative Index Numbers and Consistency in Aggregation", *Econometrica* 46, 883-900.
- Diewert, W.E. (1980), "Aggregation Problems in the Measurement of Capital", pp.433-528 in *The Measurement of Capital*, edited by D. Usher, Studies in Income and Wealth, Vol. 45, National Bureau of Economics Research, University of Chicago Press, Chicago.
- Diewert, W.E. (1983), "The Theory of the Output Price Index and the Measurement of Real Output Change", pp. 1049-1113 in *Price Level Measurement*, editors W.E. Diewert and C. Montmarquette, Ottawa: Statistics Canada.
- Diewert, W.E. (1993), "Symmetric Means and Choice Under Uncertainty", pp. 355-433 in *Essays in Index Number Theory, Volume I*, Contributions to Economic Analysis 217, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North Holland.
- Diewert, W.E. (1997), "Commentary" on Mathew D. Shapiro and David W. Wilcox, "Alternative Strategies for Aggregating Price in the CPI", *The Federal Reserve Bank of St. Louis Review*, 79:3, 127-137.
- Diewert, W.E. (2005a), "Issues in the Measurement of Capital Services, Depreciation, Asset Price Changes and Interest Rates", pp. 479-542 in *Measuring Capital in the New Economy*, C. Corrado, J. Haltiwanger and D. Sichel (eds.), Chicago: University of Chicago Press.
- Diewert, W.E. (2005b), "Accounting Theory and Alternative Methods of Asset Valuation", Chapter 3 of a *Tutorial The Measurement of Business Capital, Income and Performance* presented at the University Autonoma of Barcelona, Spain, September 21-22, 2005; revised December 2005.
- Diewert, W.E. (2006a), "Capital and Accounting Theory: The Early History", Chapter 2 of a *Tutorial The Measurement of Business Capital, Income and Performance* presented at the University Autonoma of Barcelona, Spain, September 21-22, 2005; revised February 2006.
- Diewert, W.E. (2006b), "The Measurement of Income", Chapter 7 of a *Tutorial The Measurement of Business Capital, Income and Performance* presented at the University Autonoma of Barcelona, Spain, September 21-22, 2005; revised May 2006.
- Diewert, W.E. and K.J. Fox (2005), "The New Economy and an Old Problem: Net Versus Gross Output", Center for Applied Economic Research Working Paper 2005/02, University of New South Wales, January.
- Diewert, W.E. and D. Lawrence (2006), *Measuring the Contributions of Productivity and Terms of Trade to Australia's Economic Welfare*, Consultancy Report to the Productivity Commission, Australian Government, Canberra, March.
- Diewert, W.E. and C.J. Morrison (1986), "Adjusting Output and Productivity Indexes for Changes in the Terms of Trade", *The Economic Journal* 96, 659-679.
- Edwards, E.O. and P.W. Bell (1961), *The Theory and Measurement of Business Income*, Berkeley: University of California Press.
- Eurostat, International Monetary Fund, OECD, United Nations and World Bank (1993), *System of National Accounts 1993*, Luxembourg, New York, Paris, Washington DC.
- Feenstra, R.C. (2004), *Advanced International Trade: Theory and Evidence*, Princeton N.J.: Princeton University Press.
- Fisher, F.M. and K. Shell (1972), "The Pure Theory of the National Output Deflator", pp. 49-113 in *The Economic Theory of Price Indexes*, New York: Academic Press.
- Fisher, I. (1922), *The Making of Index Numbers*, Houghton-Mifflin, Boston.

- Fox, K.J. and U. Kohli (1998), "GDP Growth, Terms of Trade Effects and Total Factor Productivity", *Journal of International Trade and Economic Development* 7, 87-110.
- Gorman, W.M. (1968), "Measuring the Quantities of Fixed Factors", pp. 141-172 in *Value, Capital and Growth: Papers in Honour of Sir John Hicks*, J.N Wolfe (ed.), Chicago: Aldine Press.
- Haig, R.M. (1959), "The Concept of Income: Economic and Legal Aspects", pp. 54-76 in *Readings in the Economics of Taxation*, R.A. Musgrave and C.S. Shoup (eds.), Homewood, Illinois: Richard D. Irwin (Haig's chapter was originally published in 1921).
- Hayek, F.A. v. (1941), "Maintaining Capital Intact: A Reply", *Economica* 8, 276-280.
- Hicks, J.R. (1939), *Value and Capital*, Oxford: The Clarendon Press.
- Hicks, J.R. (1942), "Maintaining Capital Intact: a Further Suggestion", *Economica* 9, 174-179.
- Hicks, J.R. (1946), *Value and Capital*, Second Edition, Oxford: Clarendon Press.
- Hicks, J.R. (1961), "The Measurement of Capital in Relation to the Measurement of Other Economic Aggregates", pp. 18-31 in *The Theory of Capital*, F.A. Lutz and D.C. Hague (eds.), London: Macmillan.
- Hicks, J. (1973), *Capital and Time: A Neo-Austrian Theory*, Oxford: Clarendon Press.
- Hill, P. (2000); "Economic Depreciation and the SNA"; paper presented at the 26th Conference of the International Association for Research on Income and Wealth; Cracow, Poland.
- Hill, R.J. and T.P. Hill (2003), "Expectations, Capital Gains and Income", *Economic Inquiry* 41, 607-619.
- Hotelling, H. (1932), "Edgeworth's Taxation Paradox and the Nature of Demand and Supply Functions", *Journal of Political Economy* 40, 577-616.
- Kohli, U. (1978), "A Gross National Product Function and the Derived Demand for Imports and Supply of Exports", *Canadian Journal of Economics* 11, 167-182.
- Kohli, U. (1990), "Growth Accounting in the Open Economy: Parametric and Nonparametric Estimates", *Journal of Economic and Social Measurement* 16, 125-136.
- Kohli, U. (1991), *Technology, Duality and Foreign Trade: The GNP Function Approach to Modeling Imports and Exports*, Ann Arbor: University of Michigan Press.
- Kohli, U. (2003), "Growth Accounting in the Open Economy: International Comparisons", *International Review of Economics and Finance* 12, 417-435.
- Kohli, U. (2004a), "An Implicit Törnqvist Index of Real GDP", *Journal of Productivity Analysis* 21, 337-353.
- Kohli, U. (2004b), "Real GDP, Real Domestic Income and Terms of Trade Changes", *Journal of International Economics* 62, 83-106.
- Konüs, A.A. (1924), "The Problem of the True Index of the Cost of Living", translated in *Econometrica* 7, (1939), 10-29.
- Lau, L. (1976), "A Characterization of the Normalized Restricted Profit Function", *Journal of Economic Theory*, 12:1, 131-163.
- Malinvaud, E. (1953), "Capital Accumulation and the Efficient Allocation of Resources", *Econometrica* 21, 233-268.
- Marshall, A. (1890), *Principles of Economics*, London: Macmillan.
- McFadden, D. (1978), "Cost, Revenue and Profit Functions", pp. 3-109 in *Production Economics: A Dual Approach to Theory and Applications*. Volume 1, M. Fuss and D. McFadden (eds.), Amsterdam: North-Holland.
- Middleditch, L. (1918), "Should Accounts Reflect the Changing Value of the Dollar?", *The Journal of Accountancy* 25, 114-120.

- Pigou, A.C. (1924), *The Economics of Welfare*, Second Edition, London: Macmillan.
- Pigou, A.C. (1935), "Net Income and Capital Depletion", *The Economic Journal* 45, 235-241.
- Pigou, A.C. (1941), "Maintaining Capital Intact", *Economica* 8, 271-275.
- Rymes, T.K. (1968), "Professor Read and the Measurement of Total Factor Productivity", *The Canadian Journal of Economics* 1, 359-367.
- Rymes, T.K. (1983), "More on the Measurement of Total Factor Productivity", *The Review of Income and Wealth* 29 (September), 297-316.
- Samuelson, P.A. (1953), "Prices of Factors and Goods in General Equilibrium", *Review of Economic Studies* 21, 1-20.
- Samuelson, P.A. (1961), "The Evaluation of 'Social Income': Capital Formation and Wealth", pp. 32-57 in *The Theory of Capital*, F.A. Lutz and D.C. Hague (eds.), London: Macmillan.
- Samuelson, P.A. and S. Swamy (1974), "Invariant Economic Index Numbers and Canonical Duality: Survey and Synthesis", *American Economic Review* 64, 566-593.
- Sato, K. (1976), "The Meaning and Measurement of the Real Value Added Index", *Review of Economics and Statistics* 58, 434-442.
- Sterling, R.R. (1975), "Relevant Financial Reporting in an Age of Price Changes", *The Journal of Accountancy* 139 (February), 42-51.
- Sweeney, H.W. (1934), "Approximations of Appraisal Values by Index Numbers", *Harvard Business Review* 13, 108-115.
- Sweeney, H.W. (1935), "The Technique of Stabilized Accounting", *The Accounting Review* 10, 185-205.
- Sweeney, H.W. (1964), *Stabilized Accounting*, New York: Holt, Rinehart and Winston (reissue of the 1936 original with a new foreword).
- Törnqvist, L. (1936), "The Bank of Finland's Consumption Price Index", *Bank of Finland Monthly Bulletin* 10: 1-8.
- Törnqvist, L. and E. Törnqvist (1937), "Vilket är förhållandet mellan finska markens och svenska kronans köpkraft?", *Ekonomiska Samfundets Tidskrift* 39, 1-39 reprinted as pp. 121-160 in *Collected Scientific Papers of Leo Törnqvist*, Helsinki: The Research Institute of the Finnish Economy, 1981.
- Tweedie, D. and G. Whittington (1984), *The Debate on Inflation Accounting*, London: Cambridge University Press.
- von Neumann, J. (1937), "Über ein Ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes", *Ergebnisse eines Mathematische Kolloquiums* 8, 73-83; translated as "A Model of General Economic Equilibrium", *Review of Economic Studies* (1945-6) 12, 1-9.
- Whittington, G. (1980), "Pioneers of Income Measurement and Price-Level Accounting: A Review Article", *Accounting and Business Research* Spring, 232-240.
- Woodland, A.D. (1982), *International Trade and Resource Allocation*, Amsterdam: North-Holland.



## Chapter 10

# Flexible Functional Forms

### 10.1 Introduction

In this chapter, we will take an in depth look at the problems involved in choosing functional forms for estimating systems of consumer and producer demand functions and producer supply functions. We will attempt to find functional forms that are consistent with the restrictions on supply and demand functions that are implied by economic theory but are also sufficiently flexible that elasticities of supply and demand are not arbitrarily restricted by the choice of the functional form. We will make extensive use of duality theory<sup>\*1</sup> in this chapter in order to obtain systems of demand and supply functions that are consistent with economic theory but yet can be estimated by using linear regression techniques or “slightly” nonlinear regressions. Since many problems in applied economics depend on obtaining accurate estimates of elasticities, this topic is of considerable importance for the applied economist.

Section 10.2 below starts off by giving a formal definition of a flexible functional form for a production function and a cost function. Basically, flexible functional forms are functional forms that have a second order approximation property so that elasticities of supply and demand are not a priori restricted by using a flexible functional form. Sections 10.3-10.5 give three examples of flexible functional forms for cost functions: the Generalized Leontief cost function, the Translog cost function and the Normalized Quadratic cost function. The Normalized Quadratic functional form is our preferred functional form, because convexity or concavity restrictions can be imposed on this functional form in a parsimonious way without destroying the flexibility of the functional form. We do not know of any other flexible functional form that has this property.<sup>\*2</sup>

Section 10.6 shows how cost functions can be applied to the problems involved in estimating systems of consumer demand functions that are consistent with utility maximizing behavior. Sections 10.7 and 10.8 apply the general methodology to two specific functional forms: the Generalized Leontief cost function and the Normalized Quadratic cost function. Section 10.9 discusses the problems involved in cardinalizing a measure of utility. Section 10.10 discusses how nonhomothetic preferences can be estimated and section 10.11 extends this discussion by showing how the use of spline functions can add extra flexibility.

In section 10.12, we turn our attention to the problems involved in estimating multiple output, multiple input technologies.<sup>\*3</sup> The unit (capital) profit function is a key concept that is explained in this section. Sections 10.13-10.16 apply the general framework to a number of specific functional

---

<sup>\*1</sup> See chapter 3 of these notes. Some of the material in chapter 3 will be repeated in the present chapter.

<sup>\*2</sup> For a comparison of the Normalized Quadratic functional form with other flexible functional forms, see Diewert and Wales (1993)[157].

<sup>\*3</sup> Sections 10.2-10.5 dealt with only single output, multiple input technologies.

forms. Section 10.17 is a counterpart to section 10.11 and shows how spline functions can be used to add extra flexibility.

Finally, sections 10.18 and 10.19 provide some generalizations of the basic normalized quadratic functional form. In section 10.18, we introduce a variant of the normalized quadratic profit function that can achieve flexibility at two points instead of the usual one point flexibility property.\*<sup>4</sup> In order to implement this model, the number of commodities cannot be too large, since having enough parameters to be flexible at two points instead of one point will double the number of parameters to be estimated. On the other hand, the generalization of the Normalized Quadratic functional form presented in section 10.19 is applicable to situations where the number of commodities is very large.\*<sup>5</sup>

## 10.2 The Definition of a Flexible Functional Form

Consider an  $N$  input, one output constant returns to scale *production function*  $f$  where  $y = f(x_1, x_2, \dots, x_N) = f(\mathbf{x})$  and  $y \geq 0$  denotes the output produced by the nonnegative input vector  $\mathbf{x} \geq \mathbf{0}_N$ .

The constant returns to scale assumption means that  $f$  is *linearly homogeneous*; i.e., we have

$$f(\lambda \mathbf{x}) = \lambda f(\mathbf{x}) \text{ for all scalars } \lambda \geq 0 \text{ and input vectors } \mathbf{x} \geq \mathbf{0}_N. \quad (10.1)$$

If in addition,  $f$  is twice continuously differentiable, then Euler's Theorem on homogeneous functions and Young's Theorem from calculus imply the following restrictions on the first and second order partial derivatives of  $f$ :

$$\mathbf{x}^T \nabla f(\mathbf{x}) = f(\mathbf{x}); \quad (1 \text{ restriction}) \quad (10.2)$$

$$\nabla^2 f(\mathbf{x}) \mathbf{x} = \mathbf{0}_N; \quad (N \text{ restrictions}) \quad (10.3)$$

$$\nabla^2 f(\mathbf{x}) = [\nabla^2 f(\mathbf{x})]^T \quad (N(N-1)/2 \text{ restrictions}). \quad (10.4)$$

The restrictions given by (10.2) and (10.3) are implied by Euler's Theorem and the symmetry restrictions (10.4) are implied by Young's Theorem.

A *flexible functional form*\*<sup>6</sup>  $f$  is a functional form that has enough parameters in it so that  $f$  can approximate an arbitrary twice continuously differentiable function  $f^*$  to the second order at an arbitrary point  $\mathbf{x}^*$  in the domain of definition of  $f$  and  $f^*$ . Thus  $f$  must have enough free parameters in order to satisfy the following  $1 + N + N^2$  equations:

$$f(\mathbf{x}^*) = f^*(\mathbf{x}^*); \quad (1 \text{ equation}) \quad (10.5)$$

$$\nabla f(\mathbf{x}^*) = \nabla f^*(\mathbf{x}^*); \quad (N \text{ equations}) \quad (10.6)$$

$$\nabla^2 f(\mathbf{x}^*) = \nabla^2 f^*(\mathbf{x}^*); \quad (N^2 \text{ equations}). \quad (10.7)$$

Of course, since both  $f$  and  $f^*$  are assumed to be twice continuously differentiable, we do not have to satisfy all  $N^2$  equations in (10.7) since Young's Theorem implies that  $\partial^2 f(\mathbf{x}^*)/\partial x_i \partial x_j = \partial^2 f(\mathbf{x}^*)/\partial x_j \partial x_i$  and  $\partial^2 f^*(\mathbf{x}^*)/\partial x_i \partial x_j = \partial^2 f^*(\mathbf{x}^*)/\partial x_j \partial x_i$  for all  $i$  and  $j$ . Thus the matrices of second order partial derivatives  $\nabla^2 f(\mathbf{x}^*)$  and  $\nabla^2 f^*(\mathbf{x}^*)$  are both symmetric matrices and so there are only  $N(N+1)/2$  independent equations to be satisfied in the restrictions (10.7). Thus a general flexible functional form must have at least  $1 + N + N(N+1)/2$  free parameters.

\*<sup>4</sup> This section is based on Diewert and Lawrence (2002)[143].

\*<sup>5</sup> This section is based on Diewert and Wales (1988b)[155].

\*<sup>6</sup> This terminology was introduced by Diewert (1974a; 133)[77].

The simplest example of a flexible functional form is the following *quadratic function*:

$$f(\mathbf{x}) \equiv a_0 + \mathbf{a}^T \mathbf{x} + (1/2)\mathbf{x}^T \mathbf{A} \mathbf{x}; \quad \mathbf{A} = \mathbf{A}^T \quad (10.8)$$

where  $a_0$  is a scalar parameter,  $\mathbf{a}^T \equiv [a_1, \dots, a_N]$  is a vector of parameters and  $\mathbf{A} \equiv [a_{ij}]$  is a symmetric matrix of parameters. Thus the  $f$  defined by (10.8) has  $1 + N + N(N + 1)/2$  parameters. To show that this  $f$  is flexible, we need to choose  $a_0, \mathbf{a}$  and  $\mathbf{A}$  to satisfy equations (10.5)-(10.7). Upon noting that  $\nabla f(\mathbf{x}) = \mathbf{a} + \mathbf{A} \mathbf{x}$  and  $\nabla^2 f(\mathbf{x}) = \mathbf{A}$ , equations (10.5)-(10.7) become the following equations:

$$a_0 + \mathbf{a}^T \mathbf{x}^* + (1/2)\mathbf{x}^{*T} \mathbf{A} \mathbf{x}^* = f^*(\mathbf{x}^*); \quad (10.9)$$

$$\mathbf{a} + \mathbf{A} \mathbf{x}^* = \nabla f^*(\mathbf{x}^*); \quad (10.10)$$

$$\mathbf{A} = \nabla^2 f^*(\mathbf{x}^*). \quad (10.11)$$

To satisfy these equations, choose  $\mathbf{A} \equiv \nabla^2 f^*(\mathbf{x}^*)$  (and  $\mathbf{A}$  will be a symmetric matrix since  $f^*$  is assumed to be twice continuously differentiable);  $\mathbf{a} \equiv \nabla f^*(\mathbf{x}^*) - \mathbf{A} \mathbf{x}^*$  and finally, choose  $a_0 \equiv f^*(\mathbf{x}^*) - [\mathbf{a}^T \mathbf{x}^* + (1/2)\mathbf{x}^{*T} \mathbf{A} \mathbf{x}^*]$ .

In many applications, we want to find a flexible functional form  $f$  that is also linearly homogeneous. For example, in production theory, if the minimum average cost plant size is small relative to the size of the market, then we can approximate the industry technology by means of a constant returns to scale production function. As another example, in the pure theory of international trade, we often assume that consumer preferences are *homothetic*<sup>\*7</sup>; i.e., the consumer's utility function can be represented by  $g[f(\mathbf{x})]$  where  $f$  is linearly homogeneous and  $g$  is a monotonically increasing and continuous function of one variable. In this case, we can represent the consumer's preferences equally well by using the linearly homogeneous utility function  $g[f(\mathbf{x})]$ .

If the production function  $f$  (or the utility function  $f$ ) is linearly homogeneous, then the corresponding cost function  $C$  has the following structure: for  $y > 0$  and  $\mathbf{p} \gg \mathbf{0}_N$ ,

$$\begin{aligned} C(y, \mathbf{p}) &\equiv \min_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : f(\mathbf{x}) \geq y\} \\ &= \min_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : f(\mathbf{x}) = y\} \quad \text{if } f \text{ is continuous and increasing in the components of } \mathbf{x} \\ &= \min_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : (1/y)f(\mathbf{x}) = 1\} \\ &= \min_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : f(\{1/y\}\mathbf{x}) = 1\} \quad \text{using the linear homogeneity of } f \\ &= \min_{\mathbf{x}/y} \{y\mathbf{p}^T (\mathbf{x}/y) : f(\mathbf{x}/y) = 1\} \\ &= y \min_{\mathbf{z}} \{\mathbf{p}^T \mathbf{z} : f(\mathbf{z}) = 1\} \quad \text{letting } \mathbf{z} \equiv \mathbf{x}/y \\ &= yC(1, \mathbf{p}) \\ &= yc(\mathbf{p}) \end{aligned} \quad (10.12)$$

where we define the *unit cost function*  $c(\mathbf{p})$  as  $C(1, \mathbf{p})$ , the minimum cost of producing one unit of output (or utility).

It is straightforward to show that  $C(1, \mathbf{p})$  and  $c(\mathbf{p})$  are linearly homogeneous and concave in the components of the price vector  $\mathbf{p}$ .

**Problem 1** Let  $\mathbf{y} \equiv [y_1, \dots, y_N]^T$  denote a vector of variable outputs and inputs that a firm produces or uses during a period; if the firm produces commodity  $i$ , then  $y_i > 0$  while if the firm uses commodity  $i$  as an input, then  $y_i < 0$  for  $i = 1, \dots, N$ . The vector  $\mathbf{y}$  is called a net output vector or a netput vector. Given the net output vector  $\mathbf{y}$ , the minimum amount of capital  $k \geq 0$  that is required

<sup>\*7</sup> Shephard (1953)[355] introduced this term.

to produce the vector of net outputs  $\mathbf{y}$  is  $F(\mathbf{y})$ , where  $F$  is the firm's capital requirements function.\*<sup>8</sup> Given a positive vector of variable input and output prices  $\mathbf{p} \gg \mathbf{0}_N$  and a positive amount of capital  $k > 0$ , the firm's variable profit function  $\Pi(k, \mathbf{p})$  is defined as follows\*<sup>9</sup>:

$$\Pi(k, \mathbf{p}) \equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : F(\mathbf{y}) \leq k\}. \quad (\text{i})$$

Prove that for each  $k > 0$ ,  $\Pi(k, \mathbf{p})$  is a linearly homogeneous and convex function of  $\mathbf{p}$ .

**Problem 2** (Continuation of 1.) Let  $\mathbf{y}^*$  solve the variable profit maximization problem  $\Pi(k^*, \mathbf{p}^*) \equiv \max_{\mathbf{y}} \{\mathbf{p}^{*T} \mathbf{y} : F(\mathbf{y}) \leq k^*\}$  where  $k^* > 0$  and  $\mathbf{p}^* \gg \mathbf{0}_N$ . Assume that  $\Pi(k^*, \mathbf{p}^*)$  is differentiable with respect to the components of  $\mathbf{p}$  at the point  $\mathbf{p}^*$ ; i.e., assume that the vector of first order partial derivatives  $\nabla_{\mathbf{p}} \Pi(k^*, \mathbf{p}^*)$  exists. Show that  $\mathbf{y}^* = \nabla_{\mathbf{p}} \Pi(k^*, \mathbf{p}^*)$ . This result is known as Hotelling's (1932; 594)[242] Lemma.

*Hint:* Define  $g(\mathbf{p}) \equiv \mathbf{p}^T \mathbf{y}^* - \Pi(k^*, \mathbf{p})$  and show that  $g(\mathbf{p}) \leq 0$  and  $g(\mathbf{p}^*) = 0$ .

**Problem 3** Assume that the capital requirements function  $F(\mathbf{y})$  is linearly homogeneous; i.e.,  $F(\lambda \mathbf{y}) = \lambda F(\mathbf{y})$  for all  $\lambda > 0$ . (This means that the technology exhibits *constant returns to scale*.) Under this assumption, show that  $\Pi(k, \mathbf{p})$  has the following decomposition: for  $k > 0$  and  $\mathbf{p} \gg \mathbf{0}_N$ ,

$$\Pi(k, \mathbf{p}) = k\Pi(1, \mathbf{p}).$$

The function  $\Pi(1, \mathbf{p}) \equiv \pi(\mathbf{p})$  is known as the firm's *unit (capital) profit function*. By problem 1 above, it too will be a linearly homogeneous function.

**Problem 4** Using problem 2 above, it can be seen that the firm's variable profit maximizing system of net supply functions,  $\mathbf{y}(k, \mathbf{p})$ , is equal to the vector of first order partial derivatives  $\nabla_{\mathbf{p}} \Pi(k, \mathbf{p})$ , provided that  $\Pi(k, \mathbf{p})$  is differentiable with respect to the components of the variable price vector  $\mathbf{p}$ . If  $\Pi(k, \mathbf{p})$  is twice continuously differentiable with respect to the components of  $\mathbf{p}$ , show that the  $N \times N$  matrix of price derivatives of the net supply functions,  $\nabla_{\mathbf{p}} \mathbf{y}(k, \mathbf{p}) \equiv [\partial y_i(k, \mathbf{p}) / \partial p_j]$ , has the following properties:

$$\nabla_{\mathbf{p}} \mathbf{y}(k, \mathbf{p}) = [\nabla_{\mathbf{p}} \Pi(k, \mathbf{p})]^T; \quad (\text{a})$$

$$[\nabla_{\mathbf{p}} \mathbf{y}(k, \mathbf{p})] \mathbf{p} = \mathbf{0}_N; \quad (\text{b})$$

$$\mathbf{z}^T \nabla_{\mathbf{p}} \mathbf{y}(k, \mathbf{p}) \mathbf{z} \geq 0 \text{ for every vector } \mathbf{z}; \quad (\text{c})$$

$$\mathbf{e}_i^T \nabla_{\mathbf{p}} \mathbf{y}(k, \mathbf{p}) \mathbf{e}_i \geq 0 \text{ for } i = 1, \dots, N \text{ where } \mathbf{e}_i \text{ is the } i\text{th unit vector. Provide an economic interpretation for these inequalities.} \quad (\text{d})$$

**Problem 5** (Continuation of 4.) Commodities  $i$  and  $j$  are said to be *substitutes* in production if  $\partial y_i(k, \mathbf{p}) / \partial p_j < 0$  for  $i \neq j$ . Commodities  $i$  and  $j$  are said to be *complements* in production if  $\partial y_i(k, \mathbf{p}) / \partial p_j > 0$  for  $i \neq j$ . Commodities  $i$  and  $j$  are said to be *unrelated* in production if  $\partial y_i(k, \mathbf{p}) / \partial p_j = 0$ . If  $N = 2$ , show that variable commodities 1 and 2 cannot be complements; i.e., they must be substitutes or be unrelated.

Linearly homogeneous functions arise naturally in a variety of economic applications. Moreover, even if we allow our production function or utility function  $f$  to be a general nonhomogeneous function, it is often of interest to allow  $f$  to have the capability to be flexible in the class of linearly homogeneous functions.

\*<sup>8</sup> If there is no amount of capital that can produce a given vector of net outputs  $\mathbf{y}$ , then we define  $F(\mathbf{y}) \equiv +\infty$ . For more on the properties of factor requirements functions, see Diewert (1974b)[78].

\*<sup>9</sup> We assume that for each  $k > 0$ , the lower level set of  $F$  defined as  $\{\mathbf{y} : F(\mathbf{y}) \leq k\}$  is a nonempty, closed and bounded set so that the maximum in (i) exists.

Consider what happens to the general quadratic function  $f$  defined by (10.8) if we attempt to specialize it to become a linearly homogeneous functional form. In order to make it homogeneous of degree one, we must set  $a_0 = 0$  and set  $\mathbf{A} = \mathbf{0}_{N \times N}$  and the resulting functional form collapses down to the following linear function:

$$f(\mathbf{x}) = \mathbf{a}^T \mathbf{x}. \quad (10.13)$$

But the  $f$  defined by (10.13) is not a flexible linearly homogeneous functional form! Thus finding flexible linearly homogeneous functional forms is not completely straightforward.

Let us determine the minimal number of free parameters that a flexible linearly homogeneous functional form must have. If both  $f$  and  $f^*$  are linearly homogeneous (and twice continuously differentiable), then both functions will satisfy the restrictions (10.2)-(10.4). In view of these restrictions, it can be seen that instead of  $f$  having to satisfy all  $1 + N + N^2$  of the equations (10.5)-(10.7),  $f$  need only satisfy the following  $N + N(N - 1)/2 = N(N + 1)/2$  equations:

$$\nabla f(\mathbf{x}^*) = \nabla f^*(\mathbf{x}^*); \quad (N \text{ equations}) \quad (10.14)$$

$$f_{ij}(\mathbf{x}^*) = f_{ij}^*(\mathbf{x}^*) \text{ for } 1 \leq i < j \leq N \quad (N(N - 1)/2 \text{ equations}) \quad (10.15)$$

where  $f_{ij}(\mathbf{x}^*) \equiv \partial^2 f(\mathbf{x}^*)/\partial x_i \partial x_j$ . Note that equations (10.15) are the equations in the upper triangle of the matrix equation (10.7) above. If the upper triangle equations in (10.7) are satisfied, then by Young's Theorem, the lower triangle equations will also be satisfied if equations (10.15) are satisfied. The main diagonal equations in (10.7) will also be satisfied if equations (10.15) are satisfied: the diagonal elements  $f_{ii}(\mathbf{x}^*)$  are determined by the restrictions  $\nabla^2 f(\mathbf{x}^*)\mathbf{x}^* = \mathbf{0}_N$  and the  $f_{ii}^*(\mathbf{x}^*)$  are determined by the restrictions  $\nabla^2 f^*(\mathbf{x}^*)\mathbf{x}^* = \mathbf{0}_N$ .

Thus in order for  $f(\mathbf{x})$  (or  $c(\mathbf{p})$  or  $\pi(\mathbf{p})$ ) to be a flexible linearly homogeneous functional form, it must have at least  $N + N(N - 1)/2 = N(N + 1)/2$  free parameters. If it has exactly this number of free parameters, then we say that  $f$  is a *parsimonious flexible functional form*.

In the following sections, we shall give some examples of parsimonious flexible functional forms for unit cost functions. These same functional forms can be used as parsimonious flexible functional forms for unit profit functions.\*<sup>10</sup> Thus we are looking for linearly homogeneous functions  $c(\mathbf{p})$  that can satisfy the following  $N(N + 1)/2$  equations:

$$\nabla c(\mathbf{p}^*) = \nabla c^*(\mathbf{p}^*); \quad (N \text{ equations}) \quad (10.16)$$

$$c_{ij}(\mathbf{p}^*) = c_{ij}^*(\mathbf{p}^*) \text{ for } 1 \leq i < j \leq N \quad (N(N - 1)/2 \text{ equations}). \quad (10.17)$$

Why is it important that functional forms used in applied economics be flexible? From Shephard's (1953; 11)[355] Lemma, the producer's system of cost minimizing input demand functions,  $\mathbf{x}(y, \mathbf{p})$ , is equal to the vector of first order partial derivatives of the cost function with respect of input prices,  $\nabla_{\mathbf{p}} C(y, \mathbf{p})$ . Thus the matrix of *first order input demand price derivatives*  $\nabla_{\mathbf{p}} \mathbf{x}(y, \mathbf{p})$  is equal to the matrix of second order partial derivatives  $\nabla_{\mathbf{p}\mathbf{p}}^2 C(y, \mathbf{p})$ . Hence, if the functional form for  $C$  is *not* flexible, *price elasticities of input demand will be a priori restricted in some arbitrary way*.<sup>\*11</sup> Many practical problems in applied economics depend crucially on estimates of elasticities and hence it is not appropriate to use estimates of elasticities that are restricted in some arbitrary manner.

In the following sections, we will exhibit some examples of flexible functional forms.

\*<sup>10</sup> The only difference is that the concavity in prices property for unit cost functions must be replaced by the convexity in prices property for unit profit functions.

\*<sup>11</sup> A similar comment applies in the profit function context; unless the variable profit function  $\Pi(k, \mathbf{p})$  is flexible, estimates of elasticities of net supply will be arbitrarily restricted.

### 10.3 The Generalized Leontief Cost Function

Define the *generalized Leontief unit cost function*  $c(\mathbf{p})$  as follows<sup>\*12</sup>:

$$c(p_1, \dots, p_N) \equiv \sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{1/2} p_j^{1/2}; \quad b_{ij} = b_{ji} \text{ for } 1 \leq i < j \leq N. \quad (10.18)$$

Thus  $c$  is a quadratic form in the square roots of input prices and has  $N(N+1)/2$   $b_{ij}$  parameters. We need to determine whether the unit cost function  $c(\mathbf{p})$  defined by (10.18) is flexible; i.e., whether we can choose the  $b_{ij}$  so as to satisfy equations (10.16) and (10.17). Upon differentiating (10.18), equations (10.16) and (10.17) become the following equations:

$$c_i(\mathbf{p}^*) = \sum_{j=1}^N b_{ij} (p_i^*)^{-(1/2)} (p_j^*)^{1/2} = c_i^*(\mathbf{p}^*); \quad i = 1, \dots, N; \quad (10.19)$$

$$c_{ij}(\mathbf{p}^*) = (1/2) b_{ij} (p_i^*)^{-(1/2)} (p_j^*)^{-(1/2)} = c_{ij}^*(\mathbf{p}^*); \quad 1 \leq i < j \leq N. \quad (10.20)$$

Use equations (10.20) to determine  $b_{ij}$  for  $1 \leq i < j \leq N$ . Then use equations (10.19) to solve for the  $b_{ii}$  for  $i = 1, \dots, N$ . This proves that the  $c(\mathbf{p})$  defined by (10.18) is flexible. Since it has only  $N(N+1)/2$  parameters, it is also parsimonious.

In a production study where there is only one output and  $N$  inputs and the assumption of competitive cost minimization is justified, given period  $t$  data on input demands,  $x_i^t$ , input prices,  $p_i^t$  and on output produced,  $y^t$ , then the unknown parameters in (10.18) can be estimated by using the following  $N$  estimating equations:

$$x_i^t/y^t = \sum_{j=1}^N b_{ij} (p_j^t/p_i^t)^{1/2} + e_i^t; \quad i = 1, \dots, N, \quad (10.21)$$

where the  $e_i^t$  are stochastic error terms for  $i = 1, \dots, N$ .

Note that  $b_{ij}$  in equation  $i$  should equal  $b_{ji}$  in equation  $j$ . These *cross equation symmetry restrictions* can be imposed in the estimation procedure or we could *test* for their validity.

After estimating the  $b_{ij}$ , it is necessary to check whether  $\nabla^2 c(\mathbf{p}^t)$  is *negative semidefinite* at each data point  $\mathbf{p}^t$ .<sup>\*13</sup> Thus it will be necessary to calculate the second order derivatives of  $c$  at each data point. Differentiating the  $c(\mathbf{p})$  defined by (10.18) yields the following formulae for the derivatives:

$$\begin{aligned} c_{ij}(\mathbf{p}^t) &= (1/2) b_{ij} (p_i^t p_j^t)^{-(1/2)} && \text{for } i \neq j; \\ c_{ii}(\mathbf{p}^t) &= -(1/2) \sum_{k \neq i, k=1}^N b_{ik} (p_i^t)^{-(3/2)} (p_k^t)^{(1/2)}; && \text{for } i = 1, \dots, N. \end{aligned} \quad (10.22)$$

Note that the  $b_{ii}$  do not appear in the formulae (10.22) for the second derivatives of the generalized Leontief unit cost function. Note also if all  $b_{ij} = 0$  for  $i \neq j$ , then the functional form defined by (10.18) collapses down to the no substitution Leontief (1941)[290] functional form<sup>\*14</sup>. Under these restrictions, the input demand functions defined by (10.21) collapse down to the following system of equations:

$$x_i^t/y^t = b_{ii} + e_i^t; \quad i = 1, \dots, N. \quad (10.23)$$

Thus input demands are not affected by changes in input prices if the producer's cost function has the Leontief functional form.

<sup>\*12</sup> This functional form was introduced by Diewert (1971)[73].

<sup>\*13</sup> A necessary and sufficient condition for a twice continuously differentiable  $c(\mathbf{p})$  to be concave over a convex set  $S$  is that  $\nabla^2 c(\mathbf{p})$  be negative semidefinite for all  $\mathbf{p}$  belonging to  $S$ .

<sup>\*14</sup> This functional form was actually used by Walras (1954; 243)[387]; the first edition of this book was published in 1874.

**Problem 6** Let  $N = 2$  and try to determine necessary and sufficient conditions on the parameters  $b_{11}$ ,  $b_{12}$  and  $b_{22}$  that will make the generalized Leontief unit cost function defined by (10.18),  $c(p_1, p_2)$ , concave in the input prices  $(p_1, p_2)$ . Look at the system of estimating equations (10.21) when  $N = 2$ . Can you determine a simple method for making sure that your estimated generalized Leontief unit cost function will satisfy the concavity property?

**Problem 7** Determine a simple set of sufficient conditions that will make the generalized Leontief unit cost function defined by (10.18) concave in  $\mathbf{p}$  for an arbitrary  $N$  over the set  $S \equiv \{\mathbf{p} : \mathbf{p} \gg \mathbf{0}_N\}$ .

## 10.4 The Translog Unit Cost Function

The translog unit cost function,  $c(\mathbf{p})$ , is defined as follows:<sup>\*15</sup>

$$\ln c(\mathbf{p}) \equiv \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln p_i \ln p_j \quad (10.24)$$

where the parameters  $\alpha_i$  and  $\gamma_{ij}$  satisfy the following restrictions:

$$\gamma_{ij} = \gamma_{ji}; \quad 1 \leq i < j \leq N; \quad (N(N-1)/2 \text{ symmetry restrictions}) \quad (10.25)$$

$$\sum_{i=1}^N \alpha_i = 1; \quad (1 \text{ restriction}) \quad (10.26)$$

$$\sum_{j=1}^N \gamma_{ij} = 0; \quad i = 1, \dots, N \quad (N \text{ restrictions}). \quad (10.27)$$

Note that the symmetry restrictions (10.25) and the restrictions (10.27) imply the following restrictions:

$$\sum_{i=1}^N \gamma_{ij} = 0; \quad j = 1, \dots, N. \quad (10.28)$$

There are  $1 + N$   $\alpha_i$  parameters and  $N^2$   $\gamma_{ij}$  parameters. However, the restrictions (10.25)-(10.27) mean that there are only  $N$  independent  $\alpha_i$  parameters and  $N(N-1)/2$  independent  $\gamma_{ij}$  parameters, which is the minimal number of parameters required for a unit cost function to be flexible.

We show that the translog unit cost function  $c(\mathbf{p})$  defined by (10.24)-(10.27) is linearly homogeneous; i.e., we need to show that  $c(\lambda\mathbf{p}) = \lambda c(\mathbf{p})$  for  $\lambda > 0$  and  $\mathbf{p} \gg \mathbf{0}_N$ . Thus, we need to show that

$$\ln c(\lambda\mathbf{p}) = \ln[\lambda c(\mathbf{p})] = \ln \lambda + \ln c(\mathbf{p}) \quad \text{for } \lambda > 0 \text{ and } \mathbf{p} \gg \mathbf{0}_N. \quad (10.29)$$

<sup>\*15</sup> This functional form is due to Christensen, Jorgenson and Lau (1971)[53] (1975)[55].

Using definition (10.24), we have

$$\begin{aligned}
\ln c(\lambda p_1, \dots, \lambda p_N) &= \alpha_0 + \sum_{i=1}^N \alpha_i \ln \lambda p_i + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln \lambda p_i \ln \lambda p_j \\
&= \alpha_0 + \sum_{i=1}^N \alpha_i [\ln \lambda + \ln p_i] + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} [\ln \lambda + \ln p_i] [\ln \lambda + \ln p_j] \\
&= \alpha_0 + \sum_{i=1}^N \alpha_i [\ln \lambda] + \sum_{i=1}^N \alpha_i \ln p_i + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} [\ln \lambda + \ln p_i] [\ln \lambda + \ln p_j] \\
&= \alpha_0 + 1[\ln \lambda] + \sum_{i=1}^N \alpha_i \ln p_i + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} [\ln \lambda + \ln p_i] [\ln \lambda + \ln p_j] \quad \text{using (10.26)} \\
&= \ln \lambda + \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} [\ln \lambda] [\ln \lambda] \\
&\quad + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} [\ln \lambda] [\ln p_j] + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} [\ln p_i] [\ln \lambda] \\
&\quad + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} [\ln p_i] [\ln p_j] \\
&= \ln \lambda + \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i + (1/2) \sum_{i=1}^N \left[ \sum_{j=1}^N \gamma_{ij} \right] [\ln \lambda] [\ln \lambda] \\
&\quad + (1/2) \sum_{j=1}^N \left[ \sum_{i=1}^N \gamma_{ij} \right] [\ln p_j] [\ln \lambda] + (1/2) \sum_{i=1}^N \left[ \sum_{j=1}^N \gamma_{ij} \right] [\ln p_i] [\ln \lambda] \\
&\quad + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} [\ln p_i] [\ln p_j] \\
&= \ln \lambda + \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i + (1/2) \sum_{i=1}^N [0] [\ln \lambda] [\ln \lambda] \\
&\quad + (1/2) \sum_{j=1}^N [0] [\ln p_j] [\ln \lambda] + (1/2) \sum_{i=1}^N [0] [\ln p_i] [\ln \lambda] \\
&\quad + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} [\ln p_i] [\ln p_j] \quad \text{using (10.27) and (10.28)} \\
&= \ln \lambda + \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} [\ln p_i] [\ln p_j] \\
&= \ln \lambda + \ln c(\mathbf{p}) \quad \text{using definition (10.24)} \tag{10.30}
\end{aligned}$$

which establishes the linear homogeneity property (10.29). Thus the restrictions (10.25)-(10.27) are just the right ones to imply the linear homogeneity of the translog unit cost function.

To establish the flexibility of the translog unit cost function  $c(\mathbf{p})$  defined by (10.24)-(10.27), we need only solve the following system of equations, which is equivalent to the  $N(N+1)/2$  equations defined by (10.16) and (10.17):

$$\ln c(\mathbf{p}^*) = \ln c^*(\mathbf{p}^*); \tag{1 equation} \tag{10.31}$$

$$\partial \ln c(\mathbf{p}^*) / \partial \ln p_i = \partial \ln c^*(\mathbf{p}^*) / \partial \ln p_i; \quad i = 1, 2, \dots, N-1; \quad (N-1 \text{ equations}) \tag{10.32}$$

$$\partial^2 \ln c(\mathbf{p}^*) / \partial \ln p_i \partial \ln p_j = \partial^2 \ln c^*(\mathbf{p}^*) / \partial \ln p_i \partial \ln p_j; \quad 1 \leq i < j \leq N; \quad (N(N-1)/2 \text{ equations}). \tag{10.33}$$

Upon differentiating the translog unit cost function defined by (10.24), we see that equations (10.32) are equivalent to the following equations:

$$\alpha_i + \sum_{j=1}^N \gamma_{ij} \ln p_j = \partial \ln c^*(\mathbf{p}^*) / \partial \ln p_i; \quad i = 1, 2, \dots, N-1. \tag{10.34}$$

Differentiating the translog unit cost function again, we find that equations (10.33) are equivalent

to the following equations:

$$\gamma_{ij} = \partial^2 \ln c^*(\mathbf{p}^*) / \partial \ln p_i \partial \ln p_j; \quad 1 \leq i < j \leq N. \quad (10.35)$$

Now use equations (10.35) to determine the  $\gamma_{ij}$  for  $1 \leq i < j \leq N$ . Now use the symmetry restrictions (10.25) to determine the  $\gamma_{ij}$  for  $1 \leq j < i \leq N$ . Now use equations (10.27) to determine the  $\gamma_{ii}$  for  $i = 1, 2, \dots, N$ . With the entire  $N \times N$  matrix of the  $\gamma_{ij}$  now determined, use equations (10.34) in order to determine the  $\alpha_i$  for  $i = 1, 2, \dots, N - 1$ . Now use equation (10.26) to determine  $\alpha_N$ . Finally, use equation (10.31) to determine  $\alpha_0$ .

We turn our attention to the problems involved in obtaining estimates for the unknown parameters  $\alpha_i$  and  $\gamma_{ij}$ , which occur in the definition of the translog unit cost function,  $c(\mathbf{p})$  defined by (10.24). In the producer context, the total cost function  $C(y, \mathbf{p})$  is defined in terms of the unit cost function  $c(\mathbf{p})$  as follows:

$$C(y, \mathbf{p}) \equiv y c(\mathbf{p}). \quad (10.36)$$

Taking logarithms on both sides of (10.36) yields:

$$\begin{aligned} \ln C(y, \mathbf{p}) &= \ln y + \ln c(\mathbf{p}) \\ &= \ln y + \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln p_i \ln p_j \end{aligned} \quad (10.37)$$

where we have replaced  $\ln c(\mathbf{p})$  using (10.24). The corresponding system of cost minimizing input demand functions  $\mathbf{x}(y, \mathbf{p})$  is obtained using Shephard's Lemma:

$$\mathbf{x}(y, \mathbf{p}) \equiv \nabla_{\mathbf{p}} C(y, \mathbf{p}) = y \nabla_{\mathbf{p}} c(\mathbf{p}). \quad (10.38)$$

Suppose that in period  $t$ , observed output is  $y^t$ , the vector of observed input prices is  $\mathbf{p}^t \gg \mathbf{0}_N$  and the vector of observed input demands is  $\mathbf{x}^t > \mathbf{0}_N$ . Thus the *period  $t$  observed cost* is:

$$C^t \equiv \mathbf{p}^{tT} \mathbf{x}^t \equiv \sum_{i=1}^N p_i^t x_i^t. \quad (10.39)$$

Now evaluate (10.37) at the period  $t$  data and add an error term,  $e_0^t$ . Using (10.39), (10.37) evaluated at the period  $t$  data becomes the following estimating equation:

$$\ln C^t = \ln y^t + \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i^t + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln p_i^t \ln p_j^t + e_0^t; \quad t = 1, \dots, T. \quad (10.40)$$

Note that (10.40) is linear in the unknown parameters.

In order to obtain additional estimating equations, we have to use the input demand functions,  $x_i(y, \mathbf{p}) \equiv y \partial c(\mathbf{p}) / \partial p_i$  for  $i = 1, \dots, N$ ; (see equations (10.38) above). The  $i$ th input share function,  $s_i(y, \mathbf{p})$ , is defined as:

$$\begin{aligned} s_i(y, \mathbf{p}) &\equiv p_i x_i(y, \mathbf{p}) / C(y, \mathbf{p}) && i = 1, \dots, N \\ &= p_i [y \partial c(\mathbf{p}) / \partial p_i] / C(y, \mathbf{p}) && \text{using (10.38)} \\ &= p_i [y \partial c(\mathbf{p}) / \partial p_i] / y c(\mathbf{p}) && \text{using (10.36)} \\ &= p_i [\partial c(\mathbf{p}) / \partial p_i] / c(\mathbf{p}) \\ &= \partial \ln c(\mathbf{p}) / \partial \ln p_i \\ &= \alpha_i + \sum_{j=1}^N \gamma_{ij} \ln p_j && \text{upon differentiating the } c(\mathbf{p}) \text{ defined by (10.24)}. \end{aligned} \quad (10.41)$$

Now evaluate both sides of (10.41) at the period  $t$  data and add error terms  $e_i^t$  to obtain the following system of estimating equations:

$$s_i^t \equiv p_i^t x_i^t / C^t = \alpha_i + \sum_{j=1}^N \gamma_{ij} \ln p_j^t + e_i^t; \quad i = 1, \dots, N. \quad (10.42)$$

Note that equations (10.42) are also linear in the unknown parameters. Obviously, the  $N$  estimating equations in (10.42) could be added to the single estimating equation (10.40) in order to obtain  $N + 1$  estimating equations with cross equation equality constraints on the parameters  $\alpha_i$  and  $\gamma_{ij}$ . However, since total cost in any period  $t$ ,  $C^t$ , equals the sum of the individual expenditures on the inputs<sup>\*16</sup>,  $\sum_{i=1}^N p_i^t x_i^t$ , the observed input shares  $s_i^t \equiv p_i^t x_i^t / C^t$  will satisfy the following constraint for each period  $t$ :

$$\sum_{i=1}^N s_i^t = 1. \quad (10.43)$$

Thus the stochastic error terms  $e_i^t$  in equations (10.42) cannot all be independent. Hence we must drop one estimating equation from (10.42). Thus equation (10.40) and any  $N - 1$  of the  $N$  equations in (10.42) may be used as a system of estimating equations in order to determine the parameters of the translog unit cost function.<sup>\*17</sup>

We now turn our attention to the problem of deriving a formula for the price elasticities of demand,  $\partial x_i(\mathbf{y}, \mathbf{p}) / \partial p_j$ , given that the unit cost function has the translog functional form defined by (10.24)-(10.27). Using the equations in (10.41) above, we have the following expressions for the  $i$ th input share functions,  $s_i(\mathbf{y}, \mathbf{p})$ :

$$s_i(\mathbf{y}, \mathbf{p}) = p_i x_i(\mathbf{y}, \mathbf{p}) / C(\mathbf{y}, \mathbf{p}) = \partial \ln c(\mathbf{p}) / \partial \ln p_i = \alpha_i + \sum_{j=1}^N \gamma_{ij} \ln p_j; \quad i = 1, \dots, N. \quad (10.44)$$

For  $j \neq i$ , differentiate the  $i$ th equation in (10.44) with respect to the log of  $p_j$  and we obtain the following equations:

$$\partial s_i(\mathbf{y}, \mathbf{p}) / \partial \ln p_j = p_i \partial [x_i(\mathbf{y}, \mathbf{p}) / C(\mathbf{y}, \mathbf{p})] / \partial \ln p_j = \gamma_{ij}; \quad i \neq j. \quad (10.45)$$

Hence

$$\begin{aligned} \gamma_{ij} &= p_i \partial [x_i(\mathbf{y}, \mathbf{p}) / C(\mathbf{y}, \mathbf{p})] / \partial \ln p_j \quad i \neq j \\ &= p_i p_j \partial [x_i(\mathbf{y}, \mathbf{p}) / C(\mathbf{y}, \mathbf{p})] / \partial p_j \\ &= p_i p_j \left\{ [1/C(\mathbf{y}, \mathbf{p})] [\partial x_i(\mathbf{y}, \mathbf{p}) / \partial p_j] - x_i(\mathbf{y}, \mathbf{p}) [1/C(\mathbf{y}, \mathbf{p})]^2 [\partial C(\mathbf{y}, \mathbf{p}) / \partial p_j] \right\} \\ &= [p_i x_i(\mathbf{y}, \mathbf{p}) / C(\mathbf{y}, \mathbf{p})] \left\{ \partial \ln x_i(\mathbf{y}, \mathbf{p}) / \partial \ln p_j \right\} - [p_i x_i(\mathbf{y}, \mathbf{p}) / C(\mathbf{y}, \mathbf{p})] [p_j x_j(\mathbf{y}, \mathbf{p}) / C(\mathbf{y}, \mathbf{p})] \\ &\quad \text{using Shephard's Lemma, } x_j(\mathbf{y}, \mathbf{p}) = \partial C(\mathbf{y}, \mathbf{p}) / \partial p_j \\ &= s_i(\mathbf{y}, \mathbf{p}) \left\{ \partial \ln x_i(\mathbf{y}, \mathbf{p}) / \partial \ln p_j \right\} - s_i(\mathbf{y}, \mathbf{p}) s_j(\mathbf{y}, \mathbf{p}). \end{aligned} \quad (10.46)$$

<sup>\*16</sup> This identity explains why we did not add the counterpart to (10.40) as an estimating equation to the estimating equations (10.21) in the previous section.

<sup>\*17</sup> In situations where  $N$  is large relative to the number of observations  $T$ , maximum likelihood estimation of equation (10.40) and  $N - 1$  of the equations (10.41) can fail if a general variance covariance matrix has to be estimated for the error terms in these equations. The problem is that all of the unknown economic parameters are contained in equation (10.40) and as a result, the estimated squared residuals in this equation will tend to be small relative to the estimated squared residuals in equations (10.41), where each equation has only a few unknown economic parameters. Hence equation (10.40) can suffer from multicollinearity problems and the small apparent variance of the residuals in this equation lead to the maximum likelihood estimation procedure giving too much weight to equation (10.40) relative to the other equations. Under these conditions, the resulting elasticities may be erratic and not satisfy the appropriate curvature conditions.

Equations (10.46) can be rearranged to give us the following formula for the *cross price elasticities of input demand*:

$$\partial \ln x_i(y, \mathbf{p}) / \partial \ln p_j = [s_i(y, \mathbf{p})]^{-1} \gamma_{ij} + s_j(y, \mathbf{p}); \quad i \neq j. \quad (10.47)$$

Now differentiate the  $i$ th equation in (10.44) with respect to the log of  $p_i$  and get the following equations:

$$\begin{aligned} \gamma_{ii} &= p_i \partial [p_i x_i(y, \mathbf{p}) / C(y, \mathbf{p})] / \partial p_i \quad i = 1, \dots, N \\ &= p_i \{ [x_i(y, \mathbf{p}) / C(y, \mathbf{p})] + [p_i / C(y, \mathbf{p})] [\partial x_i(y, \mathbf{p}) / \partial p_i] - [p_i x_i(y, \mathbf{p}) / C(y, \mathbf{p})^2] [\partial C(y, \mathbf{p}) / \partial p_i] \} \\ &= p_i \{ [x_i(y, \mathbf{p}) / C(y, \mathbf{p})] + [p_i / C(y, \mathbf{p})] [\partial x_i(y, \mathbf{p}) / \partial p_i] - [p_i x_i(y, \mathbf{p}) / C(y, \mathbf{p})^2] [x_i(y, \mathbf{p})] \} \\ &\quad \text{using Shephard's Lemma, } x_i(y, \mathbf{p}) = \partial C(y, \mathbf{p}) / \partial p_i \\ &= p_i x_i(y, \mathbf{p}) / C(y, \mathbf{p}) + [p_i x_i(y, \mathbf{p}) / C(y, \mathbf{p})] [\partial \ln x_i(y, \mathbf{p}) / \partial \ln p_i] - [p_i x_i(y, \mathbf{p}) / C(y, \mathbf{p})]^2 \\ &= s_i(y, \mathbf{p}) + s_i(y, \mathbf{p}) [\partial \ln x_i(y, \mathbf{p}) / \partial \ln p_i] - s_i(y, \mathbf{p})^2. \end{aligned} \quad (10.48)$$

Equations (10.48) can be rearranged to give us the following formula for the *own price elasticities of input demand*:

$$\partial \ln x_i(y, \mathbf{p}) / \partial \ln p_i = [s_i(y, \mathbf{p})]^{-1} \gamma_{ii} + s_i(y, \mathbf{p}) - 1; \quad i = 1, \dots, N. \quad (10.49)$$

Thus given econometric estimates for the  $\alpha_i$  and  $\gamma_{ij}$ , which we denote by  $\alpha_i^*$  and  $\gamma_{ij}^*$ , the estimated or fitted shares in period  $t$ ,  $s_i^{t*}$  are defined using these estimates and equations (10.44) evaluated at the period  $t$  data:

$$s_i^{t*} \equiv \alpha_i^* + \sum_{j=1}^N \gamma_{ij}^* \ln p_j^t; \quad i = 1, \dots, N; t = 1, \dots, T. \quad (10.50)$$

Now use equations (10.47) evaluated at the period  $t$  data and econometric estimates to obtain the following formula for the *period  $t$  cross elasticities of demand*,  $e_{ij}^t$ :

$$e_{ij}^t \equiv \partial \ln x_i(y^t, \mathbf{p}^t) / \partial \ln p_j = [s_i^{t*}]^{-1} \gamma_{ij}^* + s_j^{t*}; \quad i \neq j. \quad (10.51)$$

Similarly, use equations (10.49) evaluated at the period  $t$  data and econometric estimates to obtain the following formula for the *period  $t$  own elasticities of demand*,  $e_{ii}^t$ :

$$e_{ii}^t \equiv \partial \ln x_i(y^t, \mathbf{p}^t) / \partial \ln p_i = [s_i^{t*}]^{-1} \gamma_{ii}^* + s_i^{t*} - 1; \quad i = 1, \dots, N. \quad (10.52)$$

We can also obtain an estimated or *fitted period  $t$  cost*,  $C^{t*}$ , by using our econometric estimates for the parameters and by exponentiating the right hand side of the equation  $t$  in (10.40):

$$C^{t*} \equiv \exp \left[ \ln y^t + \alpha_0^* + \sum_{i=1}^N \alpha_i^* \ln p_i^t + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij}^* \ln p_i^t \ln p_j^t \right]; \quad t = 1, \dots, T. \quad (10.53)$$

Finally, our fitted period  $t$  shares  $s_i^{t*}$  defined by (10.50) and our fitted period  $t$  costs  $C^{t*}$  defined by (10.53) can be used in order to obtain estimated or *fitted period  $t$  input demands*,  $x_i^{t*}$ , as follows:

$$x_i^{t*} \equiv C^{t*} s_i^{t*} / p_i^t; \quad i = 1, \dots, N; t = 1, \dots, T. \quad (10.54)$$

Given the matrix of period  $t$  estimated input price elasticities of demand,  $[e_{ij}^t]$ , we can readily calculate the matrix of period  $t$  *estimated input price derivatives*,  $\nabla_{\mathbf{p}} \mathbf{x}(y^t, \mathbf{p}^t) = \nabla_{\mathbf{p}}^2 C(y^t, \mathbf{p}^t)$ . Our estimate for element  $ij$  of  $\nabla_{\mathbf{p}}^2 C(y^t, \mathbf{p}^t)$  is:

$$C_{ij}^{t*} \equiv e_{ij}^t x_i^{t*} / p_j^t; \quad i, j = 1, \dots, N; t = 1, \dots, T \quad (10.55)$$

where the estimated period  $t$  elasticities  $e_{ij}^t$  are defined by (10.51) and (10.52) and the fitted period  $t$  input demands  $x_i^{t*}$  are defined by (10.54). Once the estimated input price derivative matrices  $[C_{ij}^{t*}]$  have been calculated, then we may check whether each of them is negative semidefinite using determinantal conditions or by checking if all of the eigenvalues of each matrix are zero or negative. Unfortunately, *very frequently these negative semidefiniteness conditions will fail to be satisfied for both the translog and generalized Leontief functional forms*. Hence, in the following section, we study a functional form where these curvature conditions can be imposed without destroying the flexibility of the functional form.

## 10.5 The Normalized Quadratic Unit Cost Function

The normalized quadratic unit cost function  $c(\mathbf{p})$  is defined as follows for  $\mathbf{p} \gg \mathbf{0}_N$ :<sup>\*18</sup>

$$c(\mathbf{p}) \equiv \mathbf{b}^T \mathbf{p} + (1/2) \mathbf{p}^T \mathbf{B} \mathbf{p} / \alpha^T \mathbf{p} \quad (10.56)$$

where  $\mathbf{b}^T \equiv [b_1, \dots, b_N]$  and  $\alpha^T \equiv [\alpha_1, \dots, \alpha_N]$  are parameter vectors and  $\mathbf{B} \equiv [b_{ij}]$  is a matrix of parameters. The vector  $\alpha$  and the matrix  $\mathbf{B}$  satisfy the following restrictions:

$$\alpha > \mathbf{0}_N; \quad (10.57)$$

$$\mathbf{B} = \mathbf{B}^T; \quad \text{i.e., the matrix } \mathbf{B} \text{ is symmetric;} \quad (10.58)$$

$$\mathbf{B} \mathbf{p}^* = \mathbf{0}_N \text{ for some } \mathbf{p}^* \gg \mathbf{0}_N. \quad (10.59)$$

In most empirical applications, the vector of nonnegative but nonzero parameters  $\alpha$  is fixed a priori. The two most frequent a priori choices for  $\alpha$  are  $\alpha \equiv \mathbf{1}_N$ , a vector of ones or  $\alpha \equiv (1/T) \sum_{t=1}^T \mathbf{x}^t$ , the sample mean of the observed input vectors in the producer context or the sample mean of the observed commodity vectors in the consumer context. The two most frequent choices for the reference price vector  $\mathbf{p}^*$  are  $\mathbf{p}^* \equiv \mathbf{1}_N$  or  $\mathbf{p}^* \equiv \mathbf{p}^t$  for some period  $t$ ; i.e., in this second choice, we simply set  $\mathbf{p}^*$  equal to the observed period  $t$  price vector.

Assuming that  $\alpha$  has been predetermined, there are  $N$  unknown parameters in the  $\mathbf{b}$  vector and  $N(N-1)/2$  unknown parameters in the  $\mathbf{B}$  matrix, taking into account the symmetry restrictions (10.58) and the  $N$  linear restrictions in (10.59). Note that the  $c(\mathbf{p})$  defined by (10.56) is linearly homogeneous in the components of the input price vector  $\mathbf{p}$ .

Another possible way of defining the normalized quadratic unit cost function is as follows:

$$c(\mathbf{p}) \equiv (1/2) \mathbf{p}^T \mathbf{A} \mathbf{p} / \alpha^T \mathbf{p} \quad (10.60)$$

where the parameter matrix  $\mathbf{A}$  is symmetric; i.e.,  $\mathbf{A} = \mathbf{A}^T \equiv [a_{ij}]$  and  $\alpha > \mathbf{0}_N$  as before. Assuming that the vector of parameters  $\alpha$  has been predetermined, the  $c(\mathbf{p})$  defined by (10.60) has  $N(N+1)/2$  unknown  $a_{ij}$  parameters.

Comparing (10.56) with (10.60), it can be seen that (10.60) has dropped the  $\mathbf{b}$  vector but has also dropped the  $N$  linear constraints (10.59). It can be shown that the model defined by (10.56) is a special case of the model defined by (10.60). To show this, given (10.56), define the matrix  $\mathbf{A}$  in terms of  $\mathbf{B}$ ,  $\mathbf{b}$  and  $\alpha$  as follows:

$$\mathbf{A} \equiv \mathbf{B} + [\mathbf{b} \alpha^T + \alpha \mathbf{b}^T]. \quad (10.61)$$

<sup>\*18</sup> This functional form was introduced by Diewert and Wales (1987; 53)[153] where it was called the Symmetric Generalized McFadden functional form. Additional material on this functional form can be found in Diewert and Wales (1988a)[154] (1988b)[155] (1992)[156] (1993)[157].

Substituting (10.61) into (10.60), (10.60) becomes:

$$\begin{aligned}
c(\mathbf{p}) &= (1/2)\mathbf{p}^T\{\mathbf{B} + [\mathbf{b}\boldsymbol{\alpha}^T + \boldsymbol{\alpha}\mathbf{b}^T]\}\mathbf{p}/\boldsymbol{\alpha}^T\mathbf{p} \\
&= (1/2)\mathbf{p}^T\mathbf{B}\mathbf{p}/\boldsymbol{\alpha}^T\mathbf{p} + (1/2)\mathbf{p}^T[\mathbf{b}\boldsymbol{\alpha}^T + \boldsymbol{\alpha}\mathbf{b}^T]\mathbf{p}/\boldsymbol{\alpha}^T\mathbf{p} \\
&= (1/2)\mathbf{p}^T\mathbf{B}\mathbf{p}/\boldsymbol{\alpha}^T\mathbf{p} + (1/2)\{\mathbf{p}^T\mathbf{b}\boldsymbol{\alpha}^T\mathbf{p} + \mathbf{p}^T\boldsymbol{\alpha}\mathbf{b}^T\mathbf{p}\}/\boldsymbol{\alpha}^T\mathbf{p} \\
&= (1/2)\mathbf{p}^T\mathbf{B}\mathbf{p}/\boldsymbol{\alpha}^T\mathbf{p} + (1/2)\{2\mathbf{p}^T\mathbf{b}\boldsymbol{\alpha}^T\mathbf{p}\}/\boldsymbol{\alpha}^T\mathbf{p} \\
&= (1/2)\mathbf{p}^T\mathbf{B}\mathbf{p}/\boldsymbol{\alpha}^T\mathbf{p} + \mathbf{p}^T\mathbf{b}
\end{aligned} \tag{10.62}$$

which is the same functional form as (10.56). However, we prefer to work with the model (10.56) rather than with the seemingly more general model (10.60) for three reasons:

- The  $c(\mathbf{p})$  defined by (10.56) clearly contains the no substitution Leontief functional form as a special case (simply set  $\mathbf{B} = \mathbf{0}_{N \times N}$ );
- the estimating equations that correspond to (10.56) will contain constant terms and
- it is easier to establish the flexibility property for (10.56) than for (10.60).

The first and second order partial derivatives of the normalized quadratic unit cost function defined by (10.56) are given by:

$$\nabla_{\mathbf{p}}c(\mathbf{p}) = \mathbf{b} + (\boldsymbol{\alpha}^T\mathbf{p})^{-1}\mathbf{B}\mathbf{p} - (1/2)(\boldsymbol{\alpha}^T\mathbf{p})^{-2}\mathbf{p}^T\mathbf{B}\mathbf{p}\boldsymbol{\alpha}; \tag{10.63}$$

$$\nabla_{\mathbf{p}\mathbf{p}}^2c(\mathbf{p}) = (\boldsymbol{\alpha}^T\mathbf{p})^{-1}\mathbf{B} - (\boldsymbol{\alpha}^T\mathbf{p})^{-2}\mathbf{B}\mathbf{p}\boldsymbol{\alpha}^T - (\boldsymbol{\alpha}^T\mathbf{p})^{-2}\boldsymbol{\alpha}\mathbf{p}^T\mathbf{B} + (\boldsymbol{\alpha}^T\mathbf{p})^{-3}\mathbf{p}^T\mathbf{B}\mathbf{p}\boldsymbol{\alpha}\boldsymbol{\alpha}^T. \tag{10.64}$$

We now prove that the  $c(\mathbf{p})$  defined by (10.56)-(10.59) (with  $\boldsymbol{\alpha}$  predetermined) is a flexible functional form at the point  $\mathbf{p}^*$ . Using the restrictions (10.59),  $\mathbf{B}\mathbf{p}^* = \mathbf{0}_N$ , we have  $\mathbf{p}^{*T}\mathbf{B}\mathbf{p}^* = \mathbf{p}^{*T}\mathbf{0}_N = 0$ . Thus evaluating (10.63) and (10.64) at  $\mathbf{p} = \mathbf{p}^*$  yields the following equations:

$$\nabla_{\mathbf{p}}c(\mathbf{p}^*) = \mathbf{b}; \tag{10.65}$$

$$\nabla_{\mathbf{p}\mathbf{p}}^2c(\mathbf{p}^*) = (\boldsymbol{\alpha}^T\mathbf{p}^*)^{-1}\mathbf{B}. \tag{10.66}$$

We need to satisfy equations (10.16) and (10.17) above to show that the  $c(\mathbf{p})$  defined by (10.56)-(10.59) is flexible at  $\mathbf{p}^*$ . Using (10.65), we can satisfy equations (10.16) if we choose  $\mathbf{b}$  as follows:

$$\mathbf{b} \equiv \nabla c^*(\mathbf{p}^*). \tag{10.67}$$

Using (10.66), we can satisfy equations (10.17) by choosing  $\mathbf{B}$  as follows:

$$\mathbf{B} \equiv (\boldsymbol{\alpha}^T\mathbf{p}^*)\nabla^2c^*(\mathbf{p}^*). \tag{10.68}$$

Since  $\nabla^2c^*(\mathbf{p}^*)$  is a symmetric matrix,  $\mathbf{B}$  will also be a symmetric matrix and so the symmetry restrictions (10.58) will be satisfied for the  $\mathbf{B}$  defined by (10.68). Moreover, since  $c^*(\mathbf{p})$  is assumed to be a linearly homogeneous function, Euler's Theorem implies that

$$\nabla^2c^*(\mathbf{p}^*)\mathbf{p}^* = \mathbf{0}_N. \tag{10.69}$$

Equations (10.68) and (10.69) imply that the  $\mathbf{B}$  defined by (10.68) satisfies the linear restrictions (10.59). This completes the proof of the flexibility property for the normalized quadratic unit cost function.

It is convenient to define the vector of *normalized input prices*,  $\mathbf{v}^T \equiv [v_1, \dots, v_N]$  as follows:

$$\mathbf{v} \equiv (\mathbf{p}^T\boldsymbol{\alpha})^{-1}\mathbf{p}. \tag{10.70}$$

The system of input demand functions  $\mathbf{x}(y, \mathbf{p})$  that corresponds to the normalized quadratic unit cost function  $c(\mathbf{p})$  defined by (10.56) can be obtained using Shephard's Lemma in the usual way:

$$\mathbf{x}(y, \mathbf{p}) = y \nabla c(\mathbf{p}). \quad (10.71)$$

Using (10.71) and (10.63) evaluated at the period  $t$  data, we obtain the following system of *estimating equations*:

$$\mathbf{x}^t / y^t = \mathbf{b} + \mathbf{B} \mathbf{v}^t - (1/2) \mathbf{v}^{tT} \mathbf{B} \mathbf{v}^t \boldsymbol{\alpha} + \mathbf{e}^t; \quad t = 1, \dots, T \quad (10.72)$$

where  $\mathbf{x}^t$  is the observed period  $t$  input vector,  $y^t$  is the period  $t$  output,  $\mathbf{v}^t \equiv \mathbf{p}^t / \boldsymbol{\alpha}^T \mathbf{p}^t$  is the vector of period  $t$  normalized input prices and  $\mathbf{e}^t \equiv [e_1^t, \dots, e_N^t]^T$  is a vector of stochastic error terms. Equations (10.72) can be used in order to statistically estimate the parameters in the  $\mathbf{b}$  vector and the  $\mathbf{B}$  matrix. Note that equations (10.72) are linear in the unknown parameters. Note also that the symmetry restrictions (10.58) can be imposed in (10.72) or their validity can be tested.

Once estimates for  $\mathbf{b}$  and  $\mathbf{B}$  have been obtained (denote these estimates by  $\mathbf{b}^*$  and  $\mathbf{B}^*$  respectively), then equations (10.72) can be used in order to generate a period  $t$  vector of fitted input demands,  $\mathbf{x}^{t*}$  say:

$$\mathbf{x}^{t*} \equiv y^t [\mathbf{b}^* + \mathbf{B}^* \mathbf{v}^t - (1/2) \mathbf{v}^{tT} \mathbf{B}^* \mathbf{v}^t \boldsymbol{\alpha}]; \quad t = 1, \dots, T. \quad (10.73)$$

Equations (10.64) and (10.71) may be used in order to calculate the matrix of period  $t$  *estimated input price derivatives*,  $\nabla_{\mathbf{p}} \mathbf{x}(y^t, \mathbf{p}^t) = \nabla_{pp}^2 C(y^t, \mathbf{p}^t)$ . Our estimate for  $\nabla_{pp}^2 C(y^t, \mathbf{p}^t)$  is:

$$[C_{ij}^{t*}] \equiv y^t [(\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1} \mathbf{B}^* - (\boldsymbol{\alpha}^T \mathbf{p}^t)^{-2} \mathbf{B}^* \mathbf{p}^t \boldsymbol{\alpha}^T - (\boldsymbol{\alpha}^T \mathbf{p}^t)^{-2} \boldsymbol{\alpha} \mathbf{p}^{tT} \mathbf{B}^* + (\boldsymbol{\alpha}^T \mathbf{p}^t)^{-3} \mathbf{p}^{tT} \mathbf{B}^* \mathbf{p}^t \boldsymbol{\alpha} \boldsymbol{\alpha}^T]; \quad t = 1, \dots, T. \quad (10.74)$$

Equations (10.73) and (10.74) may be used in order to obtain estimates for the matrix of *period  $t$  input demand price elasticities*,  $[e_{ij}^t]$ :

$$e_{ij}^t \equiv \partial \ln x_i(y^t, \mathbf{p}^t) / \partial \ln p_j = p_j^t C_{ij}^{t*} / x_i^{t*}; \quad i, j = 1, \dots, N; t = 1, \dots, T \quad (10.75)$$

where  $x_i^{t*}$  is the  $i$ th component of the vector of fitted demands  $\mathbf{x}^{t*}$  defined by (10.73).

There is one important additional topic that we have to cover in our discussion of the normalized quadratic functional form: what conditions on  $\mathbf{b}$  and  $\mathbf{B}$  are necessary and sufficient to ensure that  $c(\mathbf{p})$  defined by (10.56)-(10.59) is concave in the components of the price vector  $\mathbf{p}$ ?

The function  $c(\mathbf{p})$  will be concave in  $\mathbf{p}$  if and only if  $\nabla^2 c(\mathbf{p})$  is a negative semidefinite matrix for each  $\mathbf{p}$  in the domain of definition of  $c$ . Evaluating (10.64) at  $\mathbf{p} = \mathbf{p}^*$  and using the restrictions (10.59) yields:

$$\nabla^2 c(\mathbf{p}^*) = (\boldsymbol{\alpha}^T \mathbf{p}^*)^{-1} \mathbf{B}. \quad (10.76)$$

Since  $\boldsymbol{\alpha} > \mathbf{0}_N$  and  $\mathbf{p}^* \gg \mathbf{0}_N$ ,  $\boldsymbol{\alpha}^T \mathbf{p}^* > 0$ . Thus in order for  $c(\mathbf{p})$  to be a concave function of  $\mathbf{p}$ , the following necessary condition must be satisfied:

$$\mathbf{B} \text{ is a negative semidefinite matrix.} \quad (10.77)$$

We now show that the *necessary condition* (10.77) is also *sufficient* to imply that  $c(\mathbf{p})$  is concave over the set of  $\mathbf{p}$  such that  $\mathbf{p} \gg \mathbf{0}_N$ . Unfortunately, the proof is somewhat involved.\*<sup>19</sup>

\*<sup>19</sup> The method of proof is due to Diewert and Wales (1987)[153].

Let  $\mathbf{p} \gg \mathbf{0}_N$ . We assume that  $\mathbf{B}$  is negative semidefinite and we want to show that  $\nabla^2 c(\mathbf{p})$  is negative semidefinite or equivalently, that  $-\nabla^2 c(\mathbf{p})$  is positive semidefinite. Thus for any vector  $\mathbf{z}$ , we want to show that  $-\mathbf{z}^T \nabla^2 c(\mathbf{p}) \mathbf{z} \geq 0$ . Using (10.64), this inequality is equivalent to:

$$-(\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{z}^T \mathbf{B} \mathbf{z} + (\boldsymbol{\alpha}^T \mathbf{p})^{-2} \mathbf{z}^T \mathbf{B} \mathbf{p} \boldsymbol{\alpha}^T \mathbf{z} + (\boldsymbol{\alpha}^T \mathbf{p})^{-2} \mathbf{z}^T \boldsymbol{\alpha} \mathbf{p}^T \mathbf{B} \mathbf{z} - (\boldsymbol{\alpha}^T \mathbf{p})^{-3} \mathbf{p}^T \mathbf{B} \mathbf{p} \mathbf{z}^T \boldsymbol{\alpha} \boldsymbol{\alpha}^T \mathbf{z} \geq 0 \quad \text{or} \quad (10.78)$$

$$-(\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{z}^T \mathbf{B} \mathbf{z} - (\boldsymbol{\alpha}^T \mathbf{p})^{-3} \mathbf{p}^T \mathbf{B} \mathbf{p} (\boldsymbol{\alpha}^T \mathbf{z})^2 \geq -2(\boldsymbol{\alpha}^T \mathbf{p})^{-2} \mathbf{z}^T \mathbf{B} \mathbf{p} \boldsymbol{\alpha}^T \mathbf{z} \quad \text{using } \mathbf{B} = \mathbf{B}^T. \quad (10.79)$$

Define  $\mathbf{A} \equiv -\mathbf{B}$ . Since  $\mathbf{B}$  is symmetric and negative semidefinite by assumption,  $\mathbf{A}$  is symmetric and positive semidefinite. Thus there exists an orthonormal matrix  $\mathbf{U}$  such that

$$\mathbf{U}^T \mathbf{A} \mathbf{U} = \boldsymbol{\Lambda}; \quad (10.80)$$

$$\mathbf{U}^T \mathbf{U} = \mathbf{I}_N \quad (10.81)$$

where  $\mathbf{I}_N$  is the  $N \times N$  identity matrix and  $\boldsymbol{\Lambda}$  is a diagonal matrix with the nonnegative eigenvalues of  $\mathbf{A}$ ,  $\lambda_i, i = 1, \dots, N$ , running down the main diagonal. Now premultiply both sides of (10.80) by  $\mathbf{U}$  and postmultiply both sides by  $\mathbf{U}^T$ . Using (10.81),  $\mathbf{U}^T = \mathbf{U}^{-1}$ , and the transformed equation (10.80) becomes the following equation:

$$\begin{aligned} \mathbf{A} &= \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T \\ &= \mathbf{U} \boldsymbol{\Lambda}^{1/2} \boldsymbol{\Lambda}^{1/2} \mathbf{U}^T \\ &= \mathbf{U} \boldsymbol{\Lambda}^{1/2} \mathbf{U}^T \mathbf{U} \boldsymbol{\Lambda}^{1/2} \mathbf{U}^T \quad \text{since } \mathbf{U}^T \mathbf{U} = \mathbf{I}_N \\ &= \mathbf{S} \mathbf{S} \end{aligned} \quad (10.82)$$

where  $\boldsymbol{\Lambda}^{1/2}$  is the diagonal matrix that has the nonnegative square roots  $\lambda_i^{1/2}$  of the eigenvalues of  $\mathbf{A}$  running down the main diagonal and the symmetric square root of  $\mathbf{A}$  matrix  $\mathbf{S}$  is defined as

$$\mathbf{S} \equiv \mathbf{U} \boldsymbol{\Lambda}^{1/2} \mathbf{U}^T. \quad (10.83)$$

If we replace  $-\mathbf{B}$  in (10.79) with  $\mathbf{A}$ , the inequality that we want to establish becomes

$$2(\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{z}^T \mathbf{A} \mathbf{p} \boldsymbol{\alpha}^T \mathbf{z} \leq \mathbf{z}^T \mathbf{A} \mathbf{z} + (\boldsymbol{\alpha}^T \mathbf{p})^{-2} \mathbf{p}^T \mathbf{A} \mathbf{p} (\boldsymbol{\alpha}^T \mathbf{z})^2 \quad (10.84)$$

where we have also multiplied both sides of (10.79) by the positive number  $\boldsymbol{\alpha}^T \mathbf{p}$  in order to derive (10.84) from (10.79).

Recall the Cauchy-Schwarz inequality for two vectors,  $\mathbf{x}$  and  $\mathbf{y}$ :

$$\mathbf{x}^T \mathbf{y} \leq (\mathbf{x}^T \mathbf{x})^{1/2} (\mathbf{y}^T \mathbf{y})^{1/2}. \quad (10.85)$$

Now we are ready to establish the inequality (10.84). Using (10.82), we have:

$$\begin{aligned} (\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{z}^T \mathbf{A} \mathbf{p} \boldsymbol{\alpha}^T \mathbf{z} &= (\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{z}^T \mathbf{S} \mathbf{S} \mathbf{p} \boldsymbol{\alpha}^T \mathbf{z} \\ &\leq (\mathbf{z}^T \mathbf{S} \mathbf{S}^T \mathbf{z})^{1/2} ([\boldsymbol{\alpha}^T \mathbf{p}]^{-2} [\boldsymbol{\alpha}^T \mathbf{z}]^2 \mathbf{p}^T \mathbf{S}^T \mathbf{S} \mathbf{p})^{1/2} \\ &\quad \text{using (10.85) with } \mathbf{x}^T \equiv \mathbf{z}^T \mathbf{S} \text{ and } \mathbf{y} \equiv (\boldsymbol{\alpha}^T \mathbf{p})^{-1} (\boldsymbol{\alpha}^T \mathbf{z}) \mathbf{S} \mathbf{p} \\ &= (\mathbf{z}^T \mathbf{S} \mathbf{S} \mathbf{z})^{1/2} ([\boldsymbol{\alpha}^T \mathbf{p}]^{-2} [\boldsymbol{\alpha}^T \mathbf{z}]^2 \mathbf{p}^T \mathbf{S} \mathbf{S} \mathbf{p})^{1/2} \quad \text{using } \mathbf{S} = \mathbf{S}^T \\ &= (\mathbf{z}^T \mathbf{A} \mathbf{z})^{1/2} ([\boldsymbol{\alpha}^T \mathbf{p}]^{-2} [\boldsymbol{\alpha}^T \mathbf{z}]^2 \mathbf{p}^T \mathbf{A} \mathbf{p})^{1/2} \quad \text{using (10.82), } \mathbf{A} = \mathbf{S} \mathbf{S} \\ &\leq (1/2)(\mathbf{z}^T \mathbf{A} \mathbf{z}) + (1/2)[\boldsymbol{\alpha}^T \mathbf{p}]^{-2} [\boldsymbol{\alpha}^T \mathbf{z}]^2 (\mathbf{p}^T \mathbf{A} \mathbf{p}) \quad (10.86) \\ &\quad \text{using the nonnegativity of } \mathbf{z}^T \mathbf{A} \mathbf{z}, \mathbf{p}^T \mathbf{A} \mathbf{p}, \text{ the positivity of } \boldsymbol{\alpha}^T \mathbf{z} \\ &\quad \text{and the Theorem of the Arithmetic and Geometric Mean.} \end{aligned}$$

The inequality (10.86) is equivalent to the desired inequality (10.84).

Thus the normalized quadratic unit cost function defined by (10.56)-(10.59) will be concave over the set of positive prices if and only if the symmetric matrix  $\mathbf{B}$  is negative semidefinite. Thus after econometric estimates of the elements of  $\mathbf{B}$  have been obtained using the system of estimating equations (10.72), we need only check that the resulting estimated  $\mathbf{B}$  matrix is negative semidefinite. However, suppose that the estimated  $\mathbf{B}$  matrix is *not* negative semidefinite. How can one reestimate the model, impose negative semidefiniteness on  $\mathbf{B}$ , but without destroying the flexibility of the normalized quadratic functional form?

The desired imposition of negative semidefiniteness can be accomplished using a technique due to Wiley, Schmidt and Bramble (1973)[402]: simply replace the matrix  $\mathbf{B}$  by

$$\mathbf{B} \equiv -\mathbf{A}\mathbf{A}^T \quad (10.87)$$

where  $\mathbf{A}$  is an  $N \times N$  lower triangular matrix; i.e.,  $a_{ij} = 0$  if  $i < j$ .<sup>\*20</sup>

We also need to take into account the restrictions (10.59),  $\mathbf{B}\mathbf{p}^* = \mathbf{0}_N$ . These restrictions on  $\mathbf{B}$  can be imposed if we impose the following restrictions on  $\mathbf{A}$ :

$$\mathbf{A}^T \mathbf{p}^* = \mathbf{0}_N. \quad (10.88)$$

To show how this curvature imposition technique works, let  $\mathbf{p}^* = \mathbf{1}_N$  and consider the case  $N = 2$ . In this case, we have:

$$\mathbf{A} \equiv \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} \text{ and } \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} \\ 0 & a_{22} \end{bmatrix}.$$

The restrictions (10.88) become:

$$\mathbf{A}^T \mathbf{1}_2 = \begin{bmatrix} a_{11} + a_{21} \\ a_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and hence we must have  $a_{21} = -a_{11}$  and  $a_{22} = 0$ . Thus in this case,

$$\mathbf{B} \equiv -\mathbf{A}\mathbf{A}^T = - \begin{bmatrix} a_{11} & 0 \\ -a_{11} & 0 \end{bmatrix} \begin{bmatrix} a_{11} & -a_{11} \\ 0 & 0 \end{bmatrix} = - \begin{bmatrix} a_{11}^2 & -a_{11}^2 \\ -a_{11}^2 & a_{11}^2 \end{bmatrix} = a_{11}^2 \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}. \quad (10.89)$$

Equations (10.89) show how the elements of the  $\mathbf{B}$  matrix can be defined in terms of the single parameter,  $a_{11}^2$ . Note that with this reparameterization of the  $\mathbf{B}$  matrix, it will be necessary to use nonlinear regression techniques rather than modifications of linear regression techniques. This turns out to be the cost of imposing the correct curvature conditions on the unit cost function.

In the following section, we indicate how the functional forms described in sections 10.2-10.4 above can be adapted to estimate consumer preferences.

## 10.6 The Estimation of Consumer Preferences: The General Framework

The cost function and production function framework described in the previous sections can be readily adapted to the problem of estimating consumer preferences: simply replace output  $y$  by utility  $u$ , reinterpret the production function  $f$  as a utility function, reinterpret the input vector  $\mathbf{x}$  as a vector of commodity demands and reinterpret the vector of input prices  $\mathbf{p}$  as a vector of

<sup>\*20</sup> Since  $\mathbf{z}^T \mathbf{A}\mathbf{A}^T \mathbf{z} = (\mathbf{A}^T \mathbf{z})^T (\mathbf{A}^T \mathbf{z}) = \mathbf{y}^T \mathbf{y} \geq 0$  for all vectors  $\mathbf{z}$ ,  $\mathbf{A}\mathbf{A}^T$  is positive semidefinite and hence  $-\mathbf{A}\mathbf{A}^T$  is negative semidefinite. Diewert and Wales (1987; 53)[153] show that any positive semidefinite matrix can be written as  $\mathbf{A}\mathbf{A}^T$  where  $\mathbf{A}$  is lower triangular. Hence, it is not restrictive to reparameterize an arbitrary negative semidefinite matrix  $\mathbf{B}$  as  $-\mathbf{A}\mathbf{A}^T$ .

commodity prices. With these changes, the producer's cost minimization problem (10.12) becomes the following problem of *minimizing the cost or expenditure of attaining a given level of utility*  $u$ :

$$C(u, \mathbf{p}) \equiv \min_{\mathbf{x}} \{\mathbf{p}^T \mathbf{x} : f(\mathbf{x}) \geq u\}. \quad (10.90)$$

If the cost function is differentiable with respect to the components of the commodity price vector  $\mathbf{p}$ , then Shephard's (1953; 11)[355] Lemma applies and the consumer's system of commodity demand functions as functions of the chosen utility level  $u$  and the commodity price vector  $\mathbf{p}$ ,  $\mathbf{x}(u, \mathbf{p})$ , is equal to the vector of first order partial derivatives of the cost or expenditure function  $C(u, \mathbf{p})$  with respect to the components of  $\mathbf{p}$ :

$$\mathbf{x}(u, \mathbf{p}) = \nabla_{\mathbf{p}} C(u, \mathbf{p}). \quad (10.91)$$

The system of demand functions  $\mathbf{x}(u, \mathbf{p})$  defined in (10.91) are known as *Hicksian*<sup>\*21</sup> *demand functions*.

Thus it seems that we can adapt the theory of cost and production functions used in sections 10.2-10.4 above in a very straightforward way and estimate consumer preferences in exactly the same way that we estimated production functions or their dual cost functions. Thus we need only replace period  $t$  output,  $y^t$ , by period  $t$  utility,  $u^t$ , in the estimating equations (10.21) (for the generalized Leontief cost function) and (10.72) (for the normalized quadratic cost function) and reinterpret the resulting equations. However, there is a problem: the period  $t$  output level  $y^t$  is an *observable* variable but the period  $t$  utility level  $u^t$  is *not observable*!

However, this problem can be solved. We need only equate the cost function  $C(u, \mathbf{p})$  to the consumer's *observable expenditure* in the period under consideration,  $Y$  say, and solve the resulting equation for  $u$  as a function of  $Y$  and  $\mathbf{p}$ , say  $u = g(Y, \mathbf{p})$ . Thus  $u = g(Y, \mathbf{p})$  is the solution to:

$$C(u, \mathbf{p}) = Y \quad (10.92)$$

and the resulting solution function  $g(Y, \mathbf{p})$  is the *consumer's indirect utility function*. Now replace the  $u$  in the system of Hicksian demand functions (10.91) by  $g(Y, \mathbf{p})$  and we obtain the consumer's system of (observable) market demand functions:

$$\mathbf{x} = \nabla_{\mathbf{p}} C(g(Y, \mathbf{p}), \mathbf{p}). \quad (10.93)$$

We will illustrate how this general framework can be implemented in the context of several specific functional forms for the cost function.

## 10.7 The Generalized Leontief Cost Function for Homothetic Preferences

We illustrate the above procedure for the generalized Leontief cost function defined in section 10.2 above. For this functional form, equation (10.92) becomes:

$$u \sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{1/2} p_j^{1/2} = Y; \quad (b_{ij} = b_{ji} \text{ for all } i \text{ and } j) \quad (10.94)$$

and the  $u$  solution to this equation is:

$$u = g(Y, \mathbf{p}) = Y \left/ \left[ \sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{1/2} p_j^{1/2} \right] \right. . \quad (10.95)$$

Substituting (10.95) into (10.91) leads to the following system of market demand functions:

$$x_i = \left[ \sum_{j=1}^N b_{ij} (p_j/p_i)^{1/2} \right] Y \left/ \left[ \sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{1/2} p_j^{1/2} \right] \right.; \quad i = 1, \dots, N. \quad (10.96)$$

<sup>\*21</sup> See Hicks (1946; 311-331)[222].

Evaluating (10.96) at the period  $t$  data and adding a stochastic error term  $e_i^t$  to equation  $i$  in (10.96) for  $i = 1, \dots, N$  leads to the following system of estimating equations:<sup>\*22</sup>

$$x_i^t = \left[ \sum_{j=1}^N b_{ij} (p_j^t / p_i^t)^{1/2} \right] Y^t / \left[ \sum_{i=1}^N \sum_{j=1}^N b_{ij} (p_i^t)^{1/2} (p_j^t)^{1/2} \right] + e_i^t; \quad t = 1, \dots, T; i = 1, \dots, N. \quad (10.97)$$

## 10.8 The Normalized Quadratic Cost Function for Homothetic Preferences

We can also illustrate the above procedure for the normalized quadratic cost function defined in section 10.4 above. For this functional form, equation (10.92) becomes:

$$u[\mathbf{b}^T \mathbf{p} + (1/2)(\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{p}^T \mathbf{B} \mathbf{p}] = Y \quad (10.98)$$

and the  $u$  solution to this equation is:

$$u = g(Y, \mathbf{p}) = Y / [\mathbf{b}^T \mathbf{p} + (1/2)(\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{p}^T \mathbf{B} \mathbf{p}]. \quad (10.99)$$

Substituting (10.99) into (10.91) leads to the following system of market demand functions:

$$\mathbf{x} = [\mathbf{b} + \mathbf{B} \mathbf{v} - (1/2) \mathbf{v}^T \mathbf{B} \mathbf{v} \boldsymbol{\alpha}] [(\boldsymbol{\alpha}^T \mathbf{p})^{-1} Y] / [\mathbf{b}^T \mathbf{v} + (1/2) \mathbf{v}^T \mathbf{B} \mathbf{v}] \quad (10.100)$$

where  $\mathbf{v} \equiv (\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{p} = \mathbf{p} / \boldsymbol{\alpha}^T \mathbf{p}$  is the vector of normalized prices. Evaluating (10.100) at the period  $t$  data and adding a vector of stochastic error terms  $\mathbf{e}^t$  to the system of equations (10.100) leads to the following system of estimating equations:

$$\mathbf{x}^t = [\mathbf{b} + \mathbf{B} \mathbf{v}^t - (1/2) \mathbf{v}^{tT} \mathbf{B} \mathbf{v}^t \boldsymbol{\alpha}] [(\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1} Y^t] / [\mathbf{b}^T \mathbf{v}^t + (1/2) \mathbf{v}^{tT} \mathbf{B} \mathbf{v}^t] + \mathbf{e}^t; \quad t = 1, \dots, T \quad (10.101)$$

where  $\mathbf{v}^t \equiv \mathbf{p}^t / \boldsymbol{\alpha}^T \mathbf{p}^t$  for  $t = 1, \dots, T$ .

In practice, period  $t$  "income"  $Y^t$  is defined to be period  $t$  expenditure,  $\mathbf{p}^{tT} \mathbf{x}^t = \sum_{i=1}^N p_i^t x_i^t$ ; i.e., we have:

$$Y^t = \mathbf{p}^{tT} \mathbf{x}^t = \sum_{i=1}^N p_i^t x_i^t; \quad t = 1, \dots, T. \quad (10.102)$$

However, the identities (10.102) create some econometric difficulties: namely, we cannot assume that all of the error terms,  $e_i^t$ , in each period are independently distributed. Thus if we premultiply both sides of equation  $i$  for period  $t$  in (10.97) by  $p_i^t$  and sum over  $i$ , we obtain the following identity using (10.102):

$$Y^t = Y^t + \sum_{i=1}^N p_i^t e_i^t; \quad t = 1, \dots, T \quad (10.103)$$

which in turn implies that the period  $t$  error terms  $e_i^t$  satisfy the following exact identity:

$$\sum_{i=1}^N p_i^t e_i^t = 0; \quad t = 1, \dots, T. \quad (10.104)$$

<sup>\*22</sup> Since  $Y^t$  will typically equal  $\sum_{i=1}^N p_i^t x_i^t$ , it can be verified that the errors in (10.97) for any period  $t$  cannot be independently distributed since they must satisfy the restriction  $\sum_{i=1}^N p_i^t e_i^t = 0$  for each  $t$ ; see (10.104) below. It is also necessary to impose a normalization on the  $b_{ij}$  since the right hand side of each equation in (10.97) is homogeneous of degree 0 in the  $b_{ij}$ . We will deal with the normalization problem in section 10.9 below.

In a similar fashion, premultiply both sides of the period  $t$  equation in (10.101) by  $\mathbf{p}^{tT}$ , we obtain the following equations:

$$\begin{aligned} \mathbf{p}^{tT} \mathbf{x}^t &= \mathbf{p}^{tT} [\mathbf{b} + \mathbf{B}\mathbf{v}^t - (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t \boldsymbol{\alpha}] [(\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1} Y^t] / [\mathbf{b}^T \mathbf{v}^t + (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t] + \mathbf{p}^{tT} \mathbf{e}^t; \\ & \qquad \qquad \qquad t = 1, \dots, T \text{ or} \\ Y^t &= \mathbf{p}^{tT} \boldsymbol{\alpha}^T \mathbf{p}^t (\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1} [\mathbf{b} + \mathbf{B}\mathbf{v}^t - (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t \boldsymbol{\alpha}] [(\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1} Y^t] / [\mathbf{b}^T \mathbf{v}^t + (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t] + \mathbf{p}^{tT} \mathbf{e}^t \text{ or} \\ Y^t &= \mathbf{v}^{tT} \boldsymbol{\alpha}^T \mathbf{p}^t [\mathbf{b} + \mathbf{B}\mathbf{v}^t - (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t \boldsymbol{\alpha}] [(\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1} Y^t] / [\mathbf{b}^T \mathbf{v}^t + (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t] + \mathbf{p}^{tT} \mathbf{e}^t \text{ or} \\ Y^t &= \mathbf{v}^{tT} [\mathbf{b} + \mathbf{B}\mathbf{v}^t - (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t \boldsymbol{\alpha}] [Y^t] / [\mathbf{b}^T \mathbf{v}^t + (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t] + \mathbf{p}^{tT} \mathbf{e}^t \text{ or} \\ Y^t &= [\mathbf{b}^T \mathbf{v}^t + (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t] [Y^t] / [\mathbf{b}^T \mathbf{v}^t + (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t] + \mathbf{p}^{tT} \mathbf{e}^t \text{ or} \\ Y^t &= Y^t + \mathbf{p}^{tT} \mathbf{e}^t \end{aligned} \tag{10.105}$$

which in turn implies that the period  $t$  error term vector  $\mathbf{e}^t$  satisfies the following exact identity,  $\mathbf{p}^{tT} \mathbf{e}^t = 0$  for  $t = 1, \dots, T$ , which is the same identity as (10.104).

Thus for both the generalized Leontief and the normalized quadratic cost function models the period  $t$  error vectors satisfy an exact identity and hence in both models, we must drop one estimating equation; i.e., we must drop one of the estimating equations in (10.97) and one of the estimating equations in (10.101). Thus there are some differences between the cost function models in the producer context and in the consumer context.

## 10.9 The Problem of Cardinalizing Utility

There is another significant difference between the producer models discussed in the previous sections and the consumer models discussed in the present section. Look closely at the estimating equations (10.97) and (10.101). From (10.97), it can be seen that the right hand side explanatory variables are *homogeneous of degree 0* in the  $b_{ij}$  coefficients. Thus the regression will not be able to determine the *scale* of the  $b_{ij}$  parameters. Similarly, by looking at the right hand side of (10.101), it can be seen that the right hand side explanatory variables are *homogeneous of degree 0* in the components of the  $\mathbf{b}$  vector and the  $\mathbf{B}$  matrix. Thus the regression will not be able to determine the *scale* of the parameters in  $\mathbf{b}$  and  $\mathbf{B}$ . This indeterminacy means that we require at least one additional restriction or normalization on the parameters of each of these models. Basically, what we have to do is *cardinalize* our measure of utility in some way.

There are two simple ways of cardinalizing utility<sup>\*23</sup>:

- Pick a positive reference quantity vector  $\mathbf{x}^* \gg \mathbf{0}_N$ . Let the period  $t$  consumption vector  $\mathbf{x}^t$  be on the indifference surface  $I(\mathbf{x}^t) \equiv \{\mathbf{x} : f(\mathbf{x}) = f(\mathbf{x}^t)\}$ . Let  $\lambda^t \mathbf{x}^*$  be on the  $I(\mathbf{x}^t)$  indifference curve. Then measure period  $t$  utility as  $\lambda^t$ .
- Pick a positive reference price vector  $\mathbf{p}^* \gg \mathbf{0}_N$ . Then normalize the consumer's cost function  $C(u, \mathbf{p})$  so that it has the following property:

$$C(u, \mathbf{p}^*) = u \text{ for all } u > 0. \tag{10.106}$$

The meaning of (10.106) is that if the consumer faces the reference price vector  $\mathbf{p}^*$ , then his or her utility will be equal to his or her "income" or expenditure on commodities at those reference prices. Thus if relative prices never changed, the consumer's utility is proportional to the size of the observed budget set. This serves to cardinalize utility for all consumption vectors. Samuelson (1974)[347] called this type of cardinalization of utility, *money metric utility*.<sup>\*24</sup>

<sup>\*23</sup> The two methods are equivalent in the case of homothetic preferences.

<sup>\*24</sup> The basic idea can be traced back to Hicks (1941-42)[220].

We will follow the money metric method of scaling utility. For the generalized Leontief model, the restriction (10.106) implies the following normalization of the  $b_{ij}$ :

$$\sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{*1/2} p_j^{*1/2} = 1. \quad (10.107)$$

For the normalized quadratic model, the restriction (10.106) implies the following normalization of the components of the  $\mathbf{b}$  vector and the  $\mathbf{B}$  matrix:

$$\mathbf{b}^T \mathbf{p}^* + (1/2) \mathbf{p}^{*T} \mathbf{B} \mathbf{p}^* / \alpha^T \mathbf{p}^* = 1. \quad (10.108)$$

If we choose the reference vector  $\mathbf{p}^*$  in (10.106) to be the same as the reference vector  $\mathbf{p}^*$  which occurred in (10.59), then  $\mathbf{B} \mathbf{p}^* = \mathbf{0}_N$  and the cardinalization restriction (10.108) becomes:

$$\mathbf{b}^T \mathbf{p}^* = 1. \quad (10.109)$$

**Problem 8** Adapt the translog unit cost function model presented in section 10.3 above to the consumer context.

*Hint:* Equations (10.42) do not depend on utility! However, you need to choose  $\mathbf{p}^*$  in a specific way in order to impose money metric utility scaling.

**Problem 9** Suppose the consumer's cost function has the following form:

$$C(u, \mathbf{p}) = uc(\mathbf{p}) \quad (i)$$

where  $c(\mathbf{p})$  is a well behaved unit cost function. Assuming that  $c(\mathbf{p})$  is differentiable, show that the consumer's system of market demand functions has the following form:

$$\mathbf{x}(Y, \mathbf{p}) = Y \nabla_{\mathbf{p}} c(\mathbf{p}) / c(\mathbf{p}). \quad (ii)$$

Show that  $\partial \ln x_i(Y, \mathbf{p}) / \partial \ln Y = 1$  for  $i = 1, \dots, N$ ; i.e., if the consumer has preferences given by (i), then all income elasticities of demand are one! This contradicts Engel's Law; i.e., that the income elasticity of demand for food is less than one.

## 10.10 Modeling Nonhomothetic Preferences

Problem 9 above shows that the assumption that the utility function is linearly homogeneous (the homothetic preferences assumption) is not a good assumption from the empirical point of view. Hence we need to generalize our functional forms in order to accommodate nonhomothetic preferences.

Let  $C^*(u, \mathbf{p})$  be an arbitrary twice continuously differentiable cost function that satisfies money metric scaling at the positive reference price vector  $\mathbf{p}^* \gg \mathbf{0}_N$ ; i.e.,  $C^*$  satisfies:

$$C^*(u, \mathbf{p}^*) = u \text{ for all } u > 0. \quad (10.110)$$

Let  $c(\mathbf{p})$  be a flexible unit cost function. Then Diewert (1980; 597)[88] showed that the following functional form could approximate  $C^*$  to the second order at  $(u^*, \mathbf{p}^*)$  where  $u^* > 0$ :

$$C(u, \mathbf{p}) \equiv \mathbf{a}^T \mathbf{p} + uc(\mathbf{p}) \quad (10.111)$$

where the vector of parameters  $\mathbf{a}$  can be chosen to satisfy the following restriction:

$$\mathbf{a}^T \mathbf{p}^* = 0. \quad (10.112)$$

The parameters of the unit cost function also satisfy the following restriction:

$$c(\mathbf{p}^*) = 1. \quad (10.113)$$

In order to derive the system of market demand functions that corresponds to the cost function defined by (10.111), we again set  $C(u, \mathbf{p})$  equal to “income”  $Y$  and solve for the  $u = g(Y, \mathbf{p})$  solution:

$$u = [Y - \mathbf{a}^T \mathbf{p}] / c(\mathbf{p}). \quad (10.114)$$

The *system of Hicksian demand functions* that corresponds to the cost function defined by (10.111) is as usual obtained using Shephard’s Lemma:

$$\mathbf{x}(u, \mathbf{p}) \equiv \nabla_{\mathbf{p}} C(u, \mathbf{p}) = \mathbf{a} + u \nabla_{\mathbf{p}} c(\mathbf{p}). \quad (10.115)$$

Now replace  $u$  in the right hand side of (10.115) by the right hand side of (10.114) and we obtain the *consumer’s system of market demand functions*:

$$\mathbf{x}(Y, \mathbf{p}) = \mathbf{a} + \nabla_{\mathbf{p}} c(\mathbf{p}) [Y - \mathbf{a}^T \mathbf{p}] / c(\mathbf{p}). \quad (10.116)$$

Letting  $c(\mathbf{p}) \equiv \sum_{i=1}^N \sum_{j=1}^N b_{ij} p_i^{1/2} p_j^{1/2}$  be the generalized Leontief unit cost function, the system of market demand functions (10.116) becomes, after adding stochastic error terms:

$$x_i^t = a_i + \left\{ \left[ \sum_{j=1}^N b_{ij} (p_j^t / p_i^t)^{1/2} \right] \left[ Y^t - \sum_{k=1}^N a_k p_k^t \right] / \left[ \sum_{i=1}^N \sum_{j=1}^N b_{ij} (p_i^t)^{1/2} (p_j^t)^{1/2} \right] \right\} + e_i^t; \quad t = 1, \dots, T; i = 1, \dots, N. \quad (10.117)$$

One of the  $a_i$  needs to be eliminated from the estimating equations (10.117) using the restriction  $\mathbf{a}^T \mathbf{p}^* = 0$  and one of the  $b_{ij}$  needs to be eliminated using the restriction  $c(\mathbf{p}^*) = 1$  in order to obtain the final system of estimating equations. However, if period  $t$  “income”  $Y^t$  is equal to period  $t$  expenditure on the commodities,  $\mathbf{p}^{tT} \mathbf{x}^t$ , then as before, we can only use  $N - 1$  of the  $N$  equations in (10.117) as estimating equations.

Letting  $c(\mathbf{p}) \equiv \mathbf{b}^T \mathbf{p} + (1/2)(\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{p}^T \mathbf{B} \mathbf{p}$  be the normalized quadratic unit cost function (with  $\mathbf{b}^T \mathbf{p}^* = 1$  and  $\mathbf{B} \mathbf{p}^* = \mathbf{0}_N$ ), the system of market demand functions (10.116) becomes, after adding stochastic error terms:

$$\mathbf{x}^t = \mathbf{a} + \{ [\mathbf{b} + \mathbf{B} \mathbf{v}^t - (1/2) \mathbf{v}^{tT} \mathbf{B} \mathbf{v}^t \boldsymbol{\alpha}] [(\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1}] [Y^t - \mathbf{a}^T \mathbf{p}^t] / [\mathbf{b}^T \mathbf{v}^t + (1/2) \mathbf{v}^{tT} \mathbf{B} \mathbf{v}^t] \} + \mathbf{e}^t; \quad t = 1, \dots, T \quad (10.118)$$

where  $\mathbf{v}^t \equiv \mathbf{p}^t / \boldsymbol{\alpha}^T \mathbf{p}^t$  for  $t = 1, \dots, T$ . Obviously, nonlinear regression techniques have to be used in order to estimate the unknown parameters in the system of estimating equations (10.118). One of the  $a_i$  needs to be eliminated from the estimating equations (10.118) using the restriction  $\mathbf{a}^T \mathbf{p}^* = 0$  and one of the  $b_i$  needs to be eliminated using the restriction  $\mathbf{b}^T \mathbf{p}^* = 1$  in order to obtain the final system of estimating equations. However, if period  $t$  “income”  $Y^t$  is equal to period  $t$  expenditure on the commodities,  $\mathbf{p}^{tT} \mathbf{x}^t$ , then as before, we can only use  $N - 1$  of the  $N$  equations in (10.118) as estimating equations. If the estimated  $\mathbf{B}$  matrix turns out to be *not* negative semidefinite, then we need to replace  $\mathbf{B}$  by  $-\mathbf{A} \mathbf{A}^T$  where  $\mathbf{A}$  is a lower triangular matrix satisfying  $\mathbf{A} \mathbf{p}^* = \mathbf{0}_N$ .

## 10.11 The Use of Linear Spline Functions to Achieve Greater Flexibility

Although the above model is flexible around the point  $\mathbf{p}^*$ , as we move away from  $\mathbf{p}^*$ , the model (10.118) may not fit the data very well. If the plots of the actual and fitted values using the

normalized quadratic model defined by the estimating equations (10.118) have a zig-zag appearance, then it may be worthwhile to try a *linear spline model*. We will indicate below how a two segment linear spline model can be implemented. For more details (and an extension to 3 segments instead of 2), see Diewert and Wales (1993; 81-85)[157].

We redefine the normalized quadratic cost function  $C(u, \mathbf{p})$  as follows:

$$C(u, \mathbf{p}) = \mathbf{a}^T \mathbf{p} + u(1/2)(\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{p}^T \mathbf{B} \mathbf{p} + d(u, \mathbf{p}) \quad (10.119)$$

where  $\mathbf{a}$  satisfies  $\mathbf{a}^T \mathbf{p}^* = 0$  and  $\boldsymbol{\alpha}$  and  $\mathbf{B}$  satisfy the restrictions (10.57)-(10.59). The function  $d(u, \mathbf{p})$  is defined as follows:

$$d(u, \mathbf{p}) \equiv \begin{cases} u \mathbf{b}^T \mathbf{p} & \text{for } 0 \leq u \leq u^* \\ u^* \mathbf{b}^T \mathbf{p} + (u - u^*) \mathbf{f}^T \mathbf{p} & \text{for } u^* \leq u. \end{cases} \quad (10.120)$$

where  $\mathbf{b}^T \equiv [b_1, \dots, b_N]$  and  $\mathbf{f}^T \equiv [f_1, \dots, f_N]$  parameter vectors to be estimated and  $u^*$  is a *break point level of utility* to be chosen by the investigator. The vectors  $\mathbf{b}$  and  $\mathbf{f}$  satisfy the restrictions:

$$\mathbf{b}^T \mathbf{p}^* = 1; \quad \mathbf{f}^T \mathbf{p}^* = 1. \quad (10.121)$$

How should one pick the break point  $u^*$ ? We examine the plots of the regression model defined by (10.118) and look for an observation number where the plot changes from a zig to a zag. Suppose that this observation number is  $t^*$ . Now compute index numbers of utility using the price and quantity data and determine what level of utility corresponds to the chosen observation and set this level equal to  $u^*$ . This choice of  $u^*$  will work satisfactorily if the observations which precede the chosen observation have estimated indirect utilities which are equal to or less than  $u^*$  and the remaining observations have indirect utilities that are greater than  $u^*$ .

The estimating equations for the first  $t^*$  observations will still be given by (10.118); i.e., for the first  $t^*$  observations, our estimating equations are:

$$\mathbf{x}^t = \mathbf{a} + \{[\mathbf{b} + \mathbf{B} \mathbf{v}^t - (1/2) \mathbf{v}^{tT} \mathbf{B} \mathbf{v}^t \boldsymbol{\alpha}] [(\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1}] [Y^t - \mathbf{a}^T \mathbf{p}^t] / [\mathbf{b}^T \mathbf{v}^t + (1/2) \mathbf{v}^{tT} \mathbf{B} \mathbf{v}^t]\} + \mathbf{e}^t; \quad t = 1, \dots, t^* \quad (10.122)$$

where as usual,  $\mathbf{v}^t \equiv \mathbf{p}^t / \boldsymbol{\alpha}^T \mathbf{p}^t$ .

In order to obtain the estimating equations for the last  $T - t^*$  observations, we need to form the Hicksian demand functions and calculate the indirect utility function. If  $t > t^*$ , then the Hicksian demand functions that correspond to the functional form defined by (10.119) and (10.120) are:

$$\begin{aligned} \mathbf{x}(u, \mathbf{p}) &\equiv \nabla_{\mathbf{p}} C(u, \mathbf{p}) = \mathbf{a} + u [(\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{B} \mathbf{p} - (1/2) (\boldsymbol{\alpha}^T \mathbf{p})^{-2} \mathbf{p}^T \mathbf{B} \mathbf{p} \boldsymbol{\alpha}] + u^* \mathbf{b} + (u - u^*) \mathbf{f} \\ &= \mathbf{a} + u^* \mathbf{b} - u^* \mathbf{f} + u [\mathbf{f} + (\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{B} \mathbf{p} - (1/2) (\boldsymbol{\alpha}^T \mathbf{p})^{-2} \mathbf{p}^T \mathbf{B} \mathbf{p} \boldsymbol{\alpha}]. \end{aligned} \quad (10.123)$$

For  $t > t^*$ , the indirect utility function  $u = g(Y, \mathbf{p})$  can be obtained by solving  $C(u, \mathbf{p}) = Y$ . The solution is:

$$u = [Y - \mathbf{a}^T \mathbf{p} - u^* \mathbf{b}^T \mathbf{p} + u^* \mathbf{f}^T \mathbf{p}] / [\mathbf{f}^T \mathbf{p} + (1/2) (\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{p}^T \mathbf{B} \mathbf{p}]. \quad (10.124)$$

Now substitute (10.124) into (10.123) in order to obtain the consumer's market demand functions for periods  $t > t^*$ . After adding stochastic error terms, we obtain the following estimating equations:

$$\begin{aligned} \mathbf{x}^t &= \mathbf{a} + u^* \mathbf{b} - u^* \mathbf{f} \\ &+ \{[\mathbf{f} + \mathbf{B} \mathbf{v}^t - (1/2) \mathbf{v}^{tT} \mathbf{B} \mathbf{v}^t \boldsymbol{\alpha}] [(\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1}] [Y^t - \mathbf{a}^T \mathbf{p}^t - u^* \mathbf{b}^T \mathbf{p}^t + u^* \mathbf{f}^T \mathbf{p}^t] / [\mathbf{f}^T \mathbf{v}^t + (1/2) \mathbf{v}^{tT} \mathbf{B} \mathbf{v}^t]\} \\ &+ \mathbf{e}^t \quad \text{for } t^* < t \leq T. \end{aligned} \quad (10.125)$$

Although the estimating equations (10.125) look rather formidable, they can be programmed with a bit of effort. The most difficult part of implementing the above spline model is choosing the “right” observation at which the break point occurs.

As usual, if “income”  $Y^t$  in period  $t$  is equal to expenditure  $\mathbf{p}^{tT} \mathbf{x}^t$ , then we must drop one equation in the system of estimating equations (10.122) and (10.125). Finally, if the estimated  $\mathbf{B}$  matrix is *not* negative semidefinite, then the model should be rerun, setting  $\mathbf{B} = -\mathbf{A}\mathbf{A}^T$ , where  $\mathbf{A}$  is lower triangular and satisfies the restrictions  $\mathbf{A}^T \mathbf{p}^* = \mathbf{0}_N$ .

## 10.12 The Estimation of Unit Profit Functions: The General Framework

Recall problems 1 to 5 above, which introduced the *capital requirements function*,  $F(\mathbf{y})$ , which gives the minimum amount of capital  $k$  that is required to produce the vector of net outputs  $\mathbf{y}$ . The corresponding variable profit (or operating profit) function  $V(k, \mathbf{p})$  can be defined as follows:

$$V(\mathbf{p}, k) \equiv \max_{\mathbf{y}} \{\mathbf{p}^T \mathbf{y} : k = F(\mathbf{y})\} \quad (10.126)$$

If commodity  $i$  is an output, then  $y_i > 0$ ; if commodity  $i$  is an input,  $y_i < 0$ . The available capital is  $k > 0$ .

The function  $V(\mathbf{p}, k)$  must be linearly homogeneous and convex in  $\mathbf{p}$  for fixed  $k$ . The economy’s system of profit maximizing supply and demand functions  $\mathbf{y}(\mathbf{p}, k)$  can be obtained by differentiating  $V(\mathbf{p}, k)$  with respect to the components of  $\mathbf{p}$ : (Hotelling’s (1932)[242] Lemma):

$$\mathbf{y}(\mathbf{p}, k) = \nabla_{\mathbf{p}} V(\mathbf{p}, k). \quad (10.127)$$

The convexity property of  $V$  in  $\mathbf{p}$  implies that:

$$\nabla_{\mathbf{p}} \mathbf{y}(\mathbf{p}, k) = \nabla_{\mathbf{p}\mathbf{p}}^2 V(\mathbf{p}, k) \text{ is a positive semidefinite matrix.} \quad (10.128)$$

If the capital requirements function  $F(\mathbf{y})$  is linearly homogeneous (so that the technology exhibits constant returns to scale), then  $V(\mathbf{p}, k)$  has the following property:

$$V(\mathbf{p}, k) = V(\mathbf{p}, 1)k. \quad (10.129)$$

The unit profit function  $v(\mathbf{p})$  is the gross return to capital we can achieve using one unit of capital; i.e., define  $v$  as:

$$v(\mathbf{p}) \equiv V(\mathbf{p}, 1). \quad (10.130)$$

With a constant returns to scale technology, we have  $V(k, \mathbf{p}) = kv(\mathbf{p})$  so that we need only pick a functional form for the unit profit function  $v$ . It turns out that we can use the functional forms for unit cost functions,  $c(\mathbf{p})$ , that we defined in sections 10.2-10.4 above as functional forms for the unit profit function  $v(\mathbf{p})$ . The only change that we need to make is that the concavity in  $\mathbf{p}$  property for the unit cost function  $c(\mathbf{p})$  must be replaced by a convexity in  $\mathbf{p}$  property for the unit profit function  $v(\mathbf{p})$ . We illustrate the use of some of the unit cost functional forms in the sections below.

## 10.13 The Translog Variable Profit Function with Constant Returns to Scale

The translog unit profit function,  $v(\mathbf{p})$ , is defined as follows:

$$\ln v(\mathbf{p}) \equiv \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln p_i \ln p_j \quad (10.131)$$

where the parameters  $\alpha_i$  and  $\gamma_{ij}$  satisfy the following restrictions:

$$\gamma_{ij} = \gamma_{ji}; \quad 1 \leq i < j \leq N; \quad (N(N-1)/2 \text{ symmetry restrictions}) \quad (10.132)$$

$$\sum_{i=1}^N \alpha_i = 1; \quad (1 \text{ restriction}) \quad (10.133)$$

$$\sum_{j=1}^N \gamma_{ij} = 0; \quad i = 1, \dots, N \quad (N \text{ restrictions}). \quad (10.134)$$

Suppose that in period  $t$ , observed capital input is  $k^t$ , the vector of observed output and variable input prices is  $\mathbf{p}^t \gg \mathbf{0}_N$  and the vector of observed net output supplies  $\mathbf{y}^t > \mathbf{0}_N$ . Thus the *period  $t$  observed variable profit or gross return to capital* is<sup>\*25</sup>:

$$V^t \equiv \mathbf{p}^{tT} \mathbf{y}^t \equiv \sum_{i=1}^N p_i^t y_i^t. \quad (10.135)$$

The log of (10.129) can act as an estimating equation:

$$\ln V^t = \ln k^t + \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i^t + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln p_i^t \ln p_j^t + e_0^t; \quad t = 1, \dots, T \text{ or} \quad (10.136)$$

$$\ln[V^t/k^t] = \alpha_0 + \sum_{i=1}^N \alpha_i \ln p_i^t + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln p_i^t \ln p_j^t + e_0^t; \quad t = 1, \dots, T. \quad (10.137)$$

Note that (10.137) is linear in the unknown parameters. As in section 10.3 above, the old estimating equations (10.42) can be adapted to yield the following estimating equations in the present context:

$$s_i^t \equiv p_i^t y_i^t / V^t = \alpha_i + \sum_{j=1}^N \gamma_{ij} \ln p_j^t + e_i^t; \quad i = 1, \dots, N. \quad (10.138)$$

As in section 10.3, only  $N-1$  of the  $N$  estimating equations in (10.138) are statistically independent.

Unfortunately, the above model is not adequate for empirical applications. The problem is that the economy becomes more efficient over time and more output is produced using the same amount of input; i.e., there is technical progress. Thus we generalize the translog unit profit function defined by (10.131) to include time trends to try and capture the effects of technical progress. Thus we now define the period  $t$  unit profit function  $v(\mathbf{p}, t)$  as follows:

$$\ln v(\mathbf{p}, t) \equiv \alpha_0 + \beta_0 t + \sum_{i=1}^N \alpha_i \ln p_i + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln p_i \ln p_j + \sum_{i=1}^N \beta_i t \ln p_i \quad (10.139)$$

where the parameters  $\alpha_i$  and  $\gamma_{ij}$  satisfy the restrictions (10.132)-(10.134) and the new  $\beta_i$  parameters satisfy the following restriction<sup>\*26</sup>:

$$\sum_{i=1}^N \beta_i = 0. \quad (10.140)$$

Using this new definition for  $v(\mathbf{p}, t)$ , defining  $V(k, \mathbf{p}, t) \equiv kv(\mathbf{p}, t)$  and using the general methodology explained above, our initial estimating equations (10.137) and (10.138) are replaced by the following estimating equations:

$$\ln[V^t/k^t] = \alpha_0 + \beta_0 t + \sum_{i=1}^N \alpha_i \ln p_i^t + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln p_i^t \ln p_j^t + \sum_{i=1}^N \beta_i t \ln p_i^t + e_0^t; \quad (10.141)$$

$$s_i^t \equiv p_i^t y_i^t / V^t = \alpha_i + \sum_{j=1}^N \gamma_{ij} \ln p_j^t + \beta_i t + e_i^t; \quad i = 1, \dots, N; t = 1, \dots, T. \quad (10.142)$$

<sup>\*25</sup> It is important to check that  $V^t > 0$  for each observation  $t$ .

<sup>\*26</sup> This restriction is required in order to ensure that  $v(\mathbf{p})$  is linearly homogeneous in the components of  $\mathbf{p}$ .

As in section 10.3, only  $N - 1$  of the  $N$  equations in (10.142) can be used in the estimation.

We have not substituted the restrictions (10.132)-(10.134) and (10.140) into the estimating equations, (10.141) and (10.142). We now do this for the case  $N = 4$  in order to show how explicit estimating equations can be derived. We use the restriction (10.133),  $\sum_{i=1}^4 \alpha_i = 1$ , in order to eliminate the parameter  $\alpha_4$  and we use the restriction (10.140),  $\sum_{i=1}^4 \beta_i = 0$ , in order to eliminate  $\beta_4$ . Finally, we use the restrictions (10.132) and (10.134) in order to eliminate the parameters  $\gamma_{i4}$  and  $\gamma_{4i}$  for  $i = 1, \dots, 4$ . With these restrictions imposed, the estimating equation (10.141) becomes:

$$\begin{aligned} \ln[V^t/p_4^t k^t] &= \alpha_0 + \beta_0 t + \sum_{i=1}^3 \alpha_i \ln(p_i^t/p_4^t) + \sum_{i=1}^3 \beta_i t \ln(p_i^t/p_4^t) + (1/2) \sum_{i=1}^3 \gamma_{ii} [\ln(p_i^t/p_4^t)]^2 \\ &+ \sum_{i=1}^3 \sum_{j=1; i < j}^3 \gamma_{ij} \ln(p_i^t/p_4^t) \ln(p_j^t/p_4^t) + e_0^t; \quad t = 1, \dots, T. \end{aligned} \quad (10.143)$$

Dropping the last equation in (10.142) and eliminating the  $\gamma_{i4}$  leads to the following system of estimating equations when  $N = 4$ :

$$s_i^t \equiv p_i^t y_i^t / V^t = \alpha_i + \sum_{j=1}^3 \gamma_{ij} \ln(p_j^t/p_4^t) + \beta_i t + e_i^t; \quad i = 1, \dots, 3; t = 1, \dots, T. \quad (10.144)$$

The unknown parameters in (10.143) and (10.144) are  $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \beta_0, \beta_1, \beta_2, \beta_3, \gamma_{11}, \gamma_{22}, \gamma_{33}, \gamma_{12}, \gamma_{13}$  and  $\gamma_{23}$  or 14 parameters in all. Note that *all* of the unknown parameters occur in the estimating equation (10.143). This fact often creates econometric problems. With a great number of parameters in equation (10.143), the fit will tend to be good but due to multicollinearity, the parameters will not be very accurately determined using this equation. However, two stage estimation procedures (or maximum likelihood estimation) will tend to give the first equation undue weight in the system estimation procedure (due to the low variance in the first equation) and hence, very inaccurate estimates of the parameters can result.

We now return to the case of a model with a general  $N$ . Our old formulae (10.47) and (10.49) in section 10.3 above for obtaining elasticities of demand can be adapted in a straightforward manner to give us the following formulae for the elasticities of net supply for variable inputs and outputs. The formulae for the *cross price elasticities of net supply* are given by:

$$\partial \ln y_i(k, \mathbf{p}) / \partial \ln p_j = [s_i(y, \mathbf{p})]^{-1} \gamma_{ij} + s_j(y, \mathbf{p}); \quad i \neq j. \quad (10.145)$$

The formulae for the *own price elasticities of net supply* are given by:

$$\partial \ln y_i(k, \mathbf{p}) / \partial \ln p_i = [s_i(y, \mathbf{p})]^{-1} \gamma_{ii} + s_i(y, \mathbf{p}) - 1; \quad i = 1, \dots, N. \quad (10.146)$$

Thus given econometric estimates for the  $\alpha_i, \beta_i$  and  $\gamma_{ij}$ , which we denote by  $\alpha_i^*, \beta_i^*$  and  $\gamma_{ij}^*$ , the estimated or fitted shares in period  $t$ ,  $s_i^{t*}$  are defined using these estimates and equations (10.142) evaluated at the period  $t$  data:

$$s_i^{t*} \equiv \alpha_i^* + \beta_i^* t + \sum_{j=1}^N \gamma_{ij}^* \ln p_j^t; \quad i = 1, \dots, N; t = 1, \dots, T. \quad (10.147)$$

Now use equations (10.147) and (10.145) evaluated at the period  $t$  data and econometric estimates to obtain the following formula for the *period  $t$  cross elasticities of net supply*,  $e_{ij}^t$ :

$$e_{ij}^t \equiv \partial \ln y_i(k^t, \mathbf{p}^t) / \partial \ln p_j = [s_i^{t*}]^{-1} \gamma_{ij}^* + s_j^{t*}; \quad i \neq j. \quad (10.148)$$

Similarly, use equations (10.146) evaluated at the period  $t$  data and econometric estimates to obtain the following formula for the *period  $t$  own elasticities of net supply*,  $e_{ii}^t$ :

$$e_{ii}^t \equiv \partial \ln y_i(k^t, \mathbf{p}^t) / \partial \ln p_i = [s_i^{t*}]^{-1} \gamma_{ii}^* + s_i^{t*} - 1; \quad i = 1, \dots, N. \quad (10.149)$$

We can also obtain an estimated or *fitted period  $t$  variable profits or gross return to capital*,  $V^{t*}$ , by using our econometric estimates for the parameters and by exponentiating the right hand side of the equation  $t$  in (10.141):

$$V^{t*} \equiv \exp \left[ \ln k^t + \alpha_0^* + t\beta_0^* + \sum_{i=1}^N \alpha_i^* \ln p_i^t + \sum_{i=1}^N \beta_i^* t \ln p_i^t + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij}^* \ln p_i^t \ln p_j^t \right];$$

$$t = 1, \dots, T. \quad (10.150)$$

Finally, our fitted period  $t$  shares  $s_i^{t*}$  defined by (10.147) and our fitted period  $t$  profits  $V^{t*}$  defined by (10.150) can be used in order to obtain estimated or *fitted period  $t$  net supplies*,  $y_i^{t*}$ , as follows:

$$y_i^{t*} \equiv V^{t*} s_i^{t*} / p_i^t; \quad i = 1, \dots, N; t = 1, \dots, T. \quad (10.151)$$

Given the matrix of period  $t$  estimated price elasticities of net supply,  $[e_{ij}^t]$ , we can readily calculate the matrix of period  $t$  *estimated net output price derivatives*,  $\nabla_p \mathbf{y}(k^t, \mathbf{p}^t) = \nabla_{pp}^2 V(k^t, \mathbf{p}^t)$ . Our estimate for element  $ij$  of  $\nabla_{pp}^2 V(k^t, \mathbf{p}^t)$  is:

$$V_{ij}^{t*} \equiv e_{ij}^t y_i^{t*} / p_j^t; \quad i, j = 1, \dots, N; t = 1, \dots, T \quad (10.152)$$

where the estimated period  $t$  elasticities  $e_{ij}^t$  are defined by (10.148) and (10.149) and the fitted period  $t$  net output supplies  $y_i^{t*}$  are defined by (10.151). Once the estimated price derivative matrices  $[V_{ij}^{t*}]$  have been calculated, then we may check whether each of them is positive semidefinite using determinantal conditions or by checking if all of the eigenvalues of each matrix are zero or positive. There remains the problem of measuring the effects of *technical progress*. Using (10.139) in order to define  $V(k, \mathbf{p}, t) \equiv kv(\mathbf{p}, t)$ , then differentiating  $V(k, \mathbf{p}, t)$  with respect to time  $t$  and evaluating the resulting expression at the period  $t$  data yields:

$$\partial \ln V(k, \mathbf{p}, t) / \partial t = \beta_0^* + \sum_{i=1}^N \beta_i^* \ln p_i^t \equiv T^t; \quad t = 1, \dots, T. \quad (10.153)$$

The right hand side of (10.153),  $T^t$ , is our desired measure of technical progress going from period  $t - 1$  to period  $t$ : it gives us an estimate of the percentage increase in variable profits due to the improvements in technology going from period  $t - 1$  to period  $t$ .<sup>\*27</sup>

In the following section, we generalize the translog model to allow for nonconstant returns to scale.

## 10.14 The Translog Variable Profit Function with Nonconstant Returns to Scale

The period  $t$  translog variable profit function is now defined as follows:

$$\ln V(k, \mathbf{p}, t) \equiv \alpha_0 + \beta_0 t + \sum_{i=1}^N \alpha_i \ln p_i + \sum_{i=1}^N \beta_i t \ln p_i + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln p_i \ln p_j$$

$$+ \delta_0 \ln k + \sum_{i=1}^N \delta_i \ln k \ln p_i + (1/2) \varepsilon [\ln k]^2 \quad (10.154)$$

<sup>\*27</sup> Since payments to capital are typically only about one third the size of payments to labour, it will turn out that "reasonable" estimates of technical progress  $T^t$  will be about 3 times the size of our index number estimates of total factor productivity growth. The total factor productivity growth rate is the rate of growth of outputs divided by the rate of growth of inputs, where the inputs are taken to be labour and capital services. Hence the denominator in this estimator of technical progress is approximately three times as big as the implicit denominator in  $T^t$  which is just capital input.

where the parameters  $\alpha_i$  and  $\gamma_{ij}$  satisfy the restrictions (10.132)-(10.134), the  $\beta_i$  parameters satisfy (10.140) and the new  $\delta_i$  parameters satisfy:

$$\sum_{i=1}^N \delta_i = 0. \quad (10.155)$$

The above restrictions ensure that the functional form is homogeneous of degree one in the prices  $\mathbf{p}$ . Comparing our new more general translog with the constant returns to scale function defined in the previous section, we see that we have added  $N$  new independent  $\delta_i$  parameters and one new  $\varepsilon$  parameter.\*28

Obviously, we can use (10.154) as an estimating equation. Defining  $V^t \equiv \mathbf{p}^{tT} \mathbf{y}^t$  as in the previous section and evaluating (10.154) at the period  $t$  data, we obtain the following estimating equation:

$$\begin{aligned} \ln V^t \equiv & \alpha_0 + \beta_0 t + \sum_{i=1}^N \alpha_i \ln p_i^t + \sum_{i=1}^N \beta_i t \ln p_i^t + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij} \ln p_i^t \ln p_j^t \\ & + \delta_0 \ln k^t + \sum_{i=1}^N \delta_i \ln k^t \ln p_i^t + (1/2) \varepsilon [\ln k^t]^2; \quad t = 1, \dots, T. \end{aligned} \quad (10.156)$$

We need to eliminate the redundant parameters in (10.156) as was done in the previous section for the case  $N = 4$ . This leads to the following estimating equations: for  $t = 1, \dots, T$ :

$$\begin{aligned} \ln(V^t/p_N^t) \equiv & \alpha_0 + \beta_0 t + \sum_{i=1}^{N-1} \alpha_i \ln(p_i^t/p_N^t) + \sum_{i=1}^{N-1} \beta_i t \ln(p_i^t/p_N^t) + \\ & (1/2) \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \gamma_{ij} \ln(p_i^t/p_N^t) \ln(p_j^t/p_N^t) + \delta_0 \ln k^t + \sum_{i=1}^{N-1} \delta_i \ln k^t \ln(p_i^t/p_N^t) + (1/2) \varepsilon [\ln k^t]^2. \end{aligned} \quad (10.157)$$

Differentiating  $V(k, \mathbf{p}, t)$  with respect to the components of the price vector  $\mathbf{p}$  and using Hotelling's Lemma leads to the following share equation counterparts to equations (10.142) in the previous section:

$$s_i^t \equiv p_i^t y_i^t / V^t = \alpha_i + \beta_i t + \sum_{j=1}^{N-1} \gamma_{ij} \ln(p_j^t/p_N^t) + \delta_i \ln k^t + e_i^t; \quad i = 1, \dots, N-1; t = 1, \dots, T. \quad (10.158)$$

Note that we have only  $N - 1$  independent estimating equations in (10.158).

It turns out that the formulae (10.145) and (10.146) in the previous section are still valid formulae for the *cross price elasticities of net supply*. Hence given econometric estimates for the  $\alpha_i, \beta_i, \delta_i, \varepsilon$  and the  $\gamma_{ij}$ , which we denote by  $\alpha_i^*, \beta_i^*, \delta_i^*, \varepsilon^*$  and  $\gamma_{ij}^*$ , the estimated or fitted shares in period  $t$ ,  $s_i^{t*}$  are defined using these estimates and equations (10.158) evaluated at the period  $t$  data:

$$s_i^{t*} \equiv \alpha_i^* + \beta_i^* t + \sum_{j=1}^N \gamma_{ij}^* \ln p_j^t + \delta_i \ln k^t; \quad i = 1, \dots, N; t = 1, \dots, T. \quad (10.159)$$

Now use equations (10.159) and (10.145) evaluated at the period  $t$  data and econometric estimates to obtain the following formula for the *period  $t$  cross elasticities of net supply*,  $e_{ij}^t$ :

$$e_{ij}^t \equiv \partial \ln y_i(k^t, \mathbf{p}^t) / \partial \ln p_j = [s_i^{t*}]^{-1} \gamma_{ij}^* + s_j^{t*}; \quad i \neq j. \quad (10.160)$$

Similarly, use equations (10.146) evaluated at the period  $t$  data and econometric estimates to obtain the following formula for the *period  $t$  own elasticities of net supply*,  $e_{ii}^t$ :

$$e_{ii}^t \equiv \partial \ln y_i(k^t, \mathbf{p}^t) / \partial \ln p_i = [s_i^{t*}]^{-1} \gamma_{ii}^* + s_i^{t*} - 1; \quad i = 1, \dots, N. \quad (10.161)$$

\*28 Thus if  $N = 4$ , we will have 19 independent parameters in all.

We can also obtain an estimated or *fitted period  $t$  variable profits or gross return to capital*,  $V^{t*}$ , by using our econometric estimates for the parameters and by exponentiating the right hand side of the equation  $t$  in (10.157):

$$V^{t*} \equiv \exp \left[ \ln k^t + \alpha_0^* + t\beta_0^* + \sum_{i=1}^N \alpha_i^* \ln p_i^t + \sum_{i=1}^N \beta_i^* t \ln p_i^t + (1/2) \sum_{i=1}^N \sum_{j=1}^N \gamma_{ij}^* \ln p_i^t \ln p_j^t + \delta_0 \ln k^t + \sum_{i=1}^N \delta_i \ln k^t \ln p_i^t + (1/2)\varepsilon [\ln k^t]^2 \right] \quad t = 1, \dots, T. \quad (10.162)$$

Our fitted period  $t$  shares  $s_i^{t*}$  defined by (10.159) and our fitted period  $t$  profits  $V^{t*}$  defined by (10.162) can be used in order to obtain estimated or *fitted period  $t$  net supplies*,  $y_i^{t*}$ , as follows:

$$y_i^{t*} \equiv V^{t*} s_i^{t*} / p_i^t; \quad i = 1, \dots, N; t = 1, \dots, T. \quad (10.163)$$

Given the matrix of period  $t$  estimated price elasticities of net supply,  $[e_{ij}^t]$ , we can readily calculate the matrix of period  $t$  *estimated net output price derivatives*,  $\nabla_p \mathbf{y}(k^t, \mathbf{p}^t) = \nabla_{pp}^2 V(k^t, \mathbf{p}^t)$ . Our estimate for element  $ij$  of  $\nabla_{pp}^2 V(k^t, \mathbf{p}^t)$  is:

$$V_{ij}^{t*} \equiv e_{ij}^t y_i^{t*} / p_j^t; \quad i, j = 1, \dots, N; t = 1, \dots, T \quad (10.164)$$

where the estimated period  $t$  elasticities  $e_{ij}^t$  are defined by (10.148) and (10.149) and the fitted period  $t$  net output supplies  $y_i^{t*}$  are defined by (10.163). Once the estimated price derivative matrices  $[V_{ij}^{t*}]$  have been calculated, then we may check whether each of them is positive semidefinite using determinantal conditions or by checking if all of the eigenvalues of each matrix are zero or positive. Differentiating  $V(k, \mathbf{p}, t)$  with respect to time  $t$  and evaluating the resulting expression at the period  $t$  data yields:

$$\partial \ln V(k, \mathbf{p}, t) / \partial t = \beta_0^* + \sum_{i=1}^N \beta_i^* \ln p_i^t \equiv T^t; \quad t = 1, \dots, T. \quad (10.165)$$

i.e., we obtain the same measure of technical progress that we obtained in the previous section.

There remains the problem of defining a measure of returns to scale. The measure we will use is one that calculates the percentage change in variable profits due to a one percent change in the use of capital. Thus our measure of returns to scale in period  $t$  is:

$$R^t \equiv \partial \ln V(k, \mathbf{p}, t) / \partial \ln k = \delta_0 + \sum_{i=1}^N \delta_i \ln p_i^t + \varepsilon \ln k^t; \quad t = 1, \dots, T. \quad (10.166)$$

Note that if we set  $\delta_0 = 1, \varepsilon = 0$  and  $\delta_i = 0$  for  $i = 1, \dots, N$ , then the model in this section collapses down to the model presented in the previous section. Under these restrictions, it can be seen that  $R^t = 1$ ; i.e., we have constant returns to scale.

When the above translog model is implemented, usually two things happen:

- The curvature conditions fail at one or more observations; i.e., the estimated period  $t$  substitution matrix  $[V_{ij}^{t*}]$  defined by (10.164) above *fails to be positive semidefinite* at one or more periods  $t$  and
- The estimates for the returns to scale  $R^t$  and for technical progress  $T^t$  are *not reasonable*.

The reason why we cannot usually determine accurate estimates for returns to scale and for technical progress is that usually  $k$  grows fairly smoothly through the sample period and hence the variables  $k$  and  $t$  tend to be highly multicollinear and so our estimates for  $R^t$  and  $T^t$  are not very well determined.

Thus in subsequent sections, we will impose constant returns to scale. We will also turn to the normalized quadratic functional form where the correct curvature conditions can be imposed without destroying the flexibility of the functional form.

## 10.15 The Normalized Quadratic Unit Profit Function Model

We adapt the normalized quadratic unit cost function defined by (10.56) in section 10.4 above into a unit profit function. Thus define the production unit's period  $t$  variable profit function  $V(k, \mathbf{p}, t)$  as follows:

$$V(k, \mathbf{p}, t) \equiv \mathbf{b}^T \mathbf{p}k + (1/2)[\mathbf{p}^T \mathbf{B} \mathbf{p} / \alpha^T \mathbf{p}]k + \mathbf{c}^T \mathbf{p}tk \quad (10.167)$$

where  $\mathbf{b}^T \equiv [b_1, \dots, b_N]$  and  $\mathbf{c}^T \equiv [c_1, \dots, c_N]$  are parameter vectors and  $\mathbf{B} \equiv [b_{ij}]$  is a matrix of parameters. The matrix  $\mathbf{B}$  satisfies the following restrictions:

$$\mathbf{B} = \mathbf{B}^T; \text{ i.e., the matrix } \mathbf{B} \text{ is symmetric;} \quad (10.168)$$

$$\mathbf{B} \mathbf{p}^* = \mathbf{0}_N \text{ for some } \mathbf{p}^* \gg \mathbf{0}_N. \quad (10.169)$$

The vector of parameters  $\alpha^T \equiv [\alpha_1, \dots, \alpha_N]$  is predetermined and satisfies  $\alpha > \mathbf{0}_N$ . We can adapt the analysis presented in section 10.4 and show that *a necessary and sufficient condition for  $V(k, \mathbf{p}, t)$  defined by (10.167) above to be convex in prices is that the matrix  $\mathbf{B}$  be positive semidefinite.*

Differentiating the normalized quadratic variable profit function defined by (10.167) with respect to the components of the price vector  $\mathbf{p}$  leads to the following system of net supply functions using Hotelling's Lemma:

$$\mathbf{y}(k, \mathbf{p}, t) = \nabla_{\mathbf{p}} V(k, \mathbf{p}, t) = \mathbf{b}k + [(\alpha^T \mathbf{p})^{-1} \mathbf{B} \mathbf{p} - (1/2)(\alpha^T \mathbf{p})^{-2} \mathbf{p}^T \mathbf{B} \mathbf{p} \alpha]k + \mathbf{c}kt. \quad (10.170)$$

Evaluating (10.170) at the period  $t$  data, dividing both sides by  $k^t$  and adding a vector of errors  $\mathbf{e}^t$  leads to the following system of estimating equations:

$$\mathbf{y}^t / k^t = \mathbf{b} + \mathbf{B} \mathbf{v}^t - (1/2) \mathbf{v}^{tT} \mathbf{B} \mathbf{v}^t \alpha + \mathbf{c}t + \mathbf{e}^t; \quad t = 1, \dots, T \quad (10.171)$$

where the vector of period  $t$  normalized prices is defined as  $\mathbf{v}^t \equiv (\alpha^T \mathbf{p}^t)^{-1} \mathbf{p}^t$ .

We have not substituted the restrictions (10.169) into the estimating equations (10.171). We shall do this substitution below assuming that  $N = 4$  and  $\mathbf{p}^* = \mathbf{1}_4$ .

We use the restrictions (10.169) to solve for the  $b_{ii}$  in terms of the off diagonal  $b_{ij}$ . Thus we have, assuming that there are 4 variable commodities and  $\mathbf{p}^* = \mathbf{1}_4$  and using  $\mathbf{B} = \mathbf{B}^T$ :

$$b_{11} = -b_{12} - b_{13} - b_{14}; \quad (10.172)$$

$$b_{22} = -b_{12} - b_{23} - b_{24}; \quad (10.173)$$

$$b_{33} = -b_{13} - b_{23} - b_{34}; \quad (10.174)$$

$$b_{44} = -b_{14} - b_{24} - b_{34}. \quad (10.175)$$

Using (10.172)-(10.175), we can write  $\mathbf{B} \mathbf{v}$  as follows:

$$\begin{aligned} [\mathbf{B} \mathbf{v}]^T &= \left[ \sum_{j=1}^4 b_{1j} v_j, \sum_{j=1}^4 b_{2j} v_j, \sum_{j=1}^4 b_{3j} v_j, \sum_{j=1}^4 b_{4j} v_j \right] \\ &= [-b_{12} w_{12} - b_{13} w_{13} - b_{14} w_{14}, b_{12} w_{12} - b_{23} w_{23} - b_{24} w_{24}, \\ &\quad b_{13} w_{13} + b_{23} w_{23} - b_{34} w_{34}, b_{14} w_{14} + b_{24} w_{24} + b_{34} w_{34}] \end{aligned} \quad (10.176)$$

where

$$w_{ij} \equiv v_i - v_j; \quad i, j = 1, 2, 3, 4. \quad (10.177)$$

Premultiplying  $\mathbf{B} \mathbf{v}$  by  $\mathbf{v}^T$  and using (10.176) and (10.177) yields the following formula:

$$\mathbf{v}^T \mathbf{B} \mathbf{v} = -[b_{12}(w_{12})^2 + b_{13}(w_{13})^2 + b_{14}(w_{14})^2 + b_{23}(w_{23})^2 + b_{24}(w_{24})^2 + b_{34}(w_{34})^2]. \quad (10.178)$$

Now substitute (10.176) and (10.178) into (10.171) and we obtain the following system of estimating equations<sup>\*29</sup> in the case where  $N = 4$  and  $\mathbf{p}^* = \mathbf{1}_N$ :

$$y_1^t/k^t = b_1 + c_1t - b_{12}w_{12}^t - b_{13}w_{13}^t - b_{14}w_{14}^t - (1/2)\mathbf{v}^{tT}\mathbf{B}\mathbf{v}^t\boldsymbol{\alpha}_1 + e_1^t; \quad t = 1, \dots, T \quad (10.179)$$

$$y_2^t/k^t = b_2 + c_2t + b_{12}w_{12}^t - b_{23}w_{23}^t - b_{24}w_{24}^t - (1/2)\mathbf{v}^{tT}\mathbf{B}\mathbf{v}^t\boldsymbol{\alpha}_2 + e_2^t; \quad t = 1, \dots, T \quad (10.180)$$

$$y_3^t/k^t = b_3 + c_3t + b_{13}w_{13}^t + b_{23}w_{23}^t - b_{34}w_{34}^t - (1/2)\mathbf{v}^{tT}\mathbf{B}\mathbf{v}^t\boldsymbol{\alpha}_3 + e_3^t; \quad t = 1, \dots, T \quad (10.181)$$

$$y_4^t/k^t = b_4 + c_4t + b_{14}w_{14}^t + b_{24}w_{24}^t + b_{34}w_{34}^t - (1/2)\mathbf{v}^{tT}\mathbf{B}\mathbf{v}^t\boldsymbol{\alpha}_4 + e_4^t; \quad t = 1, \dots, T. \quad (10.182)$$

We need to also replace  $\mathbf{v}^{tT}\mathbf{B}\mathbf{v}^t$  in equations (10.179)-(10.180) by the right hand side of (10.178) evaluated at the period  $t$  data. The resulting estimating equations turn out to be linear in the unknown  $b_i, c_i$  and  $b_{ij}$  parameters (4 plus 4 plus 6 equals 14 parameters in all).

Returning to the case of a general number of variable commodities  $N$ , we need to calculate the matrix of net supply price derivatives. Differentiating (10.170) with respect to the components of  $\mathbf{p}$  yields the following matrix of price derivatives at period  $t$ :

$$\nabla_{\mathbf{p}}\mathbf{y}(k^t, \mathbf{p}^t, t) = \nabla_{\mathbf{p}}^2V(k^t, \mathbf{p}^t, t) = (\boldsymbol{\alpha}^T\mathbf{p}^t)^{-1}[\mathbf{B} - \mathbf{B}\mathbf{v}^t\boldsymbol{\alpha}^T - \boldsymbol{\alpha}\mathbf{v}^{tT}\mathbf{B} + \mathbf{v}^{tT}\mathbf{B}\mathbf{v}^t\boldsymbol{\alpha}\boldsymbol{\alpha}^T]k^t; \quad t = 1, \dots, T \quad (10.183)$$

where, as usual, the vector of period  $t$  normalized prices is defined as  $\mathbf{v}^t \equiv (\boldsymbol{\alpha}^T\mathbf{p}^t)^{-1}\mathbf{p}^t$ . Once estimates for  $\mathbf{b}, \mathbf{c}$  and  $\mathbf{B}$  have been obtained (call these estimates  $\mathbf{b}^*, \mathbf{c}^*$  and  $\mathbf{B}^*$  respectively), we can use equations (10.171) in order to obtain period  $t$  vectors of fitted net supply vectors  $\mathbf{y}^{t*}$ :

$$\mathbf{y}^{t*} \equiv k^t[\mathbf{b}^* + \mathbf{B}^*\mathbf{v}^t - (1/2)\mathbf{v}^{tT}\mathbf{B}^*\mathbf{v}^t\boldsymbol{\alpha} + \mathbf{c}^*t]; \quad t = 1, \dots, T. \quad (10.184)$$

Equations (10.183) and (10.184) may be used to form estimated *period  $t$  price elasticity matrices*:

$$[e_{ij}^t] \equiv [\partial \ln y_i(k^t, \mathbf{p}^t, t)/\partial \ln p_j] = [(p_j^t/y_i^{t*})\partial y_i(k^t, \mathbf{p}^t, t)/\partial p_j]; \quad t = 1, \dots, T \quad (10.185)$$

where the derivative estimates  $\partial y_i(k^t, \mathbf{p}^t, t)/\partial p_j$  can be obtained from (10.183).

An estimator of variable profits in period  $t$ ,  $V^{t*}$ , can be obtained as the inner product of the period  $t$  fitted net supply vector  $\mathbf{y}^{t*}$  defined by (10.184) and the period  $t$  vector of variable commodity prices,  $\mathbf{p}^t$ :

$$V^{t*} \equiv \mathbf{p}^{tT}\mathbf{y}^{t*}; \quad t = 1, \dots, T. \quad (10.186)$$

Finally, a measure of *period  $t$  technical progress*  $T^t$  can be defined as follows:

$$T^t \equiv \partial \ln V(k^t, \mathbf{p}^t, t)/\partial t = \mathbf{p}^{tT}\mathbf{c}^*k^t/V^{t*}; \quad t = 1, \dots, T. \quad (10.187)$$

Unfortunately, the estimated  $\mathbf{B}^*$  matrix may fail to be positive semidefinite. Hence, in the following section, we adapt the technique used in section 10.4 above to impose the correct curvature conditions on the  $\mathbf{B}$  matrix.

## 10.16 The Normalized Quadratic Unit Profit Function Model with Curvature Imposed

If the estimated  $\mathbf{B}$  matrix turns out to be not positive definite, then we can rerun the model in the previous section by replacing  $\mathbf{B}$  by:

$$\mathbf{B} = \mathbf{A}\mathbf{A}^T \quad (10.188)$$

<sup>\*29</sup> An alternative system of estimating equations multiplies both sides of (10.179)-(10.182) by  $k^t$ . This alternative system of estimating equations often performs "better" in the sense that it leads to more reasonable estimates of net supply elasticities. However, in theory, the original system of estimating equations (10.179)-(10.182) should have more homoskedastic variances.

where  $\mathbf{A}$  is a lower triangular matrix which satisfies:

$$\mathbf{A}^T \mathbf{p}^* = \mathbf{0}_N. \quad (10.189)$$

For the case  $N = 4$  and for  $\mathbf{p}^* = \mathbf{1}_4$ , we can use the restrictions (10.189) and the lower triangular structure of  $\mathbf{A}$  in order to eliminate the  $a_{ii}$  as follows:

$$a_{11} = -a_{21} - a_{31} - a_{41}; \quad (10.190)$$

$$a_{22} = -a_{32} - a_{42}; \quad (10.191)$$

$$a_{33} = -a_{43}; \quad (10.192)$$

$$a_{44} = 0. \quad (10.193)$$

If we substitute (10.190)-(10.193) into (10.188), we obtain the following formulae for the  $b_{ij}$  in terms of the  $a_{ij}$ :

$$b_{11} = a_{11}^2 = [a_{21} + a_{31} + a_{41}]^2 \quad (10.194)$$

$$b_{12} = a_{11}a_{21} = -[a_{21} + a_{31} + a_{41}]a_{21} \quad (10.195)$$

$$b_{13} = a_{11}a_{31} = -[a_{21} + a_{31} + a_{41}]a_{31} \quad (10.196)$$

$$b_{14} = a_{11}a_{41} = -[a_{21} + a_{31} + a_{41}]a_{41} \quad (10.197)$$

$$b_{22} = a_{21}^2 + a_{22}^2 = a_{21}^2 + [a_{32} + a_{42}]^2 \quad (10.198)$$

$$b_{23} = a_{21}a_{31} + a_{22}a_{32} = a_{21}a_{31} - [a_{32} + a_{42}]a_{32} \quad (10.199)$$

$$b_{24} = a_{21}a_{41} + a_{22}a_{42} = a_{21}a_{41} - [a_{32} + a_{42}]a_{42} \quad (10.200)$$

$$b_{33} = a_{31}^2 + a_{32}^2 + a_{33}^2 = a_{31}^2 + a_{32}^2 + a_{43}^2 \quad (10.201)$$

$$b_{34} = a_{31}a_{41} + a_{32}a_{42} + a_{33}a_{43} = a_{31}a_{41} + a_{32}a_{42} + a_{43}^2 \quad (10.202)$$

$$b_{44} = a_{41}^2 + a_{42}^2 + a_{43}^2 = a_{41}^2 + a_{42}^2 + a_{43}^2. \quad (10.203)$$

Now we need only replace the  $b_{ij}$  parameters which occurred in the model of the previous section by the formulae on the right hand sides of (10.194)-(10.203) and run the previous model as a nonlinear regression. The parameters of the new model are the elements of the vectors  $\mathbf{b}$  and  $\mathbf{c}$  (as before) and the elements of the  $\mathbf{A}$  matrix.

Once the  $a_{ij}$  have been estimated, the  $b_{ij}$  parameters can be computed using (10.188) (or (10.194)-(10.203) if  $N = 4$  and  $\mathbf{p}^* = \mathbf{1}_4$ ) and the elasticity formulae (10.185) and the estimates of technical progress (10.187) in the previous section can be computed.

It turns out that we can use spline techniques in the production context as well as in the consumer context. In the following section, we indicate how technical progress can be modeled using spline techniques.

## 10.17 The Normalized Quadratic Unit Profit Function Model and the Use of Splines for Modeling Technical Progress

We adapt the normalized quadratic profit function defined by (10.167) in section 10.15 above into a spline model. We illustrate the technique by developing the algebra for a model with two break points. Thus define the production unit's period  $t$  variable profit function  $V(k, \mathbf{p}, t)$  as follows:

$$V(k, \mathbf{p}, t) \equiv \mathbf{b}^T \mathbf{p}k + (1/2)[\mathbf{p}^T \mathbf{B} \mathbf{p} / \alpha^T \mathbf{p}]k + d(k, \mathbf{p}, t) \quad (10.204)$$

where  $\mathbf{b}^T \equiv [b_1, \dots, b_N]$  is a parameter vectors and  $\mathbf{B} \equiv [b_{ij}]$  is a matrix of parameters. The matrix  $\mathbf{B}$  satisfies the restrictions (10.168) and (10.169) and is positive semidefinite. As usual, the vector

of parameters  $\boldsymbol{\alpha}^T \equiv [\alpha_1, \dots, \alpha_N]$  is predetermined and satisfies  $\boldsymbol{\alpha} > \mathbf{0}_N$ . The linear spline function  $d(k, \mathbf{p}, t)$  is defined as follows:

$$d(k, \mathbf{p}, t) \equiv \begin{cases} k\mathbf{c}^T \mathbf{p}t & \text{for } 1 \leq t \leq t^* \\ k\mathbf{c}^T \mathbf{p}t^* + (t - t^*)k\mathbf{f}^T \mathbf{p} & \text{for } t^* < t \leq t^{**} \\ k\mathbf{c}^T \mathbf{p}t^* + (t^{**} - t^*)k\mathbf{f}^T \mathbf{p} + (t - t^{**})k\mathbf{g}^T \mathbf{p} & \text{for } t^{**} < t \leq T \end{cases} \quad (10.205)$$

where  $\mathbf{c}^T \equiv [c_1, \dots, c_N]$ ,  $\mathbf{f}^T \equiv [f_1, \dots, f_N]$  and  $\mathbf{g}^T \equiv [g_1, \dots, g_N]$  are parameter vectors to be estimated. The periods  $t^*$  and  $t^{**}$  are *break points* where the rate of technological change shifts from one regime to another. These break points are to be chosen by the investigator.

Differentiating the normalized quadratic variable profit function defined by (10.204) with respect to the components of the price vector  $\mathbf{p}$  leads to the following system of net supply functions using Hotelling's Lemma:

$$\mathbf{y}(k, \mathbf{p}, t) = \nabla_{\mathbf{p}} V(k, \mathbf{p}, t) = \mathbf{b}k + [(\boldsymbol{\alpha}^T \mathbf{p})^{-1} \mathbf{B} \mathbf{p} - (1/2)(\boldsymbol{\alpha}^T \mathbf{p})^{-2} \mathbf{p}^T \mathbf{B} \mathbf{p} \boldsymbol{\alpha}]k + \nabla_{\mathbf{p}} d(k, \mathbf{p}, t). \quad (10.206)$$

Evaluating (10.206) at the period  $t$  data, dividing both sides by  $k^t$  and adding a vector of errors  $\mathbf{e}^t$  leads to the following system of estimating equations:

$$\mathbf{y}^t/k^t = \mathbf{b} + \mathbf{B}\mathbf{v}^t - (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t \boldsymbol{\alpha} + \mathbf{c}t + \mathbf{e}^t; \quad t = 1, \dots, t^*; \quad (10.207)$$

$$\mathbf{y}^t/k^t = \mathbf{b} + \mathbf{B}\mathbf{v}^t - (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t \boldsymbol{\alpha} + \mathbf{c}t^* + (t - t^*)\mathbf{f} + \mathbf{e}^t; \quad t = t^* + 1, \dots, t^{**}; \quad (10.208)$$

$$\mathbf{y}^t/k^t = \mathbf{b} + \mathbf{B}\mathbf{v}^t - (1/2)\mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t \boldsymbol{\alpha} + \mathbf{c}t^* + (t^{**} - t^*)\mathbf{f} + (t - t^{**})\mathbf{g} + \mathbf{e}^t; \quad t = t^{**} + 1, \dots, T. \quad (10.209)$$

where, as usual, the vector of period  $t$  normalized prices is defined as  $\mathbf{v}^t \equiv (\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1} \mathbf{p}^t$ . It can be seen that the model defined by the estimating equations (10.207)-(10.209) is linear in the unknown parameters (but their are cross equation equality constraints on the parameters in the  $\mathbf{B}$  matrix).

It turns out that equations (10.183) are still valid in order to calculate the period  $t$  matrix of price derivatives of net supply; i.e., we have the following matrix of price derivatives at period  $t$ :

$$\nabla_{\mathbf{p}} \mathbf{y}(k^t, \mathbf{p}^t, t) = \nabla_{\mathbf{p}}^2 V(k^t, \mathbf{p}^t, t) = (\boldsymbol{\alpha}^T \mathbf{p}^t)^{-1} [\mathbf{B} - \mathbf{B}\mathbf{v}^t \boldsymbol{\alpha}^T - \boldsymbol{\alpha} \mathbf{v}^{tT} \mathbf{B} + \mathbf{v}^{tT} \mathbf{B}\mathbf{v}^t \boldsymbol{\alpha} \boldsymbol{\alpha}^T] k^t; \quad t = 1, \dots, T. \quad (10.210)$$

Obviously equations (10.207)-(10.210) can be used to generate fitted net supply vectors  $\mathbf{y}^{t*}$  and then equations (10.210) and (10.185) may be used to form estimated *period  $t$  price elasticity matrices*.

How should one pick the break points  $t^*$  and  $t^{**}$ ? We examine the plots of the regression model defined by (10.207)-(10.209) and look for an observation numbers where the plot changes from a zig to a zag. Suppose that for most of the equations, these change of directions occur at periods  $t^*$  and  $t^{**}$ . This will determine the break points. Additional break points can be added if necessary.

The *period  $t$  measures of technical progress*  $T^t$  are defined as follows:

$$T^t \equiv \partial \ln V(k^t, \mathbf{p}^t, t) / \partial t = \mathbf{p}^{tT} \mathbf{c}^* k^t / V^{t*}; \quad t = 1, \dots, t^*; \quad (10.211)$$

$$T^t \equiv \partial \ln V(k^t, \mathbf{p}^t, t) / \partial t = \mathbf{p}^{tT} \mathbf{f}^* k^t / V^{t*}; \quad t = t^* + 1, \dots, t^{**}; \quad (10.212)$$

$$T^t \equiv \partial \ln V(k^t, \mathbf{p}^t, t) / \partial t = \mathbf{p}^{tT} \mathbf{g}^* k^t / V^{t*}; \quad t = t^{**} + 1, \dots, T. \quad (10.213)$$

Our estimates of the rate of technological change will change discontinuously as we cross the break points, which is perhaps a disadvantage of this spline model.

Of course, if the estimated  $\mathbf{B}$  matrix turns out to be *not* positive semidefinite, then we may replace  $\mathbf{B}$  by  $\mathbf{A}\mathbf{A}^T$  as in the previous section.

## 10.18 Allowing for Flexibility at Two Sample Points

If we differentiate the normalized quadratic profit function defined by (10.167) above with respect to the  $m$ th component of the price vector  $\mathbf{p}$ , we obtain the following equation that describes the net supply of commodity  $m$  as a function of the price vector  $\mathbf{p}$  in period  $t$ :

$$y_m(k, \mathbf{p}, t) = b_m k + c_m t k + \sum_{j=1}^N b_{mj} (p_j / \boldsymbol{\alpha}^T \mathbf{p}) k - (1/2) \alpha_m \mathbf{p}^T \mathbf{B} \mathbf{p} / (\boldsymbol{\alpha}^T \mathbf{p})^2 k. \quad (10.214)$$

Now differentiate (10.214) with respect to  $p_n$ , the  $n$ th component of the price vector  $\mathbf{p}$ :

$$\begin{aligned} \partial y_m(k, \mathbf{p}, t) / \partial p_n &= b_{mn} k / \boldsymbol{\alpha}^T \mathbf{p} - \sum_{j=1}^N b_{mj} \alpha_n p_j k / (\boldsymbol{\alpha}^T \mathbf{p})^2 - \sum_{j=1}^N b_{nj} \alpha_m p_j k / (\boldsymbol{\alpha}^T \mathbf{p})^2 \\ &\quad + (1/2) \alpha_m \alpha_n \mathbf{p}^T \mathbf{B} \mathbf{p} k / (\boldsymbol{\alpha}^T \mathbf{p})^3. \end{aligned} \quad (10.215)$$

Now turn (10.215) into the cross elasticity of net supply of commodity  $m$  with respect to a change in the price of commodity  $n$ ,  $e_{mn}$ :

$$\begin{aligned} e_{mn}(\mathbf{p}, t) &\equiv [p_n / y_m] \partial y_m(k, \mathbf{p}, t) / \partial p_n \\ &= b_{mn} (p_n / \boldsymbol{\alpha}^T \mathbf{p}) (k / y_m) - [p_n / y_m] \sum_{j=1}^N b_{mj} \alpha_n p_j k / (\boldsymbol{\alpha}^T \mathbf{p})^2 \\ &\quad - [p_n / y_m] \sum_{j=1}^N b_{nj} \alpha_m p_j k / (\boldsymbol{\alpha}^T \mathbf{p})^2 + [p_n / y_m] (1/2) \alpha_m \alpha_n \mathbf{p}^T \mathbf{B} \mathbf{p} k / (\boldsymbol{\alpha}^T \mathbf{p})^3. \end{aligned} \quad (10.216)$$

Using the restrictions (10.169), the last three terms on the right hand side of (10.216) will be zero when  $\mathbf{p} = \mathbf{p}^*$  and thus, empirically, these last three terms will typically be small in magnitude. Thus, the key determinant of the magnitude of the elasticity  $e_{mn}$  will typically be the first term on the right hand side of (10.216), namely,  $b_{mn} (p_n / \boldsymbol{\alpha}^T \mathbf{p}) (k / y_m)$ . Of course, the parameter  $b_{mn}$  will be constant over time but the other terms,  $p_n$  (the price of commodity  $n$ ),  $y_m$  (the net output of commodity  $m$ ),  $k$  (the amount of the “fixed” factor) and  $\boldsymbol{\alpha}^T \mathbf{p}$  (a fixed basket price index of all  $N$  variable input and output prices) can all have substantial trends over our sample period. Thus, our chosen functional form has built in these possible trends in elasticities.

A solution to this problem is readily at hand but at a cost in terms of using up degrees of freedom. We have followed the example of most applied production function researchers and allowed technical progress to affect the constant terms in the system of net supply functions (10.214) but we have left the substitution matrix  $\mathbf{B}$  unchanged over time. To solve the problem of trending elasticities, all we have to do is allow  $\mathbf{B}$  to change over time as well. Thus, simply set the matrix  $\mathbf{B}$  in (10.167) and (10.214) equal to a weighted average of a matrix  $\mathbf{C}$  (which characterizes substitution possibilities at the beginning of the sample period) and a matrix  $\mathbf{D}$  (which characterizes substitution possibilities at the end of the sample period); i.e., define  $\mathbf{B}$  as follows in terms of  $\mathbf{C}$  and  $\mathbf{D}$  and the time variable  $t$ :

$$\mathbf{B}^t = (1 - [t/T])\mathbf{C} + [t/T]\mathbf{D}; \quad t = 0, 1, 2, \dots, T. \quad (10.217)$$

Note that there are  $T + 1$  sample observations. Essentially, we now let technical progress affect not only the constant terms in (10.167) but we also allow it to affect substitution possibilities as well. Another way of viewing our new functional form is that we allow the functional form to be flexible at two points (the first sample point and the last) instead of the usual one point.

As usual, the correct curvature conditions can be imposed globally (globally) by setting  $\mathbf{C}$  and  $\mathbf{D}$  equal to the product of  $\mathbf{U}\mathbf{U}^T$  and  $\mathbf{V}\mathbf{V}^T$  respectively, where  $\mathbf{U}$  and  $\mathbf{V}$  are lower triangular matrices; i.e., set:

$$\mathbf{C} = \mathbf{U}\mathbf{U}^T \text{ and } \mathbf{D} = \mathbf{V}\mathbf{V}^T; \quad \mathbf{U} \text{ and } \mathbf{V} \text{ lower triangular.} \quad (10.218)$$

We can also impose the following normalizations on the matrices  $\mathbf{U}$  and  $\mathbf{V}$ :

$$\mathbf{U}^T \mathbf{p}^* = \mathbf{0}_N; \quad \mathbf{V}^T \mathbf{p}^* = \mathbf{0}_N. \quad (10.219)$$

This technique of imposing price flexibility at two points is due to Diewert and Lawrence (2002)[143].

## 10.19 Semiflexible Functional Forms

Recall the basic normalized quadratic functional form for a unit profit function that was defined in section 10.16 above and recall that most of the unknown parameters for this functional form are in the  $\mathbf{B}$  equals  $\mathbf{A}\mathbf{A}^T$  matrix where  $\mathbf{A}$  is an  $N \times N$  lower triangular matrix which satisfies the restrictions (10.189),  $\mathbf{A}^T \mathbf{p}^* = \mathbf{0}_N$ .

In models where the number of commodities  $N$  is large, it can be difficult to estimate all of the parameters of the  $\mathbf{A}$  matrix at one time. An effective way to estimate the  $\mathbf{A}$  matrix is to estimate it one column at a time. Thus in the first stage, we use the estimating equations (10.179)-(10.182) or (10.214) with the  $\mathbf{A}$  (and hence  $\mathbf{B}$ ) matrix set equal to zero. Then at the next stage we use the estimates for the parameters which are not in the  $\mathbf{B}$  matrix as starting values for the stage 2 nonlinear regression model with  $\mathbf{B}$  set equal to  $\mathbf{A}\mathbf{A}^T$  where  $\mathbf{A}$  is a rank 1 lower triangular matrix; i.e., at this second stage,  $\mathbf{A}$  is set equal to:<sup>\*30</sup>

$$\mathbf{A} \equiv \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & 0 & \cdots & 0 \end{bmatrix}. \quad (10.220)$$

The estimated parameters from this stage 2 nonlinear regression are then used as starting values in a stage 3 nonlinear regression that fills in column 2 of the lower triangular matrix  $\mathbf{A}$ ; i.e., in the stage 3 regression,  $\mathbf{A}$  is set equal to the following rank 2 lower triangular matrix:<sup>\*31</sup>

$$\mathbf{A} \equiv \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & 0 \end{bmatrix}. \quad (10.221)$$

This procedure of gradually adding nonzero columns to the  $\mathbf{A}$  matrix can be continued until the full number of  $N - 1$  nonzero columns have been added, provided that the number of time series observations  $T$  is large enough compared to  $N$ , the number of commodities in the model. However, in models where  $T$  is small relative to  $N$ , the above procedure of adding nonzero columns to  $\mathbf{A}$  will have to be stopped well before the maximum number of  $N - 1$  nonzero columns has been added, due to the lack of degrees of freedom. Suppose that we stop the above procedure after  $K < N - 1$  nonzero columns have been added. Then Diewert and Wales (1988b; 330)[155] call the resulting normalized quadratic functional form a *flexible of degree  $K$*  functional form or a *semiflexible functional form*. A

<sup>\*30</sup> We also need to use the restrictions (10.189) to express  $a_{11}$  in terms of  $a_{21}, \dots, a_{N1}$ . Thus if  $\mathbf{p}^*$  is a vector of ones, the  $a_{11}$  in (10.220) is replaced by  $-a_{21} - a_{31} \dots - a_{N1}$ . If maximum likelihood estimation is used, then in the stage 2 nonlinear regression, the starting values for  $a_{21}, \dots, a_{N1}$  are taken to be 0's so the starting log likelihood for the stage 2 nonlinear regression will be equal to the final log likelihood of the stage 1 regression. This provides a check on the programming code used. A similar strategy should be used with the subsequent stage 3, 4 and so on regressions.

<sup>\*31</sup> The starting values for the stage 3 nonlinear regression for the elements in the first column of  $\mathbf{A}$  are the final estimated values from the stage 2 nonlinear regression and the starting values for the elements in the second column of  $\mathbf{A}$  are 0's. Again, if  $\mathbf{p}^*$  is a vector of ones, the  $a_{22}$  in (10.221) is replaced by  $-a_{32} - a_{42} \dots - a_{N2}$ .

flexible of degree  $K$  functional form for a profit or cost function can approximate an arbitrary twice continuously differentiable functional form to the second order at some point, except the matrix of second order partial derivatives of the functional form with respect to prices is restricted to have maximum rank  $K$  instead of the maximum possible rank,  $N - 1$ .

What is the cost of estimating a semiflexible functional form for a profit function instead of a fully flexible functional form? When we estimate a fully flexible functional form, we need the  $\mathbf{B}$  matrix to be able to approximate an arbitrary positive semidefinite symmetric matrix  $\mathbf{B}^*$  of rank  $N - 1$ . This arbitrary  $\mathbf{B}^*$  can be represented as a sum of  $N - 1$  rank one positive semidefinite matrices as we now show.

Recall that any symmetric matrix can be diagonalized by means of an orthonormal transformation; i.e., there exists a matrix  $\mathbf{U}$  equal to  $[\mathbf{u}^1, \mathbf{u}^2, \dots, \mathbf{u}^N]$ , where the  $\mathbf{u}^n$  are the columns of  $\mathbf{U}$ , such that:

$$\mathbf{U}^T \mathbf{B} \mathbf{U} = \mathbf{\Lambda} \equiv \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix} \quad (10.222)$$

where  $\mathbf{U}$  satisfies

$$\mathbf{U}^T \mathbf{U} = \mathbf{I}_N \quad (10.223)$$

and  $\mathbf{\Lambda}$  is a diagonal matrix with the nonnegative eigenvalues of  $\mathbf{B}$ , the  $\lambda_n$ , running down the main diagonal. We order these eigenvalues starting with the biggest and ending up with the smallest (which is equal to 0):

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N-1} \geq \lambda_N = 0. \quad (10.224)$$

Now premultiply both sides of (10.222) by  $\mathbf{U}$  and post multiply both sides of (10.222) by  $\mathbf{U}^T$ . Using (10.223), we find that:

$$\begin{aligned} \mathbf{B} &= \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \\ &= [\mathbf{u}^1 \lambda_1, \mathbf{u}^2 \lambda_2, \dots, \mathbf{u}^N \lambda_N] [\mathbf{u}^1, \mathbf{u}^2, \dots, \mathbf{u}^N]^T \\ &= \sum_{n=1}^N \lambda_n \mathbf{u}^n \mathbf{u}^{nT} \\ &= \sum_{n=1}^{N-1} \lambda_n \mathbf{u}^n \mathbf{u}^{nT} \end{aligned} \quad (10.225)$$

where the last equality in (10.225) follows from the fact that  $\lambda_N = 0$ .

If we estimate a normalized quadratic that is flexible of degree  $K$ , then it turns out that the resulting  $\mathbf{A} \mathbf{A}^T$  matrix can approximate  $\mathbf{B}$  defined by (10.225) as follows:

$$\mathbf{A} \mathbf{A}^T = \sum_{n=1}^K \lambda_n \mathbf{u}^n \mathbf{u}^{nT}. \quad (10.226)$$

Thus the cost of using a semiflexible functional form of degree  $K$  where  $K$  is less than  $N - 1$  is that we will miss out on the part of  $\mathbf{B}$  that corresponds to the smallest eigenvalues of  $\mathbf{B}$ ; i.e., our estimating  $\mathbf{A} \mathbf{A}^T$  will be too small by the positive semidefinite matrix  $\sum_{n=K}^{N-1} \lambda_n \mathbf{u}^n \mathbf{u}^{nT}$ . In many situations, this cost will be very small; i.e., as we go through the various stages of estimating  $\mathbf{A}$  by adding an extra nonzero column to  $\mathbf{A}$  at each stage, we can monitor the increase in the final log likelihood (if we use maximum likelihood estimation) and when the increase in stage  $k + 1$  over stage  $k$  is "small", we can stop adding extra columns, secure in the knowledge that we are not underestimating the size of  $\mathbf{B}$  by a large amount.

This semiflexible technique has not been widely applied but it would seem to offer some big advantages in estimating substitution matrices in situations where there are a large number of commodities in the model.\*<sup>32</sup>

## 10.20 References

- Christensen, L.R., D.W. Jorgenson and L.J. Lau (1971), "Conjugate Duality and the Transcendental Logarithmic Production Function," *Econometrica* 39, 255-256.
- Christensen, L.R., D.W. Jorgenson and L.J. Lau (1975), "Transcendental Logarithmic Utility Functions", *American Economic Review* 65, 367-383.
- Diewert, W.E. (1971), "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function", *Journal of Political Economy* 79, 481-507.
- Diewert, W.E. (1974a), "Applications of Duality Theory," pp. 106-171 in M.D. Intriligator and D.A. Kendrick (ed.), *Frontiers of Quantitative Economics*, Vol. II, Amsterdam: North-Holland.
- Diewert, W.E. (1974b), "Functional Forms for Revenue and Factor Requirements Functions", *International Economic Review* 15, 119-130.
- Diewert, W.E. (1980), "Symmetry Conditions for Market Demand Functions", *The Review of Economic Studies* 47, 595-601.
- Diewert, W.E. and D. Lawrence (2002), "The Deadweight Costs of Capital Taxation in Australia", pp. 103-167 in *Efficiency in the Public Sector*, Kevin J. Fox (ed.), Boston: Kluwer Academic Publishers.
- Diewert, W.E. and T.J. Wales (1987), "Flexible Functional Forms and Global Curvature Conditions", *Econometrica* 55, 43-68.
- Diewert, W.E. and T.J. Wales (1988a), "Normalized Quadratic Systems of Consumer Demand Functions", *Journal of Business and Economic Statistics* 6, 303-12.
- Diewert, W.E. and T.J. Wales (1988b), "A Normalized Quadratic Semiflexible Functional Form", *Journal of Econometrics* 37, 327-42.
- Diewert, W.E. and T.J. Wales (1992), "Quadratic Spline Models For Producer's Supply and Demand Functions", *International Economic Review* 33, 705-722.
- Diewert, W.E. and T.J. Wales (1993), "Linear and Quadratic Spline Models for Consumer Demand Functions", *Canadian Journal of Economics* 26, 77-106.
- Hicks, J.R. (1941-42), "Consumers' Surplus and Index Numbers", *The Review of Economic Studies* 9, 126-137.
- Hicks, J.R. (1946), *Value and Capital*, Second Edition, Oxford: Clarendon Press.
- Hotelling, H. (1932), "Edgeworth's Taxation Paradox and the Nature of Demand and Supply Functions", *Journal of Political Economy* 40, 577-616.
- Leontief, W.W. (1941), *The Structure of the American Economy 1919-1929*, Cambridge, Massachusetts: Harvard University Press.
- Samuelson, P.A. (1974), "Complementarity—An Essay on the 40th Anniversary of the Hicks-Allen Revolution in Demand Theory", *Journal of Economic Literature* 12, 1255-1289.
- Shephard, R.W. (1953), *Cost and Production Functions*, Princeton: Princeton University Press.
- Walras, L. (1954), *Elements of Pure Economics*, (a translation by W. Jaffé of the Edition Définitive (1926) of the *Eléments d'économie pure*, first edition published in 1874), Homewood, Illinois: Richard D. Irwin.

\*<sup>32</sup> Diewert and Lawrence in some unpublished work have successfully estimated semiflexible models for profit functions for 40 to 45 commodities.

---

Wiley, D.E., W.H. Schmidt and W.J. Bramble (1973), "Studies of a Class of Covariance Structure Models", *Journal of the American Statistical Association* 68, 317-323.



## Chapter 11

# Linear Programming

### 11.1 Introduction

The theory of linear programming provides a good introduction to the study of constrained maximization (and minimization) problems where some or all of the constraints are in the form of inequalities rather than equalities. Many models in economics can be expressed as inequality constrained optimization problems. A linear program is a special case of this general class of problems where both the objective function and the constraint functions are linear in the decision variables.

Linear programming problems are important for a number of reasons:

- Many general constrained optimization problems can be approximated by a linear program.
- The mathematical prerequisites for studying linear programming are minimal; only a knowledge of matrix algebra is required.
- Linear programming theory provides a good introduction to the theory of duality in nonlinear programming.

Linear programs appear in many economic contexts but the exact form of the problems varies across applications. We shall present several equivalent formulations of the basic linear programming problem in this introductory section. In the following section, we provide a geometric interpretation of a linear program (LP) in activities space. In subsequent sections, we will present George Dantzig's (1963)[65] *simplex algorithm* for solving an LP.\*<sup>1</sup>

Our *first formulation* of the basic linear programming problem is:

$$\min_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A} \mathbf{x} = \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N \} \quad (11.1)$$

where  $\mathbf{c}^T \equiv [c_1, c_2, \dots, c_N]$  and  $\mathbf{b}^T \equiv [b_1, b_2, \dots, b_M]$  are  $N$  and  $M$  dimensional vectors of constants,  $\mathbf{A} \equiv [a_{mn}]$  is an  $M \times N$  matrix of constants and  $\mathbf{x}^T \equiv [x_1, x_2, \dots, x_N]$  is a *nonnegative*  $N$  dimensional vector of decision or choice variables. Thus this first form for a linear programming problem is the problem of minimizing a linear function  $\mathbf{c}^T \mathbf{x}$  in the vector of nonnegative variables  $\mathbf{x} \geq \mathbf{0}_N$  subject to  $M$  linear equality constraints, which are written in the form  $\mathbf{A} \mathbf{x} = \mathbf{b}$ .

Our *second formulation* of an LP is:

$$\max_{x_0, \mathbf{x}} \{ x_0 : x_0 + \mathbf{c}^T \mathbf{x} = 0; \mathbf{A} \mathbf{x} = \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N \} \quad (11.2)$$

---

\*<sup>1</sup> "Linear programming was developed by George B. Dantzig in 1947 as a technique for planning the diversified activities of the U.S. Air Force." Robert Dorfman, Paul A. Samuelson and Robert M. Solow (1958; 3)[160]. Dorfman, Samuelson and Solow go on to note that Dantzig's fundamental paper was circulated privately for several years and finally published as Dantzig (1951)[64]. A complete listing of Dantzig's early contributions to developing the theory of linear programming can be found in Dantzig (1963; 597-597)[65].

where  $x_0$  is a new scalar variable, which is defined by the equality constraint in (11.2); i.e.,  $x_0 \equiv -\mathbf{c}^T \mathbf{x}$  and so minimizing  $\mathbf{c}^T \mathbf{x}$  is equivalent to maximizing  $x_0$ . It can be seen that the first and second formulations of an LP are completely equivalent.

Our *third formulation* of an LP is the following problem:

$$\max_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A} \mathbf{x} \leq \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N \}. \quad (11.3)$$

The above problem can be transformed into a problem of the type defined by (11.2) as follows:

$$\max_{x_0, \mathbf{x}, \mathbf{s}} \{ x_0 : x_0 - \mathbf{c}^T \mathbf{x} = 0; \mathbf{A} \mathbf{x} + \mathbf{I}_M \mathbf{s} = \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N; \mathbf{s} \geq \mathbf{0}_M \} \quad (11.4)$$

where  $\mathbf{I}_M$  is the  $M \times M$  identity matrix. Note that we have converted the inequality constraints  $\mathbf{A} \mathbf{x} \leq \mathbf{b}$  into equality constraints by adding an  $M$  dimensional vector of nonnegative *slack variables*  $\mathbf{s}$  to  $\mathbf{A} \mathbf{x}$ .

Note that equality constraints such as  $\mathbf{A} \mathbf{x} = \mathbf{b}$  can be converted into inequality constraints by writing  $\mathbf{A} \mathbf{x} = \mathbf{b}$  as two sets of inequality constraints:  $\mathbf{A} \mathbf{x} \leq \mathbf{b}$  and  $-\mathbf{A} \mathbf{x} \leq -\mathbf{b}$ . Note also that it does not matter whether we are maximizing the objective function  $\mathbf{c}^T \mathbf{x}$  or minimizing  $\mathbf{c}^T \mathbf{x}$  since a problem involving the maximization of  $\mathbf{c}^T \mathbf{x}$  is equivalent to a problem involving the minimization of  $-\mathbf{c}^T \mathbf{x}$ .

In the above problems, the vectors of variables  $\mathbf{x}$  and  $\mathbf{s}$  were restricted to be nonnegative. This is not an essential restriction, since if a variable,  $x_1$  say, is allowed to be unrestricted, it may be written in terms of two nonnegative variables, say  $s_1$  and  $s_2$ , as  $x_1 = s_1 - s_2$ . However, in most economic problems, restricting the decision variables  $\mathbf{x}$  to be nonnegative will be a reasonable assumption and it will not be necessary to resort to the  $x_1 = s_1 - s_2$  construction.

Thus the essence of a *linear programming problem* is that the objective function (the function that we are maximizing or minimizing) is *linear* and the constraint functions are *linear equalities or inequalities*.

It turns out that our second formulation of an LP is the most convenient one for actually solving an LP (the simplex algorithm due to Dantzig) but the third formulation is the most convenient one for proving the duality theorem for linear programs.

## 11.2 The Geometric Interpretation of a Linear Program in Activities Space

Consider the following LP:

$$\min_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A} \mathbf{x} \geq \mathbf{b}; \mathbf{x} \geq \mathbf{0}_2 \} \quad (11.5)$$

where  $\mathbf{c}^T \equiv [1, 1]$ ,  $\mathbf{b}^T \equiv [1, 2]$  and the two rows of  $\mathbf{A}$  are  $A_1 \equiv [1, 0]$  and  $A_2 \equiv [-1, 1]$ . Thus we wish to minimize  $x_1 + x_2$  subject to the four inequality constraints:  $x_1 \geq 1$ ;  $-x_1 + x_2 \geq 2$ ;  $x_1 \geq 0$ ; and  $x_2 \geq 0$ . We define the constraint set or *feasible region* in  $\mathbf{x}$  space or *activities space* to be the set of  $\mathbf{x}$ 's which satisfy the constraints in (11.5). It is the shaded set in Figure 11.1 below. We also graph the level sets of the objective function  $x_1 + x_2$ ; i.e., these are the family of straight lines indexed by  $k$ ,  $L(k) \equiv \{ \mathbf{x} : \mathbf{c}^T \mathbf{x} = k \}$ . These level sets form a system of parallel straight lines. Obviously, the optimal  $\mathbf{x}$  solution to the LP problem will be that  $\mathbf{x}$  which belongs to the feasible region and which also belongs to the lowest level set of the objective function.

Note that the optimal solution to the LP lies on the boundary of the feasible region and that the optimal level set of the objective function is tangent to the feasible region. Finally, note that the optimal solution to the LP is at a *vertex* of the feasible region. This is typically what happens.

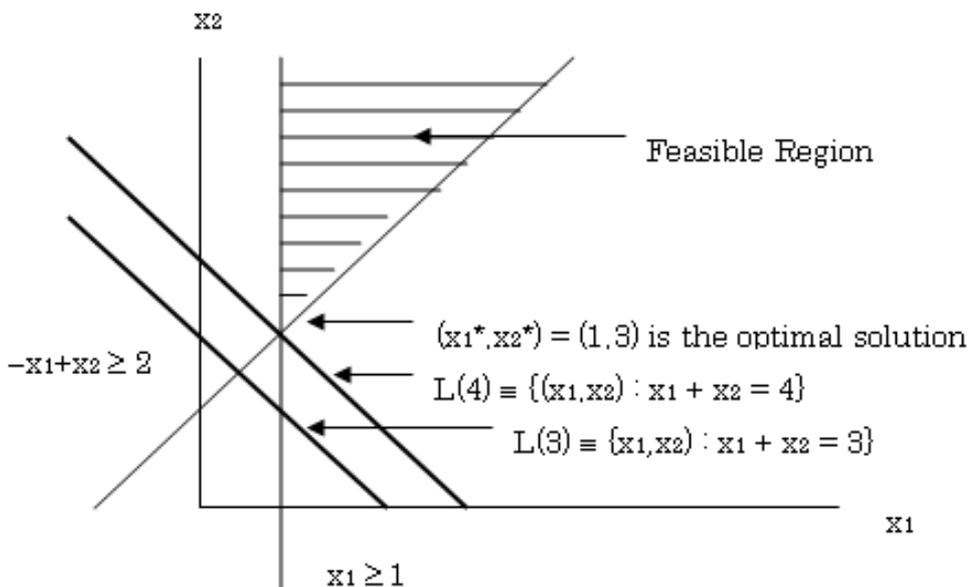


Fig. 11.1 A Linear Program in Activities Space

### 11.3 The Simplex Algorithm for Solving Linear Programs

In this section, we outline Dantzig’s (1963; chapters 5-7)[65] simplex algorithm for solving linear programming problems.\*<sup>2</sup> Dantzig’s method is not only of interest from a computational point of view, but also from a theoretical point of view, since it enables us to present an entirely algebraic proof of the duality theorem for linear programming problems as we shall see later. However, before the method can be explained, we need some additional definitions and results.

**Definition 1** Any solution  $\mathbf{x}^0$  to the system of linear equations  $\mathbf{Ax} = \mathbf{b}$  and inequalities  $\mathbf{x} \geq \mathbf{0}_N$  is called a *feasible solution* to the LP defined by (11.2).

We are considering the system of linear equations,  $\mathbf{Ax} = \mathbf{b}$ , which occurs in (11.2) above where  $\mathbf{A}$  is an  $M \times N$  matrix. In what follows, we assume that the number of equations  $M$  is less than the number of  $x_n$  variables, which is  $N$ . Hence  $\mathbf{A}$  is a singular matrix.

**Definition 2** Any nonnegative solution  $\mathbf{x}^0 \geq \mathbf{0}_N$  to the system of equations  $\mathbf{Ax} = \mathbf{b}$  with at least  $N - M$  components equal to zero is called a *basic feasible solution* to the LP defined by (11.2).

**Theorem 1** Carathéodory (1911; 200)[45], Fenchel (1953; 37)[179], Dantzig (1963; 113-114)[65]: If  $\mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}_N$  has a feasible solution, then it has a basic feasible solution.

**Proof.** Assume that  $M < N$  and that there exists an  $\mathbf{x}^* \equiv [x_1^*, \dots, x_N^*]^T \geq \mathbf{0}_N$  such that  $\mathbf{Ax}^* = \mathbf{b}$ . If  $\mathbf{x}^*$  has  $M$  or less nonzero components, then  $\mathbf{x}^*$  is a basic feasible solution and we are done. Thus we assume that  $\mathbf{x}^*$  has  $K > M$  nonzero components. By reordering the variables if necessary, we assume that the first  $K$  components of  $\mathbf{x}^*$  are the nonzero components. Now look at the following

\*<sup>2</sup> Actually, we present a version of Dantzig’s (1963; chapter 9)[65] revised simplex algorithm.

system of equations:

$$\sum_{k=1}^K A_{.k} \lambda_k = \mathbf{0}_M \quad (11.6)$$

where  $A_{.k}$  is the  $k$ th column of the  $\mathbf{A}$  matrix. The matrix  $[A_{.1}, A_{.2}, \dots, A_{.K}]$  is  $M \times K$  where  $K > M$  and hence this matrix is singular or alternatively, the  $K$  columns in this matrix are linearly dependent. Hence, there is a nonzero  $\boldsymbol{\lambda}^* \equiv [\lambda_1^*, \dots, \lambda_K^*]^T$  solution to (11.6). Without loss of generality, we can assume that at least one component of  $\boldsymbol{\lambda}^*$  is positive.\*<sup>3</sup> We also have  $x_k^* > 0$  for  $k = 1, 2, \dots, K$ . Hence the number  $\alpha$  defined by (11.7) below must be positive:

$$\alpha \equiv \max_k \{\lambda_k^*/x_k^* : k = 1, 2, \dots, K\} > 0. \quad (11.7)$$

Note that:

$$\alpha \geq \lambda_k^*/x_k^* \text{ or } \alpha x_k^* \geq \lambda_k^* \text{ for } k = 1, 2, \dots, K. \quad (11.8)$$

Since  $\mathbf{x}^* \geq \mathbf{0}_N$  is a feasible solution for  $\mathbf{Ax} = \mathbf{b}$ , we have:

$$\begin{aligned} \mathbf{b} &= \mathbf{Ax}^* \\ &= \sum_{k=1}^K A_{.k} x_k^* \\ &= \sum_{k=1}^K A_{.k} x_k^* - \alpha^{-1} \sum_{k=1}^K A_{.k} \lambda_k^* \quad \text{using (11.6)} \\ &= \alpha^{-1} \sum_{k=1}^K A_{.k} [\alpha x_k^* - \lambda_k^*]. \end{aligned} \quad (11.9)$$

But for all  $k$  which attain the max in (11.7), we will have  $\alpha x_k^* - \lambda_k^*$  equal to 0. Thus  $y_k^* \equiv \alpha^{-1}[\alpha x_k^* - \lambda_k^*]$  for  $k = 1, 2, \dots, K$  is a feasible solution for  $\mathbf{Ax} = \mathbf{b}$  with at least one additional zero component compared to our initial feasible solution  $\mathbf{x}^*$ .

We can continue this process of creating extra zero components as long as  $K$ , the number of nonzero variables, exceeds  $M$ , the number of equations. The procedure may stop when  $K = M$ , but at that stage, we have a basic feasible solution to  $\mathbf{Ax} = \mathbf{b}$ . ■

In order to initiate the simplex algorithm, we need to start the algorithm with a basic feasible solution. But the above results say that given an arbitrary feasible solution, we can construct a basic feasible solution in a finite number of steps and thus initiate the simplex algorithm.

We can rewrite the basic linear program defined by (11.2) above as follows:

$$\max_{x_0, x_1 \geq 0, x_2 \geq 0, \dots, x_N \geq 0} \{x_0 : \mathbf{e}_0 x_0 + \sum_{n=1}^N A_{.n}^* x_n = \mathbf{b}^*\} \quad (11.10)$$

where  $\mathbf{c}^T \equiv [c_1, \dots, c_N]$ ,  $\mathbf{A} \equiv [A_{.1}, A_{.2}, \dots, A_{.N}]$  is the original  $\mathbf{A}$  matrix and

$$A_{.n}^* \equiv \begin{bmatrix} c_n \\ A_{.n} \end{bmatrix} \text{ for } n = 1, \dots, N, \quad \mathbf{b}^* \equiv \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} \text{ and } \mathbf{e}_0 \equiv \begin{bmatrix} 1 \\ \mathbf{0}_M \end{bmatrix}. \quad (11.11)$$

In what follows, we also need to define the following  $M$  unit vectors of dimension  $M + 1$ :

$$\mathbf{e}_1^T \equiv [0, 1, 0, \dots, 0]; \mathbf{e}_2^T \equiv [0, 0, 1, 0, \dots, 0]; \dots; \mathbf{e}_M^T \equiv [0, 0, 0, \dots, 1]. \quad (11.12)$$

Thus  $\mathbf{e}_m$  is an  $M + 1$  dimensional vector consisting of 0 elements except entry  $m + 1$  is 1.

Assume that a feasible solution to (11.10) exists. Then by Theorem 1, a basic feasible solution exists. By relabelling variables if necessary, we can assume that the first  $M$  columns of  $\mathbf{A}$  correspond to the

\*<sup>3</sup> If all components of  $\boldsymbol{\lambda}^*$  were 0 or negative, then  $-\boldsymbol{\lambda}^*$  would satisfy (11.6), with all components 0 or positive.

nonzero variables in this basic feasible solution. Thus we have  $x_0^0, x_1^0 \geq 0, x_2^0 \geq 0, \dots, x_M^0 \geq 0$  such that:

$$\mathbf{e}_0 x_0^0 + \sum_{m=1}^M A_{.m}^* x_m^0 = \mathbf{b}^*. \quad (11.13)$$

At this point, we make two additional assumptions that will be relaxed at a later stage.

*Nonsingular Basis Matrix Assumption:*  $[\mathbf{e}_0, A_{.1}^*, A_{.2}^*, \dots, A_{.M}^*]^{-1} \equiv \mathbf{B}^{-1}$  exists.

*Nondegenerate Basis Matrix Assumption:*  $x_1^0 > 0, x_2^0 > 0, \dots, x_M^0 > 0$ ; i.e., the  $x_m^0$  which satisfy (11.13) for  $m = 1, 2, \dots, M$  are all positive.

We define the initial  $(M+1) \times (M+1)$  basis matrix  $\mathbf{B}$  as  $\mathbf{B} \equiv [\mathbf{e}_0, A_{.1}^*, A_{.2}^*, \dots, A_{.M}^*]$ . By the nonsingular basis matrix assumption, we have the existence of  $\mathbf{B}^{-1}$  and

$$\mathbf{B}^{-1}\mathbf{B} = \mathbf{B}^{-1}[\mathbf{e}_0, A_{.1}^*, A_{.2}^*, \dots, A_{.M}^*] = \mathbf{I}_{M+1} = [\mathbf{e}_0, \mathbf{e}_1, \dots, \mathbf{e}_M]. \quad (11.14)$$

Now premultiply both sides of the constraint equations in (11.10) by  $\mathbf{B}^{-1}$ . In view of (11.14), the resulting system of  $M+1$  equations is the following one:

$$\begin{aligned} \mathbf{e}_0 x_0 + \mathbf{e}_1 x_1 + \mathbf{e}_2 x_2 + \dots + \mathbf{e}_M x_M \\ + \mathbf{B}^{-1} A_{.M+1}^* x_{M+1} + \mathbf{B}^{-1} A_{.M+2}^* x_{M+2} + \dots + \mathbf{B}^{-1} A_{.N}^* x_N = \mathbf{B}^{-1} \mathbf{b}^*. \end{aligned} \quad (11.15)$$

If we set  $x_{M+1} = 0, x_{M+2} = 0, \dots, x_N = 0$  in (11.15), then we obtain our initial basic feasible solution to the LP problem:

$$x_m^0 = B_{m.}^{-1} \mathbf{b}^* > 0 \text{ for } m = 1, 2, \dots, M \quad (11.16)$$

where  $B_{m.}^{-1}$  is row  $m+1$  of  $\mathbf{B}^{-1}$  and the strict inequalities in (11.16) follow from the nondegenerate basis matrix assumption.

The question we now ask is: is our initial basic feasible solution to the LP (11.10) the optimal solution to the problem or not? To answer this question, look at equations (11.15) with  $x_{M+1}, x_{M+2}, \dots, x_N$  all equal to zero. If we try to increase any of these last  $N-M$   $x_n$  variables from its initial 0 level, then obviously, we will increase  $x_0$  only if  $B_{0.}^{-1} A_{.n}^* < 0$  for some  $n > M$ . Thus we have the following:

*Optimality Criterion:* If  $B_{0.}^{-1} A_{.n}^* \geq 0$  for  $n = M+1, M+2, \dots, N$ , then the current basis matrix and the corresponding solution (11.16) solve the LP (11.10). If  $B_{0.}^{-1} A_{.n}^* > 0$  for  $n = M+1, M+2, \dots, N$ , then the current basis matrix and the corresponding solution provide the unique solution to the LP (11.10). (11.17)

Since  $B_{0.}^{-1} A_{.n}^* = 0$  for  $n = 1, 2, \dots, M$  using (11.14), the first part of the optimality criterion can be changed to:

$$B_{0.}^{-1} A_{.n}^* \geq 0 \text{ for } n = 1, 2, \dots, N. \quad (11.18)$$

Suppose now that column  $s$  (not in the basis matrix) is such that

$$B_{0.}^{-1} A_{.s}^* < 0. \quad (11.19)$$

In view of the nondegeneracy assumption, it can be seen that we can increase the value of our objective function  $x_0$  and satisfy the inequality constraints in the LP (11.10) by increasing  $x_s$  from its initial 0 level. The question now is: which column should be dropped from the initial basis matrix  $\mathbf{B}$ ? To answer this question, look at the following system of equations:

$$\mathbf{e}_0 x_0 + \mathbf{e}_1 x_1 + \mathbf{e}_2 x_2 + \dots + \mathbf{e}_M x_M + \mathbf{B}^{-1} A_{.s}^* x_s = \mathbf{B}^{-1} \mathbf{b}^*. \quad (11.20)$$

Suppose that  $B_m^{-1}A_s^* \leq 0$  for  $m = 1, 2, \dots, M$ . Then as we increase  $x_s$  from its initial 0 level, the last  $M$  components of  $\mathbf{B}^{-1}A_s^*$  are either 0 or negative and these nonpositive numbers can be offset by increasing  $x_1, x_2, \dots, x_M$  from their initial values  $x_1^0, x_2^0, \dots, x_M^0$ . Thus we have a feasible solution to (11.10) which will allow the objective function to attain any large positive value. In this case, we obtain the following:

*Unbounded Solution Criterion:* If we have a column  $s$  such that  $B_0^{-1}A_s^* < 0$  and  $B_m^{-1}A_s^* \leq 0$  for  $m = 1, 2, \dots, M$ , then we may increase  $x_s$  to  $+\infty$  and obtain an unbounded solution to the LP (11.10). (11.21)

Suppose now that (11.19) holds but that  $B_m^{-1}A_s^* > 0$  for at least one index  $m \geq 1$ . In this case, as we increase  $x_s$  from its initial 0 level, in order to satisfy equations (11.20), we must decrease  $x_m$  from its initial positive level  $x_m^0 = B_m^{-1}\mathbf{b}^* > 0$ . Thus an upper bound to the amount that we can increase  $x_s$  before we violate the nonnegativity constraints  $\mathbf{x} \geq \mathbf{0}_N$  is  $B_m^{-1}\mathbf{b}^*/B_m^{-1}A_s^*$ . Now take the minimum of all such upper bounds over all indexes  $m$  such that  $B_m^{-1}A_s^* > 0$ :

$$x_s^* \equiv \min_m \{B_m^{-1}\mathbf{b}^*/B_m^{-1}A_s^* : m \geq 1 \text{ and } m \text{ is such that } B_m^{-1}A_s^* > 0\}. \quad (11.22)$$

The algebra in the above paragraph can be summarized as the following:

*Dropping Criterion:* If  $B_0^{-1}A_s^* < 0$  for some column index  $s$  not in the initial basis matrix and  $B_m^{-1}A_s^* > 0$  for at least one  $m \geq 1$ , then add column  $s$  to the basis matrix and drop any column  $r$  such that column  $r$  attains the minimum in (11.22); i.e.,  $r$  is such that  $r \geq 1$  and (11.23)

$$B_r^{-1}\mathbf{b}^*/B_r^{-1}A_s^* = \min_m \{B_m^{-1}\mathbf{b}^*/B_m^{-1}A_s^* : m \geq 1 \text{ and } m \text{ is such that } B_m^{-1}A_s^* > 0\}. \quad (11.24)$$

Using the nondegeneracy assumption, it can be seen that  $x_s^*$  defined by (11.22) is positive and introducing column  $s$  into the basis matrix leads to a strict increase in the objective function for the LP (11.10).

If the minimum in (11.22) or (11.24) is attained by a unique column  $r$ , then it can be seen that the new basis matrix is uniquely determined. It is also possible to show that the new basis matrix will also satisfy the nondegeneracy assumption.

The above criteria form the core of an effective algorithm for solving the LP (11.10).<sup>\*4</sup> Obviously, a finite but tremendously inefficient algorithm for solving (11.10) would involve solving the following system of  $M$  equations in  $M$  unknowns:

$$\sum_{j=1}^M A_{n_j} x_{n_j} = \mathbf{b} \quad (11.25)$$

for all possible choices of  $M$  of the  $N$  columns of the  $\mathbf{A}$  matrix, checking to see if the resulting  $x_{n_j}$  are nonnegative, evaluating the objective function at these basic feasible solutions and then the nonnegative solution that gave the lowest objective function  $\sum_{j=1}^M c_{n_j} x_{n_j}$  would be picked. What the simplex algorithm does is to search through basis matrices in such a way that we always increase our objective function.<sup>\*5</sup> Experience with the simplex algorithm has shown that an optimal basis matrix is generally reached in the order of  $M$  to  $3M$  iterations of the algorithm. *Thus the simplex algorithm is tremendously more efficient than simply searching through all of the basic feasible solutions for the linear program.*

<sup>\*4</sup> There are a few gaps but we will fill them in later.

<sup>\*5</sup> In general, we either increase the objective function or leave it unchanged at each iteration of the simplex algorithm or we find an unbounded solution.

## 11.4 An Example of the Simplex Algorithm

Let us use the simplex algorithm to solve the LP that was graphed in section 11.2 above. Rewriting the problem in the form (11.10) leads to the following problem:

$$\max_{x_0, x_1 \geq 0, x_2 \geq 0, s_1 \geq 0, s_2 \geq 0} \{x_0 : \mathbf{e}_0 x_0 + A_{.1}^* x_1 + A_{.2}^* x_2 + A_{.3}^* s_1 + A_{.4}^* s_2 = \mathbf{b}^*\} \quad (11.26)$$

$$\text{where } \mathbf{e}_0 \equiv \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, A_{.1}^* \equiv \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, A_{.2}^* \equiv \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, A_{.3}^* \equiv \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}, A_{.4}^* \equiv \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \text{ and } \mathbf{b}^* \equiv \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}.$$

Define the initial basis matrix to be

$$\mathbf{B} \equiv [\mathbf{e}_0, A_{.1}^*, A_{.2}^*] = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \equiv \begin{bmatrix} 1 & \mathbf{c}^T \\ \mathbf{0}_2 & \mathbf{A} \end{bmatrix} \quad (11.27)$$

where  $\mathbf{A}$  is a  $2 \times 2$  matrix. Using determinants, if  $\mathbf{A} \equiv \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ , we know that the formula for  $\mathbf{A}^{-1}$  is given by (provided that  $a_{11}a_{22} - a_{12}a_{21} \neq 0$ ):

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \text{ for our example.} \quad (11.28)$$

Using (11.28), we can readily calculate  $\mathbf{B}^{-1}$  as follows:

$$\mathbf{B}^{-1} = \begin{bmatrix} 1 & \mathbf{c}^T \\ \mathbf{0}_2 & \mathbf{A} \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -\mathbf{c}^T \mathbf{A}^{-1} \\ \mathbf{0}_2 & \mathbf{A}^{-1} \end{bmatrix} = \begin{bmatrix} 1 & -2 & -1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}. \quad (11.29)$$

Check for feasibility:

$$\begin{bmatrix} x_0^* \\ x_1^* \\ x_2^* \end{bmatrix} = \mathbf{B}^{-1} \mathbf{b}^* = \begin{bmatrix} 1 & -2 & -1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \\ 3 \end{bmatrix}. \quad (11.30)$$

To check that we have a basic feasible solution that satisfies the nonnegativity constraints, we need to check that  $x_1^* \geq 0$  and  $x_2^* \geq 0$ . Equations (11.30) show that these constraints are indeed satisfied.

Check for optimality:

We need to check that the inner product of the first row of  $\mathbf{B}^{-1}$  with each column not in the basis matrix is positive or zero.

$$B_0^{-1} A_{.3}^* = [1 \quad -2 \quad -1] \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} = 2 > 0; \quad B_0^{-1} A_{.4}^* = [1 \quad -2 \quad -1] \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} = 1 > 0. \quad (11.31)$$

The optimality conditions are satisfied and thus  $x_1^* = 1$  and  $x_2^* = 3$  solve the LP.

In the above example, it was easy to find a starting basic feasible solution so that we could start the simplex algorithm. However, finding a starting basic feasible solution may not be all that easy in a complicated problem. In fact, it may be the case that a given LP may not even have a feasible solution. Thus in the following section, we discuss Dantzig's Phase I procedure for finding a starting basic feasible solution.

## 11.5 Finding a Starting Basic Feasible Solution

Consider again the first form (11.1) of a linear programming problem,  $\min_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N \}$ . Now multiply both sides of each constraint equation through by  $-1$  if necessary so that the right hand side vector  $\mathbf{b}$  is nonnegative; i.e., so that  $\mathbf{b} \geq \mathbf{0}_M$ .

Now consider the following *Phase I* LP where we have introduced a vector of *artificial variables*  $\mathbf{s}$ :

$$\min_{\mathbf{x}, \mathbf{s}} \{ \mathbf{1}_M^T \mathbf{s} : \mathbf{A}\mathbf{x} + \mathbf{I}_M \mathbf{s} = \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N; \mathbf{s} \geq \mathbf{0}_M \} \quad (11.32)$$

where  $\mathbf{1}_M$  is an  $M$  dimensional column vector of ones and  $\mathbf{I}_M$  is an  $M \times M$  identity matrix. A basic feasible starting solution to (11.32) is  $\mathbf{s} = \mathbf{b} \geq \mathbf{0}_M$  and  $\mathbf{x} = \mathbf{0}_N$ . Note that this starting basis matrix is nonsingular. Now use the simplex algorithm to solve (11.32).<sup>\*6</sup>

Several cases can occur.

**Case 1:** The minimum for (11.32) is greater than 0.

In this case, there is no feasible solution for the original LP and we can stop at this point.

**Case 2:** The minimum for (11.32) equals 0.

In this case, the original LP has feasible solutions. However, we need to consider two subcases.

**Case 2A:** The minimum for (11.32) is 0 and no columns corresponding to the artificial variables are in the final basis matrix.

In this case, we have an  $\mathbf{x}^0 \geq \mathbf{0}_N$  such that  $\mathbf{A}\mathbf{x}^0 = \mathbf{b}$  and at most  $M$  components of  $\mathbf{x}^0$  are nonzero. Thus we have a starting basic feasible solution for the original LP (11.1).

**Case 2B:** The minimum for (11.32) is 0 and one or more columns corresponding to the artificial variables are in the final basis matrix.

In this case, we have a starting basic feasible solution for the following linear programming problem, which turns out to be equivalent to the original problem (11.1):

$$\max_{x_0, \mathbf{x} \geq \mathbf{0}_N, \mathbf{s} \geq \mathbf{0}_M} \{ x_0 : x_0 + \mathbf{c}^T \mathbf{x} = 0; \mathbf{A}\mathbf{x} + \mathbf{I}_M \mathbf{s} = \mathbf{b}; \mathbf{1}_M^T \mathbf{s} = 0 \}. \quad (11.33)$$

Note that (11.33) is the same as (11.1) except that the vector of artificial variables  $\mathbf{s}$  has been inserted into the constraint equations  $\mathbf{A}\mathbf{x} = \mathbf{b}$  and the extra constraint  $\mathbf{1}_M^T \mathbf{s} = 0$  has been added to the problem, along with the nonnegativity restrictions on  $\mathbf{s}$ ,  $\mathbf{s} \geq \mathbf{0}_M$ . However, note that the constraints  $\mathbf{1}_M^T \mathbf{s} = 0$  and  $\mathbf{s} \geq \mathbf{0}_M$  together imply that  $\mathbf{s} = \mathbf{0}_M$ . Thus the artificial variables will be kept at 0 levels for all iterations of the simplex algorithm applied to (11.33) and so solving (11.33) will also solve (11.1).

Thus solving the Phase I linear programming problem (11.32) will tell us whether a feasible solution to the original problem (11.1) exists or not and if a feasible solution to the original problem does exist, the solution to the Phase I problem will give us a starting basic feasible solution to solve the original linear programming problem (11.1), which is called the *Phase II problem* by Dantzig (1963)[65].

We conclude this section by solving the Phase I problem for the example in the previous section. Our Phase I problem is the following one:

$$\max_{x_0, x_1 \geq 0, x_2 \geq 0, s_1 \geq 0, s_2 \geq 0} \{ x_0 : \mathbf{e}_0 x_0 + A_{.1}^* x_1 + A_{.2}^* x_2 + A_{.3}^* x_3 + A_{.4}^* x_4 + A_{.5}^* s_1 + A_{.6}^* s_2 = \mathbf{b}^* \} \quad (11.34)$$

<sup>\*6</sup> Note that the unbounded solution case cannot occur for this Phase I problem because the objective function is bounded from below by 0.

where  $\mathbf{e}_0 \equiv \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ ,  $A_{\cdot 1}^* \equiv \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$ ,  $A_{\cdot 2}^* \equiv \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ ,  $A_{\cdot 3}^* \equiv \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}$ ,  $A_{\cdot 4}^* \equiv \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}$ ,  $A_{\cdot 5}^* \equiv \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ ,  $A_{\cdot 6}^* \equiv \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$  and  $\mathbf{b}^* \equiv \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$ .

Our starting basis matrix is  $\mathbf{B}$  defined as follows:

$$\mathbf{B} \equiv [\mathbf{e}_0, A_{\cdot 5}^*, A_{\cdot 6}^*] = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and so } \mathbf{B}^{-1} = \begin{bmatrix} 1 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (11.35)$$

Feasibility Check:

$$\begin{bmatrix} x_0^{(0)} \\ s_1^{(0)} \\ s_2^{(0)} \end{bmatrix} = \mathbf{B}^{-1} \mathbf{b}^* = \begin{bmatrix} 1 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -3 \\ 1 \\ 2 \end{bmatrix}. \quad (11.36)$$

Since  $s_1^{(0)} \geq 0$  and  $s_2^{(0)} \geq 0$ , the feasibility conditions are satisfied.

Optimality Check:

$$\left. \begin{array}{l} B_0^{-1} A_{\cdot 1}^* = 0; \quad \text{OK} \\ B_0^{-1} A_{\cdot 2}^* = -1; \quad \text{Can introduce } A_{\cdot 2}^* \text{ into the basis matrix} \\ B_0^{-1} A_{\cdot 3}^* = 1; \quad \text{OK} \\ B_0^{-1} A_{\cdot 4}^* = 1; \quad \text{OK} \\ B_0^{-1} A_{\cdot 5}^* = 0; \quad \text{OK (must equal 0 since } A_{\cdot 5}^* \text{ is in the initial basis matrix)} \\ B_0^{-1} A_{\cdot 6}^* = 0; \quad \text{OK (must equal 0 since } A_{\cdot 6}^* \text{ is in the initial basis matrix).} \end{array} \right\} \quad (11.37)$$

There is only one column  $A_{\cdot s}^*$  such that  $B_0^{-1} A_{\cdot s}^* < 0$ , namely  $A_{\cdot 2}^*$ . If there were two or more columns with  $B_0^{-1} A_{\cdot s}^* < 0$ , then pick the  $s$  such that  $B_0^{-1} A_{\cdot s}^*$  is the most negative. Thus column  $A_{\cdot 2}^*$  will enter the new basis matrix but which column will leave? We need to calculate:

Dropping Criterion:

$$\begin{aligned} x_2^{(1)} &\equiv \min_m \{ B_m^{-1} \mathbf{b}^* / B_m^{-1} A_{\cdot 2}^* : m \geq 1 \text{ and } m \text{ is such that } B_m^{-1} A_{\cdot 2}^* > 0 \} \\ &= B_2^{-1} \mathbf{b}^* / B_2^{-1} A_{\cdot 2}^* \\ &= 2/1 \\ &= 2 \end{aligned} \quad (11.38)$$

since  $B_1^{-1} A_{\cdot 2}^* = 0$ ,  $B_2^{-1} A_{\cdot 2}^* = 1$  and  $B_2^{-1} \mathbf{b}^* = 2$ . In other words, the minimum in (11.38) is attained for  $m = 2$  and the corresponding column of the original  $\mathbf{B}$  matrix leaves the basis matrix; i.e., the last column  $A_{\cdot 6}^*$  leaves. Thus the new basis matrix becomes:

$$\mathbf{B}^{(1)} \equiv [\mathbf{e}_0, A_{\cdot 5}^*, A_{\cdot 2}^*] = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and so } [\mathbf{B}^{(1)}]^{-1} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (11.39)$$

Feasibility Check:

$$\begin{bmatrix} x_0^{(1)} \\ s_1^{(1)} \\ x_2^{(1)} \end{bmatrix} = \mathbf{B}^{(1)-1} \mathbf{b}^* = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}. \quad (11.40)$$

Since  $s_1^{(1)} \geq 0$  and  $x_2^{(1)} \geq 0$ , the feasibility conditions are satisfied.

Optimality Check:

$$\left. \begin{array}{l} B_0^{(1)-1} A_{.1}^* = -1; \quad \text{Can introduce } A_{.1}^* \text{ into the basis matrix} \\ B_0^{(1)-1} A_{.2}^* = 0; \quad \text{OK (must equal 0 since } A_{.2}^* \text{ is in the basis matrix)} \\ B_0^{(1)-1} A_{.3}^* = 1; \quad \text{OK} \\ B_0^{(1)-1} A_{.4}^* = 0; \quad \text{OK} \\ B_0^{(1)-1} A_{.5}^* = 0; \quad \text{OK (must equal 0 since } A_{.5}^* \text{ is in the basis matrix)} \\ B_0^{(1)-1} A_{.6}^* = 1; \quad \text{OK.} \end{array} \right\} \quad (11.41)$$

There is only one column  $A_{.s}^*$  such that  $B_0^{(1)-1} A_{.s}^* < 0$ , namely  $A_{.1}^*$ . Thus column  $A_{.1}^*$  will enter the new basis matrix but which column will leave? We need to calculate:

Dropping Criterion:

$$\begin{aligned} x_1^{(2)} &\equiv \min_m \{ B_m^{(1)-1} \mathbf{b}^* / B_m^{(1)-1} A_{.1}^* : m \geq 1 \text{ and } m \text{ is such that } B_m^{(1)-1} A_{.1}^* > 0 \} \\ &= B_1^{(1)-1} \mathbf{b}^* / B_1^{(1)-1} A_{.1}^* \\ &= 1/1 \\ &= 1 \end{aligned} \quad (11.42)$$

since  $B_1^{(1)-1} A_{.1}^* = 1$ ,  $B_2^{(1)-1} A_{.1}^* = -1$  and  $B_1^{(1)-1} \mathbf{b}^* = 1$ . In other words, the minimum in (11.42) is attained for  $m = 1$  and the corresponding column of the  $\mathbf{B}^{(1)}$  matrix leaves the basis matrix; i.e., the second column  $A_{.5}^*$  leaves and  $A_{.1}^*$  enters. Thus the new basis matrix becomes:

$$\mathbf{B}^{(2)} \equiv [\mathbf{e}_0, A_{.1}^*, A_{.2}^*] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \text{ and so } [\mathbf{B}^{(2)}]^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}. \quad (11.43)$$

Feasibility Check:

$$\begin{bmatrix} x_0^{(2)} \\ x_1^{(2)} \\ x_2^{(2)} \end{bmatrix} = \mathbf{B}^{(2)-1} \mathbf{b}^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}. \quad (11.44)$$

Since  $x_1^{(2)} \geq 0$  and  $x_2^{(2)} \geq 0$ , the feasibility conditions are satisfied.

Optimality Check:

$$\left. \begin{aligned} B_{0\cdot}^{(2)-1} A_{\cdot 1}^* &= 0; && \text{OK (must equal 0 since } A_{\cdot 2}^* \text{ is in the basis matrix)} \\ B_{0\cdot}^{(2)-1} A_{\cdot 2}^* &= 0; && \text{OK (must equal 0 since } A_{\cdot 2}^* \text{ is in the basis matrix)} \\ B_{0\cdot}^{(2)-1} A_{\cdot 3}^* &= 0; && \text{OK} \\ B_{0\cdot}^{(2)-1} A_{\cdot 4}^* &= 0; && \text{OK} \\ B_{0\cdot}^{(2)-1} A_{\cdot 5}^* &= 1; && \text{OK} \\ B_{0\cdot}^{(2)-1} A_{\cdot 6}^* &= 1; && \text{OK.} \end{aligned} \right\} \quad (11.45)$$

Thus the optimality conditions for solving the Phase I problem are satisfied with  $A_{\cdot 1}^*$  and  $A_{\cdot 2}^*$  the final columns in the optimal basis matrix. Note that the objective function  $x_0$  increased at each iteration of the Phase I problem and we ended up with a basis matrix that we could use to start the Phase II problem. Note also that the basis matrix was invertible at each stage of the Phase I algorithm. In the following section, we will show that the basis matrix is always invertible, provided that we start with a basis matrix that is invertible, which we can always do. Thus our nonsingular basis matrix assumption made in section 11.2 is not a restrictive assumption.

## 11.6 Nonsingularity of the Basis Matrix in the Simplex Algorithm

Suppose that at some iteration of the simplex algorithm, the basis matrix is  $\mathbf{B} \equiv [e_0, A_{\cdot 1}^*, \dots, A_{\cdot M}^*]$  is nonsingular; i.e.,  $\mathbf{B}^{-1}$  exists. If  $B_{0\cdot}^{-1} A_{\cdot n}^* \geq 0$  for  $n = 1, 2, \dots, N$ , then we have an optimal solution and the algorithm stops. If  $B_{0\cdot}^{-1} A_{\cdot s}^* < 0$  for some  $s$  with  $B_{m\cdot}^{-1} A_{\cdot s}^* \leq 0$  for  $m = 1, 2, \dots, M$ , then we have an unbounded optimal solution and the algorithm stops. Now consider the remaining case; i.e., suppose that there exists a column index  $s$  such that

$$B_{0\cdot}^{-1} A_{\cdot s}^* < 0 \text{ and } B_{m\cdot}^{-1} A_{\cdot s}^* > 0 \text{ for at least one } m \geq 1. \quad (11.46)$$

Thus  $A_{\cdot s}^*$  enters the basis matrix and according to the dropping criterion,  $A_{\cdot r}^*$  leaves the basis matrix where  $r$  is such that  $B_{r\cdot}^{-1} \mathbf{b}^* / B_{r\cdot}^{-1} A_{\cdot s}^* = \min_m \{B_{m\cdot}^{-1} \mathbf{b}^* / B_{m\cdot}^{-1} A_{\cdot s}^* : m \geq 1 \text{ and } m \text{ is such that } B_{m\cdot}^{-1} A_{\cdot s}^* > 0\}$ . Thus we have:

$$B_{r\cdot}^{-1} A_{\cdot s}^* > 0. \quad (11.47)$$

Since the initial basis matrix  $\mathbf{B} \equiv [e_0, A_{\cdot 1}^*, \dots, A_{\cdot M}^*]$  is nonsingular, it must be possible to express the entering column  $A_{\cdot s}^*$  as a linear combination of the columns of  $\mathbf{B}$ ; i.e., there exists  $\boldsymbol{\alpha} \equiv [\alpha_0, \alpha_1, \dots, \alpha_M]^T$  such that

$$A_{\cdot s}^* = \mathbf{B}\boldsymbol{\alpha} \text{ or } \boldsymbol{\alpha} = \mathbf{B}^{-1} A_{\cdot s}^*. \quad (11.48)$$

Using (11.47) and (11.48), we have

$$\alpha_r = B_{r\cdot}^{-1} A_{\cdot s}^* > 0. \quad (11.49)$$

Now form the new basis matrix  $\mathbf{B}^{(1)}$  defined as follows:

$$\mathbf{B}^{(1)} \equiv [e_0, A_{\cdot 1}^*, A_{\cdot 2}^*, \dots, A_{\cdot r-1}^*, A_{\cdot s}^*, A_{\cdot r+1}^*, A_{\cdot r+2}^*, \dots, A_{\cdot M}^*]. \quad (11.50)$$

It is easy to see that (11.48) and  $\alpha_r > 0$  imply that  $A_{\cdot r}^*$  can be expressed as a linear combination of the columns of  $\mathbf{B}^{(1)}$ :

$$A_{\cdot r}^* = \mathbf{B}^{(1)}\boldsymbol{\beta}; \quad \boldsymbol{\beta} \equiv [\beta_0, \beta_1, \dots, \beta_M]^T \quad (11.51)$$

where

$$\beta_r = 1/\alpha_r = 1/B_r^{-1}A_{.s}^* \text{ and } \beta_m = -\alpha_m/\alpha_r = -B_m^{-1}A_{.s}^*/B_r^{-1}A_{.s}^* \text{ for } m \neq r. \quad (11.52)$$

Comparing  $\mathbf{B}$  with  $\mathbf{B}^{(1)}$ , it can be seen that

$$\begin{aligned} \mathbf{B} &\equiv [e_0, A_{.1}^*, \dots, A_{.r-1}^*, A_{.r}^*, A_{.r+1}^*, \dots, A_{.M}^*] \\ &= [e_0, A_{.1}^*, \dots, A_{.r-1}^*, A_{.s}^*, A_{.r+1}^*, \dots, A_{.M}^*][e_0, e_1, \dots, e_{r-1}, \boldsymbol{\beta}, e_{r+1}, \dots, e_M] \\ &= \mathbf{B}^{(1)}\mathbf{E} \end{aligned} \quad (11.53)$$

where  $\mathbf{E} \equiv [e_0, e_1, \dots, e_{r-1}, \boldsymbol{\beta}, e_{r+1}, \dots, e_M]$  and the components of  $\boldsymbol{\beta}$  are defined by (11.52). Note that since  $\beta_r = 1/\alpha_r > 0$ , the determinant of  $\mathbf{E}$  is equal to  $\beta_r > 0$  and hence  $\mathbf{E}^{-1}$  exists. Thus  $\mathbf{B}^{(1)} = \mathbf{B}\mathbf{E}^{-1}$  and

$$\mathbf{B}^{(1)-1} = \mathbf{E}\mathbf{B}^{-1} \quad (11.54)$$

and hence the inverse of the new basis matrix exists. Thus we have proven:

**Theorem 2** Dantzig (1963)[65]: The basis matrix is nonsingular at each iteration of the simplex algorithm.

Note that the inverse of the new basis matrix  $\mathbf{B}^{(1)}$  is easy to compute using (11.54): we need only pre-multiply the inverse of the old basis matrix  $\mathbf{B}^{-1}$  by the matrix  $\mathbf{E} \equiv [e_0, e_1, \dots, e_{r-1}, \boldsymbol{\beta}, e_{r+1}, \dots, e_M]$ , which differs from the identity matrix in only one column.\*7

There is one remaining detail to be filled in to complete our discussion of Dantzig's simplex algorithm for solving a linear program.

## 11.7 The Degeneracy Problem

Recall the nondegeneracy assumption made in section 11.3. Recall that degeneracy occurs when the dropping criterion does not yield a *unique* column to be dropped; i.e., when we compute the minimum in (11.24), the minimum is attained by more than one column index  $m$ . When this situation occurs, then it can happen at the next stage of the simplex algorithm that we do not get a drop in the objective function and two or more columns cycle in and out of the basis matrix without causing a drop in the objective function. Thus there is the theoretical possibility that the simplex algorithm could enter an infinite cycle and the algorithm might fail to converge.

It seems from an a priori point of view that it would be extremely unlikely that this cycling phenomenon could occur. However, Dantzig reports that the degeneracy phenomenon is common:

“It is common experience, based on the solutions of thousands of practical linear programming problems by the simplex method, that nearly every problem at some stage of the process is degenerate.” George Dantzig (1963; 231)[65].

When degeneracy occurs, it is conceivable that we could have a sequence of basis matrices where the objective function does not change such that the initial basis matrix reappears at some stage, which could then lead to an endless cycle of the same sequence of basis matrices. There are at least two artificially constructed examples of linear programming problems where this cycling phenomenon occurred; see Hoffman (1953)[240] and Beale (1955)[30]. Dantzig (1963; 228-230)[65] presents these two examples but he comments on the phenomenon as follows:

\*7 See (11.52) for the definition of the components of the  $\boldsymbol{\beta}$  vector.

“To date, there has not been one single case of circling, except in the specially constructed examples of Hoffman and Beale. Apparently, circling is a very rare phenomenon in practice. For this reason, most instruction codes for electronic computers use no special device for perturbing the problem to avoid degeneracy and the possibility of circling.” George Dantzig (1963; 231)[65].

It turns out that cycling or circling can be avoided if we try dropping a different sequence of columns after going through a cycle of column changes where the objective function did not change the first time around. Dantzig (1963)[65] develops various rules for choosing which column to drop which work and he has references to rules developed by others. For our purposes, the exact form of the dropping rule is not important; all we need to know is that a dropping rule exists such that the simplex algorithm will terminate in a finite number of iterations, even when nondegeneracy is not assumed.

We turn now to a topic of great economic interest.

## 11.8 The Dual Linear Program

In this section, we use the third formulation for a linear programming problem as our starting point. Thus we suppose that we are given as data an  $M \times N$  matrix  $\mathbf{A}$ , an  $N$  dimensional vector  $\mathbf{c}$  and an  $M$  dimensional vector  $\mathbf{b}$ . Then the *primal linear programming problem* is defined as the problem of maximizing the linear function  $\mathbf{c}^T \mathbf{x}$  with respect to the vector of primal decision variables  $\mathbf{x}$  subject to the inequality constraints  $\mathbf{A}\mathbf{x} \leq \mathbf{b}$  and the nonnegativity constraints  $\mathbf{x} \geq \mathbf{0}_N$ :

$$\text{Primal Problem: } \max_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N \}. \quad (11.55)$$

The *dual* to the above problem switches the roles of  $\mathbf{b}$  and  $\mathbf{c}$ , changes a maximization problem into a minimization problem, switches from a vector of primal variables  $\mathbf{x}$  that operate on the columns of  $\mathbf{A}$  to a vector of dual variables  $\mathbf{y}$  which operate on the rows of  $\mathbf{A}$  and writes the inequality constraints in the opposite direction:

$$\text{Dual Problem: } \min_{\mathbf{y}} \{ \mathbf{y}^T \mathbf{b} : \mathbf{y}^T \mathbf{A} \geq \mathbf{c}^T; \mathbf{y} \geq \mathbf{0}_M \}. \quad (11.56)$$

Note that the dual problem is also a linear programming problem. The primal and dual problems are related by the following theorem:

**Theorem 3** von Neumann (1947)[386], Dantzig (1963; chapter 6)[65]: If the primal problem has a finite optimal solution  $\mathbf{x}^*$  say, then the dual problem also has a finite optimal solution  $\mathbf{y}^*$  say and moreover:

$$\mathbf{c}^T \mathbf{x}^* = \mathbf{y}^{*T} \mathbf{b}; \quad (11.57)$$

i.e., the optimal values of the primal and dual objective functions are equal.

**Proof.** We first put the primal problem into the standard simplex algorithm format. Define  $x_0 \equiv \mathbf{c}^T \mathbf{x}$ . After adding the vector of slack variables  $\mathbf{s}$  to  $\mathbf{A}\mathbf{x}$ , we find that the primal problem (11.55) is equivalent to the following problem:

$$\max_{x_0, \mathbf{x} \geq \mathbf{0}_N, \mathbf{s} \geq \mathbf{0}_M} \{ x_0 : \begin{bmatrix} 1 \\ \mathbf{0}_M \end{bmatrix} x_0 + \begin{bmatrix} -\mathbf{c}^T & \mathbf{0}_M^T \\ \mathbf{A} & \mathbf{I}_M \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{s} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} \}. \quad (11.58)$$

Apply Phase I of the simplex algorithm to the LP defined by (11.58). Since we assume that the primal problem has a finite optimal solution, the optimal objective function for the Phase I problem

will attain its lower bound of 0. Thus we will have:

$$0 = \max_{x_0, \mathbf{x} \geq \mathbf{0}_N, \mathbf{s} \geq \mathbf{0}_M, \mathbf{z} \geq \mathbf{0}_M} \left\{ x_0 : \begin{bmatrix} 1 \\ \mathbf{0}_M \end{bmatrix} x_0 + \begin{bmatrix} \mathbf{0}_N^T & \mathbf{0}_M^T & \mathbf{1}_M^T \\ \mathbf{A} & \mathbf{I}_M & -\mathbf{I}_M \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{s} \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} \right\} \quad (11.59)$$

where we have added the vector of artificial variables  $\mathbf{z} \geq \mathbf{0}_M$  to the constraints of (11.58). There are two cases that can occur at this point.

*Case 1:* In this case, the Phase I problem yields a nonsingular initial basis matrix for the Phase II problem (11.58), which does not have any columns in it that correspond to the artificial variables  $\mathbf{z}$ . Thus in this case, we may apply the simplex algorithm to (11.58) without including any artificial variables and since by assumption, there is a finite optimal solution to (11.58), the simplex algorithm will terminate and there will exist an  $(M+1) \times (M+1)$  nonsingular basis matrix  $\mathbf{B}$  (where the first column of  $\mathbf{B}$  is the unit vector  $\mathbf{e}_0$ ) such that the following *optimality conditions* hold:

$$B_0^{-1} \begin{bmatrix} -\mathbf{c}^T & \mathbf{0}_M^T \\ \mathbf{A} & \mathbf{I}_M \end{bmatrix} \geq [\mathbf{0}_N^T \quad \mathbf{0}_M^T] \quad \text{and} \quad B_0^{-1} \mathbf{e}_0 = 1 \quad (11.60)$$

where  $B_0^{-1} \equiv [1, \mathbf{y}^{*T}]$  say, is the first row of  $\mathbf{B}^{-1}$ .<sup>\*8</sup> Thus the optimality conditions (11.60) imply that  $\mathbf{y}^*$  satisfies the following inequalities:

$$\mathbf{y}^{*T} \mathbf{A} \geq \mathbf{c}^T \quad \text{and} \quad \mathbf{y}^{*T} \mathbf{I}_M \geq \mathbf{0}_M^T. \quad (11.61)$$

Thus  $\mathbf{y}^*$  is a feasible solution for the dual problem (11.56).

At this point, we need the following *auxiliary result*: if  $\mathbf{x}^*$  and  $\mathbf{y}^*$  are feasible solutions for the primal and dual problems respectively, then we have the following bounds for the objective functions for the two problems:

$$\mathbf{c}^T \mathbf{x}^* \leq \mathbf{y}^{*T} \mathbf{b}; \quad (11.62)$$

i.e., the dual objective function evaluated at a feasible solution to the dual,  $\mathbf{y}^{*T} \mathbf{b}$ , is an upper bound for the primal objective function evaluated at any feasible solution and the primal objective function evaluated at a feasible solution for the primal,  $\mathbf{c}^T \mathbf{x}^*$ , is a lower bound for the dual objective function evaluated at any feasible solution. To prove (11.62), note that  $\mathbf{x}^*$  feasible for the primal problem and  $\mathbf{y}^*$  feasible for the dual problem means that  $\mathbf{x}^*$  and  $\mathbf{y}^*$  satisfy the following inequalities:

$$\mathbf{x}^* \geq \mathbf{0}_N; \quad \mathbf{A} \mathbf{x}^* \leq \mathbf{b}; \quad \mathbf{y}^* \geq \mathbf{0}_M; \quad \mathbf{y}^{*T} \mathbf{A} \geq \mathbf{c}^T. \quad (11.63)$$

Thus

$$\begin{aligned} \mathbf{c}^T \mathbf{x}^* &\leq \mathbf{c}^T \mathbf{x}^* + \mathbf{y}^{*T} (\mathbf{b} - \mathbf{A} \mathbf{x}^*) && \text{since } \mathbf{y}^* \geq \mathbf{0}_M \text{ and } \mathbf{A} \mathbf{x}^* \leq \mathbf{b} \\ &= \mathbf{y}^{*T} \mathbf{b} + (\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}) \mathbf{x}^* && \text{rearranging terms} \\ &\leq \mathbf{y}^{*T} \mathbf{b} && \text{since } \mathbf{x}^* \geq \mathbf{0}_N \text{ and } \mathbf{y}^{*T} \mathbf{A} \geq \mathbf{c}^T. \end{aligned} \quad (11.64)$$

Now (11.61) shows that the  $\mathbf{y}^*$  defined by  $B_0^{-1} \equiv [1, \mathbf{y}^{*T}]$  is a feasible solution for the dual. In view of (11.64), we need only show that  $\mathbf{y}^{*T} \mathbf{b} = \mathbf{c}^T \mathbf{x}^*$  where  $\mathbf{x}^*$  solves the primal, and we will have  $\mathbf{y}^*$  as an optimal solution to the dual problem.

Substitute the optimal Phase II solution to (11.58),  $x_0^*$ ,  $\mathbf{x}^*$  and  $\mathbf{s}^*$  into the constraints listed in (11.58) and then premultiply both sides of this matrix equation by  $B_0^{-1} \equiv [1, \mathbf{y}^{*T}]$ . We obtain the following equation:

$$[1, \mathbf{y}^{*T}] \begin{bmatrix} 1 \\ \mathbf{0}_M \end{bmatrix} x_0^* + [1, \mathbf{y}^{*T}] \begin{bmatrix} -\mathbf{c}^T & \mathbf{0}_M^T \\ \mathbf{A} & \mathbf{I}_M \end{bmatrix} \begin{bmatrix} \mathbf{x}^* \\ \mathbf{s}^* \end{bmatrix} = [1, \mathbf{y}^{*T}] \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix}. \quad (11.65)$$

<sup>\*8</sup> We can deduce that the first component of  $B_0^{-1}$  is 1 using  $B_0^{-1} \mathbf{e}_0 = 1$  since  $\mathbf{e}_0$  is the first column of  $\mathbf{B}$ .

Obviously,  $[1, \mathbf{y}^{*T}][1, \mathbf{0}_M^T]^T = 1$  and so the first term on the left hand side of (11.65) is simply equal to  $x_0^*$ . It turns out that the next set of terms on the left hand side of (11.65) is equal to 0 since if a particular column in the matrix  $\begin{bmatrix} -\mathbf{c}^T & \mathbf{0}_M^T \\ \mathbf{A} & \mathbf{I}_M \end{bmatrix}$  is in the optimal basis matrix, the inner product of this column with the first row of  $\mathbf{B}^{-1}$ ,  $[1, \mathbf{y}^{*T}]$ , will be  $0^{*9}$  and if a particular column in this matrix is not in the optimal basis matrix, then the corresponding  $x_n^*$  or  $s_m^*$  is zero and again the term will be 0. The term on the right hand side of (11.65) is equal to  $\mathbf{y}^{*T}\mathbf{b}$ . Thus (11.65) simplifies to:

$$x_0^* = \mathbf{y}^{*T}\mathbf{b}. \tag{11.66}$$

Since  $x_0^* = \mathbf{c}^T\mathbf{x}^*$ , we have  $\mathbf{c}^T\mathbf{x}^* = \mathbf{y}^{*T}\mathbf{b}$  and thus  $\mathbf{y}^*$  is indeed an optimal solution to the dual problem.

*Case 2:* In this case, the solution to the Phase I problem contains one or more columns corresponding to the artificial variables  $\mathbf{z}$ . In this case, the Phase I solution gives us a starting basic feasible solution to the following problem, which is equivalent to the primal problem (11.55):

$$\max_{x_0, \mathbf{x} \geq \mathbf{0}_N, \mathbf{s} \geq \mathbf{0}_M, \mathbf{z} \geq \mathbf{0}_M} \left\{ x_0 : \begin{bmatrix} 1 \\ \mathbf{0}_M \\ 0 \end{bmatrix} x_0 + \begin{bmatrix} -\mathbf{c}^T \\ \mathbf{A} \\ \mathbf{0}_N^T \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{0}_M^T \\ \mathbf{I}_M \\ \mathbf{0}_M^T \end{bmatrix} \mathbf{s} + \begin{bmatrix} \mathbf{0}_M^T \\ -\mathbf{I}_M \\ \mathbf{1}_M^T \end{bmatrix} \mathbf{z} = \begin{bmatrix} 0 \\ \mathbf{b} \\ 0 \end{bmatrix} \right\}. \tag{11.67}$$

We have set up the artificial variables in a slightly different way than was suggested in section 11.5 above. Since we already have added  $+\mathbf{I}_M \mathbf{s}$  to  $\mathbf{A}\mathbf{x}$ , we added  $-\mathbf{I}_M \mathbf{z}$  as the artificial variables. Between the  $\mathbf{s}$  and  $\mathbf{z}$  variables, we can obviously find a Phase I starting basic feasible solution to the constraints  $\mathbf{A}\mathbf{x} + \mathbf{I}_M\mathbf{s} - \mathbf{I}_M\mathbf{z} = \mathbf{b}$ , no matter what the signs are for the components of the  $\mathbf{b}$  vector. Now apply the Phase II simplex algorithm to (11.67). The constraints  $\mathbf{1}_M^T\mathbf{z} = 0$  and  $\mathbf{z} \geq \mathbf{0}_M$  will force the  $\mathbf{z}$  variables to be kept at 0 levels at all iterations of the Phase II algorithm. Since we assumed that a finite optimal solution to the original primal problem existed, the Phase II simplex algorithm will eventually terminate and there will exist an  $(M + 2) \times (M + 2)$  basis matrix  $\mathbf{C}$  such that the following optimality conditions will hold:

$$C_0^{-1} \begin{bmatrix} -\mathbf{c}^T & \mathbf{0}_M^T & \mathbf{0}_M^T \\ \mathbf{A} & \mathbf{I}_M & -\mathbf{I}_M \\ \mathbf{0}_N^T & \mathbf{0}_M^T & \mathbf{1}_M^T \end{bmatrix} \geq [\mathbf{0}_N^T, \mathbf{0}_M^T, \mathbf{0}_M^T] \tag{11.68}$$

where  $C_0^{-1} \equiv [1, \mathbf{y}^{*T}, y_0^*]$  is the first row of  $\mathbf{C}^{-1}$ . The optimal solution to (11.67),  $x_0^*, \mathbf{x}^*, \mathbf{s}^*, \mathbf{z}^*$  will satisfy the following relations:

$$x_0^* = \mathbf{c}^T\mathbf{x}^*; \mathbf{x}^* \geq \mathbf{0}_N; \mathbf{s}^* \geq \mathbf{0}_M; \mathbf{z}^* = \mathbf{0}_M. \tag{11.69}$$

The first two sets of inequalities in (11.68) imply that  $\mathbf{y}^*$  satisfies:

$$\mathbf{y}^{*T}\mathbf{A} \geq \mathbf{c}^T \text{ and } \mathbf{y}^* \geq \mathbf{0}_M \tag{11.70}$$

which implies that  $\mathbf{y}^*$  is feasible for the dual problem. Now evaluate the constraints in (11.67) at the optimal Phase II solution,  $x_0^*, \mathbf{x}^*, \mathbf{s}^*, \mathbf{z}^*$  and premultiply the constraint equations through by  $C_0^{-1} \equiv [1, \mathbf{y}^{*T}, y_0^*]$ . We find that as in Case 1:

$$x_0^* + 0 = \mathbf{y}^{*T}\mathbf{b} \tag{11.71}$$

Where the second term is 0 because columns not in the final Phase II basis matrix are multiplied by 0 while the columns in the final basis matrix have a 0 inner product with the first row of the inverse

---

<sup>\*9</sup> This follows from  $\mathbf{B}^{-1}\mathbf{B} = \mathbf{I}_{M+1}$ .

of the basis matrix,  $C_0^{-1} \equiv [1, \mathbf{y}^{*T}, y_0^*]$ . Since  $\mathbf{x}^*$  is feasible (and optimal) for the primal problem and  $\mathbf{y}^*$  is feasible for the dual problem, (11.69) and (11.71) and the auxiliary result imply that  $\mathbf{y}^*$  is a solution to the dual with  $\mathbf{y}^{*T} \mathbf{b} = \mathbf{c}^T \mathbf{x}^*$ . ■

**Corollary 1** If the primal problem has a feasible solution and the dual problem has a feasible solution, then both the primal and dual problems have finite optimal solutions and the values of the optimal objective functions coincide.

**Proof.** The auxiliary result (11.62) shows that if the primal and dual both have feasible solutions, then the primal objective function is bounded from above. Hence the unbounded solution case for the primal cannot occur. Now repeat the proof of Theorem 3. ■

Note that the above proof of the duality theorem for linear programs is entirely algebraic and rests on the mechanics of the simplex algorithm for solving linear programs. The proof of the above theorem is essentially due to Dantzig (1963)[65].

A further implication of the above duality theorem is that if the primal solution has an unbounded optimal solution, then the dual cannot have a feasible solution.

The duality theorem for linear programs has many economic applications as will be seen by studying the problems at the end of the chapter.

We conclude this section by proving another result that will be helpful in providing an economic interpretation for the dual variables.

**Theorem 4** *Basis Theorem for Linear Programs*; Dantzig (1963; 121)[65]: Suppose that a nonsingular basis matrix  $\mathbf{B}$  is *optimal* for the problem:

$$\max_{x_0, \mathbf{x}} \{x_0 : \mathbf{e}_0 x_0 + \mathbf{A}^* \mathbf{x} = \mathbf{b}^*; \mathbf{x} \geq \mathbf{0}_N\}. \quad (11.72)$$

Suppose further that this basis matrix  $\mathbf{B}$  generates a *feasible* solution for the following problem where the right hand side vector  $\mathbf{b}^*$  has been replaced by  $\mathbf{b}^1$ :

$$\max_{x_0, \mathbf{x}} \{x_0 : \mathbf{e}_0 x_0 + \mathbf{A}^* \mathbf{x} = \mathbf{b}^1; \mathbf{x} \geq \mathbf{0}_N\}. \quad (11.73)$$

Then the basis matrix  $\mathbf{B}$  is also optimal for the new problem (11.73).

**Proof.** Since  $\mathbf{B}$  is an optimal basis matrix for (11.72), then we have:

$$B_m^{-1} \mathbf{b}^* \geq 0 \text{ for } m = 1, 2, \dots, M \text{ (feasibility conditions satisfied);} \quad (11.74)$$

$$B_0^{-1} A_n^* \geq 0 \text{ for } n = 1, 2, \dots, N \text{ (optimality conditions satisfied).} \quad (11.75)$$

But by assumption, the initial optimal basis matrix  $\mathbf{B}$  generates a feasible solution for the new  $\mathbf{b}$  vector,  $\mathbf{b}^1$ , so that we have:

$$B_m^{-1} \mathbf{b}^1 \geq 0 \text{ for } m = 1, 2, \dots, M. \quad (11.76)$$

But since the basis matrix has not changed, the optimality conditions (11.75) are still satisfied (these conditions do not depend directly on  $\mathbf{b}$ ) and so (11.75) and (11.76) imply that the old basis matrix  $\mathbf{B}$  is still optimal for the new problem (11.73). ■

Theorem 4 is helpful in providing an economic interpretation for the dual variables. However, to see the connection, we have to consider a slightly different primal and dual LP problems compared to (11.55) and (11.56). Thus consider an equality constrained primal problem of the following form:

$$\max_{\mathbf{x}} \{\mathbf{c}^T \mathbf{x} : \mathbf{A} \mathbf{x} = \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N\}. \quad (11.77)$$

In order to find the dual problem to (11.77), we need to write the equality constraints  $\mathbf{A}\mathbf{x} = \mathbf{b}$  in an inequality format. Thus (11.77) is equivalent to the following problem:

$$\max_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \begin{bmatrix} \mathbf{A} \\ -\mathbf{A} \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} \mathbf{b} \\ -\mathbf{b} \end{bmatrix}; \mathbf{x} \geq \mathbf{0}_N \}. \quad (11.78)$$

Using our rules for forming the dual problem, it can be seen that the dual to (11.78) is:

$$\min_{\mathbf{y}^1, \mathbf{y}^2} \{ \mathbf{y}^{1T} \mathbf{b} - \mathbf{y}^{2T} \mathbf{b} : [\mathbf{y}^{1T}, \mathbf{y}^{2T}] \begin{bmatrix} \mathbf{A} \\ -\mathbf{A} \end{bmatrix} \geq \mathbf{c}^T; \mathbf{y}^1 \geq \mathbf{0}_M; \mathbf{y}^2 \geq \mathbf{0}_M \} \quad (11.79)$$

Define the unrestricted vector of variables  $\mathbf{y} \equiv \mathbf{y}^1 - \mathbf{y}^2$  and using this definition, we can rewrite the dual problem (11.79) as follows:

$$\min_{\mathbf{y}} \{ \mathbf{y}^T \mathbf{b} : \mathbf{y}^T \mathbf{A} \geq \mathbf{c}^T \}. \quad (11.80)$$

Thus (11.80) is the dual problem to the primal problem (11.77): *equality constraints in the primal lead to a dual problem with unrestricted in sign dual variables* (instead of nonnegative dual variables as before).

Assume that there is a finite optimal solution to (11.77). Putting (11.77) into the standard simplex algorithm format leads to the following equivalent problem:

$$\max_{x_0, \mathbf{x}} \{ x_0 : \begin{bmatrix} 1 \\ \mathbf{0}_M \end{bmatrix} x_0 + \begin{bmatrix} -\mathbf{c}^T \\ \mathbf{A} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix}; \mathbf{x} \geq \mathbf{0}_N \}. \quad (11.81)$$

Suppose that the final optimal basis matrix for (11.81) turns out to include the first  $M$  columns of the  $\mathbf{A}$  matrix. Thus we have

$$\mathbf{B} = \begin{bmatrix} 1 & -c_1 & \cdots & -c_M \\ \mathbf{0}_M & A_{.1} & \cdots & A_{.M} \end{bmatrix} \quad (11.82)$$

and we assume that  $\mathbf{B}$  is nonsingular. Our final assumption is that the final basis matrix  $\mathbf{B}$  for (11.81) is *nondegenerate* so that

$$x_m^* = B_{m\cdot}^{-1} \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} > 0 \text{ for } m = 1, 2, \dots, M. \quad (11.83)$$

We may repeat the proof of the Duality Theorem for linear programs in this set up. As before, we find that

$$B_{m\cdot}^{-1} \equiv [1, \mathbf{y}^{*T}] \quad (11.84)$$

where  $\mathbf{y}^*$  turns out to be a solution to the dual problem (11.80). The nondegeneracy assumptions (11.83) imply that  $\mathbf{y}^*$  must satisfy the following equations:

$$[1, \mathbf{y}^{*T}] \begin{bmatrix} -c_m \\ A_{.m} \end{bmatrix} = 0 \text{ for } m = 1, 2, \dots, M. \quad (11.85)$$

Equations (11.85) imply that the dual variables must be unique. Thus *nondegeneracy of the final basis matrix implies that the optimal solution to the dual problem is unique*.

Recall that the feasibility restrictions (11.83) are satisfied by our basis matrix, which we rewrite as follows:

$$B_{m\cdot}^{-1} \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} > 0 \text{ for } m = 1, 2, \dots, M. \quad (11.86)$$

The following optimality conditions are also satisfied by our basis matrix  $\mathbf{B}$ :

$$B_{0\cdot}^{-1} \begin{bmatrix} -c_n \\ A_{.n} \end{bmatrix} \geq 0 \text{ for } n = 1, 2, \dots, N. \quad (11.87)$$

Now we can apply the Basis Theorem to the above setup. If  $\mathbf{b}^1$  is sufficient close to the initial  $\mathbf{b}$ , the nonnegativity restrictions (11.86) will continue to hold using the old basis matrix  $\mathbf{B}$ ; i.e., we will have,  $\mathbf{b}^1$  close to  $\mathbf{b}$ :

$$B_m^{-1} \begin{bmatrix} 0 \\ \mathbf{b}^1 \end{bmatrix} > 0 \text{ for } m = 1, 2, \dots, M. \quad (11.88)$$

The optimality conditions (11.87) will continue to hold for this new primal problem where  $\mathbf{b}^1$  replaces  $\mathbf{b}$  and hence the basis matrix  $\mathbf{B}$  continues to be optimal for the new problem. Setting  $x_{M+1} = x_{M+2} = \dots = x_N = 0$  for the constraints in (11.81) and premultiplying both sides of the constraints by  $[1, \mathbf{y}^{*T}]$  where  $\mathbf{y}^*$  is defined by (11.84) leads to the following equation, where we have also used (11.85):

$$x_0 = \mathbf{y}^{*T} \mathbf{b}. \quad (11.89)$$

All of this shows that for  $\mathbf{b}^1$  close to  $\mathbf{b}$ , the optimal primal and dual objective functions, regarded as functions of  $\mathbf{b}^1$ , are given by

$$V(\mathbf{b}^1) = \mathbf{y}^{*T} \mathbf{b}^1. \quad (11.90)$$

Thus for  $\mathbf{b}^1 = \mathbf{b}$ , the vector of first order partial derivatives of the optimized primal and dual objective functions, with respect to the components of  $\mathbf{b}$  is equal to:

$$\nabla_{\mathbf{b}} V(\mathbf{b}) = \mathbf{y}^* \quad (11.91)$$

where the unique vector of dual variables  $\mathbf{y}^*$  is defined by (11.84). Thus  $y_m^*$  is the marginal increase in the primal (and dual) objective functions due to a small increase in the  $m$ th component of  $\mathbf{b}$ ,  $b_m$ , provided that we have a nondegenerate primal solution. Note that this interpretation for a dual variable is entirely analogous to the economic interpretation of a Lagrange multiplier in classical constrained optimization.\*<sup>10</sup>

## 11.9 The Geometric Interpretation of a Linear Program in Requirements Space

Recall the basic linear program defined by (11.2) above and recall that we rewrote in the form (11.10) above. For convenience, we repeat (11.10) below as problem (11.92):

$$\max_{x_0, x_1 \geq 0, x_2 \geq 0, \dots, x_N \geq 0} \{x_0 : \mathbf{e}_0 x_0 + \sum_{n=1}^N A_{.n}^* x_n = \mathbf{b}^*\} \quad (11.92)$$

where  $\mathbf{c}^T \equiv [c_1, \dots, c_N]$ ,  $\mathbf{A} \equiv [A_{.1}, A_{.2}, \dots, A_{.N}]$  is the original  $\mathbf{A}$  matrix and

$$A_{.n}^* \equiv \begin{bmatrix} c_n \\ A_{.n} \end{bmatrix} \text{ for } n = 1, \dots, N, \quad \mathbf{b}^* \equiv \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} \text{ and } \mathbf{e}_0 \equiv \begin{bmatrix} 1 \\ \mathbf{0}_M \end{bmatrix}. \quad (11.93)$$

Define the *super feasibility set*  $S$  as the following set in  $M + 1$  dimensional space:\*<sup>11</sup>

$$S \equiv \{z : z = \sum_{n=1}^N A_{.n}^* x_n; \mathbf{x}_1 \geq 0, \mathbf{x}_2 \geq 0, \dots, \mathbf{x}_N \geq 0\}. \quad (11.94)$$

\*<sup>10</sup> See Samuelson (1974; 65)[348] and Diewert (1984; 148)[99] for this interpretation for a Lagrange multiplier in classical (equality constrained) optimization theory. Dorfman, Samuelson and Solow (1958; 52-59)[160] have a nice discussion on the meaning and interpretation of dual variables. Problem 16 at the end of the chapter provides an economic interpretation for the dual prices in the general case where conditions (11.86) hold only as weak inequalities rather than as strict inequalities.

\*<sup>11</sup> Contrast this definition with the feasible set of  $\mathbf{x}$ 's, which is the set of  $\mathbf{x}$  such that  $\mathbf{x} \geq \mathbf{0}_N$  and  $\mathbf{A}\mathbf{x} = \mathbf{b}$ .

The set  $S$  is the set of all nonnegative linear combinations of the  $N$  column vectors  $A_n^*$ . Thus  $S$  is a set in  $M + 1$  dimensional space.

A set  $S$  is a *cone* if and only if it has the following property:

$$z \in S, \lambda \geq 0 \text{ implies } \lambda z \in S. \tag{11.95}$$

It is easy to verify that the super feasibility set  $S$  defined by (11.94) is a cone.\*<sup>12</sup> Note that this set  $S$  does not involve the right hand side vector  $\mathbf{b}$  which appeared in the original constraints for the LP,  $\mathbf{Ax} = \mathbf{b}$ . The vector  $\mathbf{b}$  is often called the *requirements vector*; i.e., a nonnegative vector  $\mathbf{x}$  must be chosen so that “production”  $\mathbf{Ax}$  meets the “requirements”  $\mathbf{b}$ . The *requirements line*  $L$  is defined as follows:

$$L \equiv \left\{ \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_M \end{bmatrix} : \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_M \end{bmatrix} = \begin{bmatrix} z_0 \\ b_1 \\ \vdots \\ b_M \end{bmatrix} \text{ where } z_0 \text{ is an unrestricted scalar variable} \right\}. \tag{11.96}$$

Note that points  $\mathbf{z}$  that belong to *both*  $L$  and  $S$  correspond to feasible solutions for the LP (11.92). An *optimal solution* for (11.92) will correspond to a point  $\mathbf{z}$  such that:

- $\mathbf{z} \in S$
- $\mathbf{z} \in L$  and
- The first component of  $\mathbf{z}$  is as small as possible subject to  $\mathbf{z} \in S$  and  $\mathbf{z} \in L$ .

Thus to solve (11.92), *we look for the lowest point* (with respect to the first component of  $\mathbf{z}$ ) *on the requirements line that also belong to the super feasibility set*  $S$ . It also turns out that any hyperplane that supports the set  $S$  (is tangent to  $S$ ) at this lowest point of  $L$  provides a set of optimal dual variables. An example will help to illustrate these points.

Suppose that we have 4 production units in a region or a country that uses varying amounts of two inputs and each production unit when run at unit scale produces output that is valued at 1 dollar. The input requirements of the 4 production units when run at unit scale are the  $A_n$  listed below:

$$A_{.1} \equiv \begin{bmatrix} 0 \\ 2 \end{bmatrix}; A_{.2} \equiv \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}; A_{.3} \equiv \begin{bmatrix} 1 \\ 1/2 \end{bmatrix}; A_{.4} \equiv \begin{bmatrix} 2 \\ 0 \end{bmatrix}. \tag{11.97}$$

Thus the first production unit uses 0 units of the first input and 2 units of the second input in order to produce outputs worth 1 dollar, the second production unit uses 1/2 of a unit of the first input and 1 unit of the second input in order to produce outputs worth 1 dollar and so on. The 4 production units are controlled by a single firm which can allocate  $b_1 > 0$  units of input 1 and  $b_2 > 0$  units of input 2 across the 4 production units. Production in each plant is subject to constant returns to scale. The firm is interested in allocating the amounts of the two inputs across the 4 plants in order to maximize revenue; i.e., the firm would like to solve the following linear programming problem:

$$\max_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{Ax} = \mathbf{b}; \mathbf{x} \geq \mathbf{0}_4 \} \tag{11.98}$$

where  $\mathbf{c}^T \equiv [1, 1, 1, 1]$ ,  $\mathbf{b}^T \equiv [b_1, b_2]$ ,  $\mathbf{A} \equiv [A_{.1}, A_{.2}, A_{.3}, A_{.4}]$  and  $\mathbf{x}^T \equiv [x_1, x_2, x_3, x_4]$  is the vector of nonnegative plant scales.

To put the problem defined by (11.97) into standard simplex algorithm format, define the  $A_n^*$  by adding  $-c_n$  as an extra component to the  $A_n$ :

$$A_{.1}^* \equiv \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}; A_{.2}^* \equiv \begin{bmatrix} -1 \\ 1/2 \\ 1 \end{bmatrix}; A_{.3}^* \equiv \begin{bmatrix} -1 \\ 1 \\ 1/2 \end{bmatrix}; A_{.4}^* \equiv \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix}; \mathbf{b}^* \equiv \begin{bmatrix} 0 \\ b_1 \\ b_2 \end{bmatrix}. \tag{11.99}$$

\*<sup>12</sup> It is also a convex set so it is a convex cone.

Using the above notation, the LP (11.98) can be rewritten as follows:

$$\max_{x_0, x \geq 0} \{x_0 : e_0 x_0 + A_1^* x_1 + A_2^* x_2 + A_3^* x_3 + A_4^* x_4 = \mathbf{b}^*\}. \tag{11.100}$$

Now recall the definition of the super feasibility set  $S$ , (11.94). For our example LP (11.100), this set will be a cone in three dimensional space. Now note that each of the 4 column vectors  $A_1^*, A_2^*, A_3^*$  and  $A_4^*$  will belong to  $S$  and the first component for each of these vectors is equal to  $-1$ . Hence if we intersect the cone  $S$  with the plane  $\{(z_0, z_1, z_2) : z_0 = -1, z_1 \text{ and } z_2 \text{ are unrestricted}\}$ , we obtain the shaded set in Figure 11.2 below.\*<sup>13</sup>

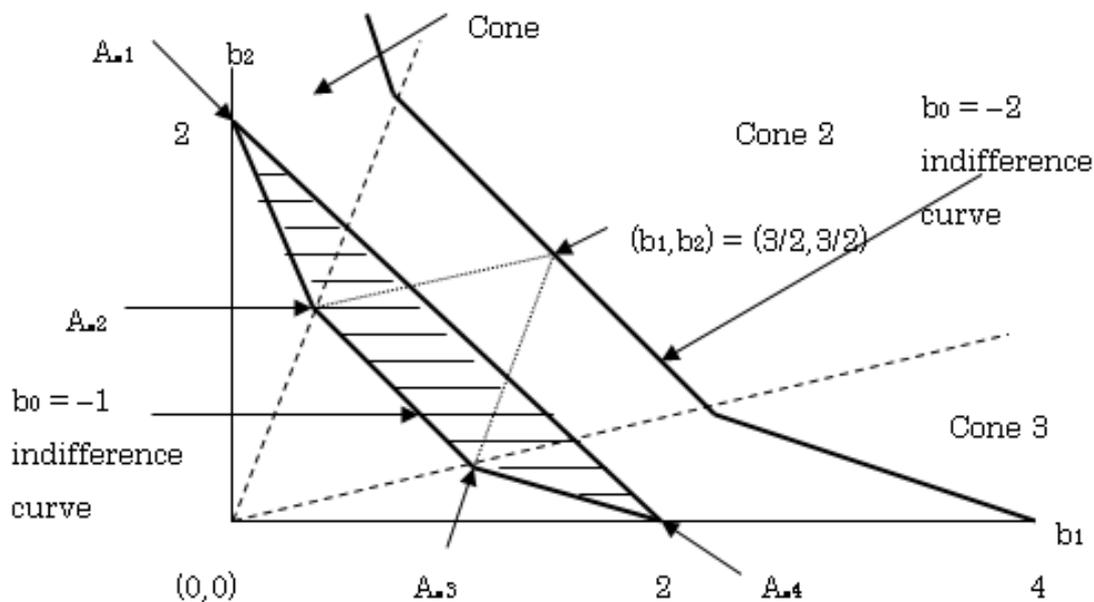


Fig. 11.2 The Simplex Algorithm in Requirements Space

The lower boundary of the shaded set is the counterpart to an indifference curve in consumer theory; this curve gives the set of  $(b_1, b_2)$  points which will allow the firm to earn one unit of revenue without wasting resources.\*<sup>14</sup> The point  $A_1$  is the top point on this curve and it is joined by a straight line to the point  $A_2$ , which in turn is joined to the point  $A_3$ , which in turn is joined to the lowest point on the curve,  $A_4$ . Now suppose that the firm has  $(b_1, b_2) = (3/2, 3/2)$  units of the two inputs available to use in the 4 production units. One can go through the algebra of the simplex algorithm in this case and one will find that  $A_2$  and  $A_3$  are the efficient plants to use in this case and the optimal  $x$ 's are  $x_1^* = 0, x_2^* = 1, x_3^* = 1$  and  $x_4^* = 0$ . The optimal objective function in this case is  $x_0^* = 2$ . In terms of Figure 11.1, it can be seen that we get to the point  $(b_1, b_2) = (3/2, 3/2)$  by moving up the dashed line through  $A_2$  until we hit  $A_2$  and by moving up the dashed line through  $A_3$  until we hit  $A_3$  and then we complete the parallelogram with the dotted lines until we hit  $(b_1, b_2) = (3/2, 3/2)$ . Note that the two dashed lines enclose a region of  $(b_1, b_2)$  that we called *cone 2*. Using the Basis Theorem for linear programs, it can be seen that we will use only production units 2 and 3 provided that  $(b_1, b_2)$  falls into this region. Using a similar argument, it can be seen only production units 1

\*<sup>13</sup> A three dimensional diagram would be more instructive but it is more difficult to graph. The reader should think of  $S$  as a 3 dimensional set in  $(b_0, b_1, b_2)$  space. What we are doing in Figure 11.2 is intersecting this three dimensional set with various (negative) heights or levels for  $b_0$  and then we plot these level sets in  $(b_1, b_2)$  space.

\*<sup>14</sup> In terms of the set  $S$ , this curve is the set of  $(b_1, b_2)$  points such that  $(-1, b_1, b_2)$  belongs to the boundary of  $S$ .

and 2 will be used if  $(b_1, b_2)$  falls into the *cone 1* region and only production units 3 and 4 will be used if  $(b_1, b_2)$  falls into the *cone 3* region.

Suppose that  $(b_1, b_2)$  falls into the interior of the cone 2 region. The corresponding (unique) dual (input) prices can be calculated by solving the following two equations in two unknowns:

$$[1, y_1, y_2]A_{.2}^* = 0; [1, y_1, y_2]A_{.3}^* = 0. \quad (11.101)$$

Using our data listed in (11.99), we find that the cone 2 dual solution is:

$$y_1^{(2)} = 2/3; y_2^{(2)} = 2/3. \quad (11.102)$$

Thus the input prices are equal in this case where the resource availability vector falls into the cone 2 region. This corresponds to the fact that the parallel “indifference” curves in this region all have slope equal to  $-1$  (which is equal to  $-y_1^{(2)}/y_2^{(2)}$ ).

Now suppose that  $(b_1, b_2)$  falls into the interior of the cone 1 region. The corresponding (unique) dual (input) prices can be calculated by solving the following two equations in two unknowns:

$$[1, y_1, y_2]A_{.1}^* = 0; [1, y_1, y_2]A_{.2}^* = 0. \quad (11.103)$$

Using our data listed in (11.99), we find that the cone 1 dual solution is:

$$y_1^{(1)} = 1; y_2^{(1)} = 1/2. \quad (11.104)$$

Thus the input 1 price is twice the size of the input 2 price in this region where input 1 is relatively scarce compared to input 2. This corresponds to the fact that the parallel “indifference” curves in this region all have slope equal to  $-2$  (which is equal to  $-y_1^{(1)}/y_2^{(1)}$ ).

Finally suppose that  $(b_1, b_2)$  falls into the interior of the cone 3 region. The corresponding (unique) dual (input) prices can be calculated by solving the following two equations in two unknowns:

$$[1, y_1, y_2]A_{.3}^* = 0; [1, y_1, y_2]A_{.4}^* = 0. \quad (11.105)$$

Using our data listed in (11.99), we find that the cone 3 dual solution is:

$$y_1^{(3)} = 1/2; y_2^{(3)} = 1. \quad (11.106)$$

Thus the input 2 price is twice the size of the input 1 price in this region where input 2 is relatively scarce compared to input 1. This corresponds to the fact that the parallel “indifference” curves in this region all have slope equal to  $-1/2$  (which is equal to  $-y_1^{(3)}/y_2^{(3)}$ ).

What happens if the endowment vector  $(b_1, b_2)$  happens to fall on the dashed line through the origin and  $A_{.2}$  so that  $(b_1, b_2)$  is on the boundary of both cone 1 and cone 2? In this case, although the primal solution is unique ( $x_2^*$  solves the primal where  $A_{.2}x_2^* = \mathbf{b}$ ), the dual solution is not. In this case the dual solution set is:

$$(y_1, y_2) = \lambda(y_1^{(1)}, y_2^{(1)}) + (1 - \lambda)(y_1^{(2)}, y_2^{(2)}); \quad 0 \leq \lambda \leq 1; \quad (11.107)$$

i.e., it is the set of all convex combinations of the cone 1 and cone 2 dual prices. Similarly, if the endowment vector  $(b_1, b_2)$  happens to fall on the dashed line through the origin and  $A_{.3}$  so that  $(b_1, b_2)$  is on the boundary of both cone 2 and cone 3, then in this case, although again the primal solution is unique ( $x_3^*$  solves the primal where  $A_{.3}x_3^* = \mathbf{b}$ ), the dual solution is not. In this case the dual solution set is:

$$(y_1, y_2) = \lambda(y_1^{(2)}, y_2^{(2)}) + (1 - \lambda)(y_1^{(3)}, y_2^{(3)}); \quad 0 \leq \lambda \leq 1; \quad (11.108)$$

i.e., it is the set of all convex combinations of the cone 2 and cone 3 dual prices.

The reader is now encouraged to visualize how the simplex algorithm would work if we had additional “inefficient” production units in the model. If our starting basis matrix had one or more inefficient columns in the basis matrix, then there would exist at least one more efficient column that would lie *below* the dual hyperplane that was generated by the starting columns. Introducing this more efficient column into the basis matrix would lead to a lower dual hyperplane at the next iteration and eventually, the dual hyperplane would lie on or below all of the columns in the model and the algorithm would come to a halt.

The geometry of the simplex algorithm in requirements space is illustrated in Dantzig (1963; 160-165)[65] and Van Slyke (1968)[381].<sup>\*15</sup> The representation of a linear programming problem in requirements space is a very useful one because we can simultaneously visualize the solution to the primal problem as well as to the dual problem. Also this requirements space geometric approach makes it easy to visualize the comparative statics effects of changing the right hand side vector  $\mathbf{b}$ .<sup>\*16</sup>

## 11.10 The Saddlepoint Criterion for Solving a Linear Program

In this section, we develop some criteria for solving the primal dual problems that do not rely on the simplex algorithm. However, the material in this section is very closely related to the auxiliary result that appeared in the proof of Theorem 3, the duality theorem for linear programs.

Consider the following primal and dual problems:

$$P: \max_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A} \mathbf{x} \leq \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N \}; \quad (11.109)$$

$$D: \min_{\mathbf{y}} \{ \mathbf{y}^T \mathbf{b} : \mathbf{y}^T \mathbf{A} \geq \mathbf{c}^T; \mathbf{y} \geq \mathbf{0}_M \} \quad (11.110)$$

where  $\mathbf{A}$  is an  $M \times N$  matrix,  $\mathbf{b}$  and  $\mathbf{y}$  are  $M$  dimensional vectors and  $\mathbf{c}$  and  $\mathbf{x}$  are  $N$  dimensional vectors.

Theorem 3 above, the Duality Theorem for Linear Programs, developed optimality criteria for  $\mathbf{x}^*$  to solve the primal problem (11.109) and for  $\mathbf{y}^*$  to solve the dual problem (11.110), using the mechanics of the simplex algorithm as a tool for establishing the criteria. In this section, we develop some alternative optimality criteria that do not rely on the simplex algorithm.

In order for  $\mathbf{x}^*$  to solve P and  $\mathbf{y}^*$  to solve D, it is obviously necessary that  $\mathbf{x}^*$  and  $\mathbf{y}^*$  satisfy the feasibility constraints for their respective problems:

$$\mathbf{A} \mathbf{x}^* \leq \mathbf{b}; \mathbf{x}^* \geq \mathbf{0}_N; \mathbf{y}^{*T} \mathbf{A} \geq \mathbf{c}^T; \mathbf{y}^* \geq \mathbf{0}_M. \quad (11.111)$$

Consider the following condition, which relates the primal and dual objective functions:

$$\mathbf{c}^T \mathbf{x}^* = \mathbf{y}^{*T} \mathbf{b}. \quad (11.112)$$

**Theorem 5** Goldman and Tucker (1956)[199]: Conditions (11.111) and (11.112) are necessary and sufficient for  $\mathbf{x}^*$  to solve P and for  $\mathbf{y}^*$  to solve D.

**Proof.** Recall the auxiliary result that was proven in Theorem 3 above, which may be restated as follows: if  $\mathbf{x}^*$  is feasible for the primal problem P and  $\mathbf{y}^*$  is feasible for the dual problem D, then

$$\begin{aligned} \mathbf{c}^T \mathbf{x}^* &\leq \mathbf{c}^T \mathbf{x}^* + \mathbf{y}^{*T} (\mathbf{b} - \mathbf{A} \mathbf{x}^*) && \text{since } \mathbf{y}^* \geq \mathbf{0}_M \text{ and } \mathbf{A} \mathbf{x}^* \leq \mathbf{b} \\ &= \mathbf{y}^{*T} \mathbf{b} + (\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}) \mathbf{x}^* && \text{rearranging terms} \\ &\leq \mathbf{y}^{*T} \mathbf{b} && \text{since } \mathbf{x}^* \geq \mathbf{0}_N \text{ and } \mathbf{y}^{*T} \mathbf{A} \geq \mathbf{c}^T. \end{aligned} \quad (11.113)$$

<sup>\*15</sup> Van Slyke was a student of Dantzig's.

<sup>\*16</sup> The Figure 11.1 geometry in activities space is useful for visualizing the comparative statics effects of changing the objective function vector  $\mathbf{c}$ .

Now suppose that  $\mathbf{x}^*$  and  $\mathbf{y}^*$  satisfy (11.111) and (11.112). Then repeating the proof of the auxiliary result in Theorem 3 shows that  $\mathbf{x}^*$  solves P and  $\mathbf{y}^*$  solves D.

Conversely, let  $\mathbf{x}^*$  solve P and let  $\mathbf{y}^*$  solve D. Then  $\mathbf{x}^*$  and  $\mathbf{y}^*$  satisfy (11.111) and (11.113) and so we have  $\mathbf{c}^T \mathbf{x}^* \leq \mathbf{y}^{*T} \mathbf{b}$ . Suppose  $\mathbf{c}^T \mathbf{x}^* < \mathbf{y}^{*T} \mathbf{b}$ , then we could repeat the proof of Theorem 3 and find a dual solution  $\mathbf{y}^{**}$  such that  $\mathbf{c}^T \mathbf{x}^* = \mathbf{y}^{**T} \mathbf{b} < \mathbf{y}^{*T} \mathbf{b}$ . But this last inequality contradicts the assumed optimality of  $\mathbf{y}^*$  for the dual problem D. Thus our *supposition* is false and  $\mathbf{c}^T \mathbf{x}^* = \mathbf{y}^{*T} \mathbf{b}$ , which is (11.112). ■

**Theorem 6** Goldman and Tucker (1956)[199]: Consider the following *complementary slackness conditions*:

$$\mathbf{y}^{*T} (\mathbf{b} - \mathbf{A}\mathbf{x}^*) = 0; \quad (11.114)$$

$$(\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}) \mathbf{x}^* = 0. \quad (11.115)$$

The feasibility conditions (11.111) and the complementary slackness conditions (11.114) and (11.115) are necessary and sufficient for  $\mathbf{x}^*$  to solve P and for  $\mathbf{y}^*$  to solve D.

**Proof.** By Theorem 5, we need only show that conditions (11.114) and (11.115) are equivalent to condition (11.112) when conditions (11.111) hold. But given that (11.111) holds, we know the inequality (11.113) holds. But now it can be seen that conditions (11.114) and (11.115) are precisely the conditions that are necessary and sufficient to convert the inequality (11.113) into the equality (11.112). ■

In order to state the next theorem, it is first necessary to define the *Lagrangian*  $L$  that corresponds to the primal problem P:

$$L(\mathbf{x}, \mathbf{y}) \equiv \mathbf{c}^T \mathbf{x} + \mathbf{y}^T [\mathbf{b} - \mathbf{A}\mathbf{x}]. \quad (11.116)$$

Note that the vector  $\mathbf{y}$  in (11.116) plays the role of a vector of *Lagrange multipliers* for the constraints in the primal problem P.

**Definition 3**  $(\mathbf{x}^*, \mathbf{y}^*)$  is a *saddle point* of the Lagrangian  $L$  if and only if

$$\mathbf{x}^* \geq \mathbf{0}_N; \mathbf{y}^* \geq \mathbf{0}_M \text{ and} \quad (11.117)$$

$$L(\mathbf{x}, \mathbf{y}^*) \leq L(\mathbf{x}^*, \mathbf{y}^*) \leq L(\mathbf{x}^*, \mathbf{y}) \text{ for all } \mathbf{x} \geq \mathbf{0}_N \text{ and } \mathbf{y} \geq \mathbf{0}_M. \quad (11.118)$$

Looking at conditions (11.118), we see that  $L(\mathbf{x}, \mathbf{y}^*)$  attains a *maximum* with respect to  $\mathbf{x}$  at  $\mathbf{x} = \mathbf{x}^*$  over all  $\mathbf{x} \geq \mathbf{0}_N$ . Conversely,  $L(\mathbf{x}^*, \mathbf{y})$  attains a *minimum* with respect to  $\mathbf{y}$  at  $\mathbf{y} = \mathbf{y}^*$  over all  $\mathbf{y} \geq \mathbf{0}_M$ .

**Theorem 7** Goldman and Tucker (1956; 77)[199]: The saddle point conditions (11.117) and (11.118) are necessary and sufficient for  $\mathbf{x}^*$  to solve P and for  $\mathbf{y}^*$  to solve D.

**Proof.** In view of Theorem 6, we need only show that (11.117) and (11.118) are equivalent to the feasibility restrictions (11.111) and the complementary slackness conditions (11.114) and (11.115).

Assume that (11.111), (11.114) and (11.115) hold. Conditions (11.111) imply that conditions (11.117) hold so we need only show that conditions (11.118) hold. Using the definition of  $L(\mathbf{x}, \mathbf{y})$ , conditions (11.118) can be rewritten as follows:

$$\mathbf{c}^T \mathbf{x} + \mathbf{y}^{*T} [\mathbf{b} - \mathbf{A}\mathbf{x}] \leq \mathbf{c}^T \mathbf{x}^* + \mathbf{y}^{*T} [\mathbf{b} - \mathbf{A}\mathbf{x}^*] \leq \mathbf{c}^T \mathbf{x}^* + \mathbf{y}^T [\mathbf{b} - \mathbf{A}\mathbf{x}^*] \quad \text{for all } \mathbf{x} \geq \mathbf{0}_N \text{ and } \mathbf{y} \geq \mathbf{0}_M. \quad (11.119)$$

Using (11.114), the right hand set of inequalities in (11.119) is equivalent to:

$$0 \leq \mathbf{y}^T [\mathbf{b} - \mathbf{A}\mathbf{x}^*] \text{ for all } \mathbf{y} \geq \mathbf{0}_M. \quad (11.120)$$

But the feasibility conditions (11.111) imply that  $\mathbf{b} - \mathbf{A}\mathbf{x}^* \geq \mathbf{0}_M$  so (11.120) must hold.

Using (11.115), the left hand set of inequalities in (11.119) is equivalent to:

$$[\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}] \mathbf{x} \leq [\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}] \mathbf{x}^* = 0 \text{ for all } \mathbf{x} \geq \mathbf{0}_N. \quad (11.121)$$

But the feasibility conditions (11.111) imply that  $\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A} \leq \mathbf{0}_N^T$  so that (11.121) holds. This completes the proof of the first half of the theorem.

Now assume that the saddle point conditions (11.117) and (11.118) hold. The right hand set of inequalities in (11.118) is equivalent to:

$$\mathbf{y}^{*T} [\mathbf{b} - \mathbf{A}\mathbf{x}^*] \leq \mathbf{y}^T [\mathbf{b} - \mathbf{A}\mathbf{x}^*] \text{ for all } \mathbf{y} \geq \mathbf{0}_M. \quad (11.122)$$

But it is easy to see that (11.122) implies that  $\mathbf{b} - \mathbf{A}\mathbf{x}^* \geq \mathbf{0}_M$ . In a similar manner, we can show that the left hand set of inequalities in (11.118) implies that  $\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A} \leq \mathbf{0}_N^T$ . Thus using also (11.117), we have shown that  $\mathbf{x}^*$  and  $\mathbf{y}^*$  satisfy the feasibility conditions (11.111).

We have shown that  $\mathbf{y}^* \geq \mathbf{0}_M$  and  $\mathbf{b} - \mathbf{A}\mathbf{x}^* \geq \mathbf{0}_M$ . Hence  $\mathbf{y}^{*T} [\mathbf{b} - \mathbf{A}\mathbf{x}^*] \geq 0$ . Suppose  $\mathbf{y}^{*T} [\mathbf{b} - \mathbf{A}\mathbf{x}^*] > 0$ . Now set  $\mathbf{y} = \mathbf{0}_M$  and (11.122) becomes  $0 < \mathbf{y}^{*T} [\mathbf{b} - \mathbf{A}\mathbf{x}^*] \leq 0$ , which is a contradiction. Hence our *supposition* is false and  $\mathbf{y}^{*T} [\mathbf{b} - \mathbf{A}\mathbf{x}^*] = 0$ , which is the complementary slackness condition (11.114).

Finally, we have shown that  $\mathbf{x}^* \geq \mathbf{0}_N$  and  $\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A} \leq \mathbf{0}_N^T$ . Hence  $(\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}) \mathbf{x}^* \leq 0$ . Note that the left hand set of inequalities in (11.118) or (11.119) is equivalent to:

$$[\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}] \mathbf{x} \leq [\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}] \mathbf{x}^* \text{ for all } \mathbf{x} \geq \mathbf{0}_N. \quad (11.123)$$

Now suppose  $(\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}) \mathbf{x}^* < 0$ . Set  $\mathbf{x} = \mathbf{0}_N$  and (11.123) becomes  $0 \leq (\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}) \mathbf{x}^* < 0$ , which is a contradiction. Hence our *supposition* is false and  $[\mathbf{c}^T - \mathbf{y}^{*T} \mathbf{A}] \mathbf{x}^* = 0$ , which is the complementary slackness condition (11.115). ■

The saddle point characterization for optimal solutions to the primal and dual linear programs turns out to be very useful as will be seen in the problems at the end of the chapter.

## 11.11 Programming with Variable Coefficients

Consider the following *programming problem with variable coefficients* which has the structure (11.10) of a standard LP with one exception:

$$\max_{x_0, x_1 \geq 0, x_2 \geq 0, \dots, x_N \geq 0, A_{\cdot 1}^*, \dots, A_{\cdot N}^*} \{x_0 : \mathbf{e}_0 x_0 + \sum_{n=1}^N A_{\cdot n}^* x_n = \mathbf{b}^*; A_{\cdot 1}^* \in C^1, \dots, A_{\cdot N}^* \in C^N\}. \quad (11.124)$$

The only difference between (11.124) and our standard simplex algorithm primal LP problem (11.10) is that in (11.10), the columns  $A_{\cdot n}$  were fixed but now columns are variable; i.e.,  $A_{\cdot n}^*$  may be chosen from a closed and bounded set<sup>\*17</sup> of columns  $C^n$ . Obviously, if each column set  $C^n$  has only a single member in it, then (11.124) boils down to our standard LP (11.10). Dantzig (1963; chapter 22)[65] considered problems of the form (11.124) where the sets  $C^n$  consisted of convex combinations of a finite set of columns. Van Slyke (1968)[381] considered general problems of the type (11.124).

<sup>\*17</sup> These restrictions on the sets  $C^n$  can be relaxed under certain additional hypotheses. It is sometimes assumed that the sets  $C^n$  are convex, but this additional assumption is not necessary for the simplex algorithm to work in this context.

Consider now how to solve a problem of the type (11.124). Suppose that  $A_{.1}^* \in C^1, \dots, A_{.M}^* \in C^M$  exist such that  $\mathbf{B} \equiv [\mathbf{e}_0, A_{.1}^*, \dots, A_{.M}^*]$  is an initial basis matrix for (11.124); i.e.,  $\mathbf{B}^{-1}$  exists and the last  $M$  components of the  $M+1$  vector  $\mathbf{B}^{-1}\mathbf{b}^*$  are nonnegative. Consider the following  $N$  subproblems:

$$f_n(\mathbf{B}) \equiv \min_{A_{.n}^*} \{B_{0.}^{-1}A_{.n} : A_{.n} \in C^n\}; \quad n = 1, 2, \dots, N. \quad (11.125)$$

Assuming that we can solve the subproblems (11.125) (a nontrivial assumption), then the variable coefficients simplex algorithm works as follows:

- (i) If  $f_n(\mathbf{B}) \geq 0$  for  $n = 1, 2, \dots, N$ , then the present basis matrix is optimal and the algorithm terminates.
- (ii) If  $f_n(\mathbf{B}) < 0$  for some  $n$ , then there exists  $A_{.n}^* \in C^n$  such that  $f_n(\mathbf{B}) = B_{0.}^{-1}A_{.n}^* < 0$ . Introduce  $A_{.n}^*$  into the basis matrix using the usual unbounded solutions criterion or the dropping criterion. If we end up in the unbounded solutions criterion case, then the algorithm terminates. Otherwise, we solve the following LP:

$$\max_{x_0, x_1 \geq 0, x_2 \geq 0, \dots, x_M \geq 0, x_n \geq 0} \{x_0 : \mathbf{e}_0 x_0 + \sum_{m=1}^M A_{.m}^* x_m + A_{.n}^* x_n = \mathbf{b}^*\} \quad (11.126)$$

and we obtain a new basis matrix. We use this new basis matrix to solve the subproblems (11.126) and we return to (i) above and repeat the cycle.

We can no longer assert that the simplex algorithm will converge to a solution in a finite number of steps. However, Van Slyke (1968)[381] proved that the algorithm will converge to an optimal solution of (11.124) under relatively weak regularity conditions.

The above algorithm has been used in the theory of economic planning; see Dorfman, Samuelson and Solow (1958; 59-63)[160] and Malinvaud (1967)[300]. The basic idea is the following one. The central planner has a preliminary vector of resource prices  $[1, \mathbf{y}^{*T}]$ , which is transmitted to the various sectors in the economy. Given these prices, the manager of sector  $n$  solves a (net) cost minimization problem similar to (11.126), where  $B_{0.}^{-1}$  is replaced by  $[1, \mathbf{y}^{*T}]$ . Each manager then sends back a solution vector  $A_{.n}^*$  back to the center. The central planner then solves a simple linear program involving this sectoral data, obtains a new set of prices, transmits this new set of prices to the sectors and so on.\*<sup>18</sup>

The variable coefficients simplex algorithm does depend on the ease with which the subproblems (11.126) can be solved. If the sets  $C^n$  are unit scale production possibilities sets (these sets are dual to a unit cost function or to a unit profit function) and if the dual unit cost or profit functions are available to the programmer, then the subproblems (11.126) can be solved by a simple application of Shephard's (1953)[356] Lemma or Hotelling's (1932)[242] Lemma. In this case, the variable coefficients simplex algorithm is just as easy to use as the ordinary simplex algorithm. Some applications along these lines may be found in Diewert (1975)[81], Diewert and Woodland (1977)[158] and Woodland (1982)[405].

**Problem 1** Consider the following linear program:

$$\max_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A}^{(1)} \mathbf{x} \leq \mathbf{b}^{(1)}; \mathbf{A}^{(2)} \mathbf{x} = \mathbf{b}^{(2)}; \mathbf{x} \geq \mathbf{0}_N \}. \quad (i)$$

\*<sup>18</sup> Why has this theory of decentralized economic planning not worked in the real world? The assumption of constant returns to scale is somewhat problematic in this theory but probably the main reason it has not worked very well in applications is the *problem of dimensionality*; i.e., real life economic planning problems at the national scale involve thousands if not millions of variables and the resulting programming problems are just too difficult to solve. However, at the level of a firm which has several more or less independent divisions, the above theory of decentralized economic planning could work.

Show that the dual to (i) is:

$$\min_{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}} \{ \mathbf{y}^{(1)T} \mathbf{b}^{(1)} + \mathbf{y}^{(2)T} \mathbf{b}^{(2)} : \mathbf{y}^{(1)T} \mathbf{A}^{(1)} + \mathbf{y}^{(2)T} \mathbf{A}^{(2)} \geq \mathbf{c}^T; \mathbf{y}^{(1)} \geq \mathbf{0}_{M_1}; \mathbf{y}^{(2)} \text{ unrestricted} \}. \quad (\text{ii})$$

*Hint:* Write  $\mathbf{A}^{(2)} \mathbf{x} = \mathbf{b}^{(2)}$  as  $\mathbf{A}^{(2)} \mathbf{x} \leq \mathbf{b}^{(2)}$  and  $-\mathbf{A}^{(2)} \mathbf{x} \leq -\mathbf{b}^{(2)}$ .

*Comment:* If  $M_1 = 0$ , then the primal problem has equality constraints and the corresponding dual variables are unrestricted.

**Problem 2** Consider the following linear program:

$$\max_{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}} \{ \mathbf{c}^{(1)T} \mathbf{x}^{(1)} + \mathbf{c}^{(2)T} \mathbf{x}^{(2)} : \mathbf{A}^{(1)} \mathbf{x}^{(1)} + \mathbf{A}^{(2)} \mathbf{x}^{(2)} \leq \mathbf{b}; \mathbf{x}^{(1)} \geq \mathbf{0}_{N_1}; \mathbf{x}^{(2)} \text{ unrestricted} \}. \quad (\text{i})$$

Show that the dual to (i) is

$$\min_{\mathbf{y}} \{ \mathbf{y}^T \mathbf{b} : \mathbf{y}^T \mathbf{A}^{(1)} \geq \mathbf{c}^{(1)T}; \mathbf{y}^T \mathbf{A}^{(2)} = \mathbf{c}^{(2)T}; \mathbf{y} \geq \mathbf{0}_M \}. \quad (\text{ii})$$

**Problem 3** Consider a firm that has 4 production processes (or activities) that can be used to produce two outputs.<sup>\*19</sup> Each process uses one primary input (call it labour) and produces one of the two outputs. Some processes use the other output as an intermediate input. The firm wishes to minimize the aggregate labour cost of producing given positive amounts,  $b_1 > 0$  and  $b_2 > 0$ , of the two outputs. The linear program that the firm wishes to solve is:

$$\min_{x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0} \{ 2x_1 + 1x_2 + 1x_3 + 2x_4 : \begin{bmatrix} 1 & 1 & -2 & 0 \\ 0 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \geq \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \}. \quad (\text{i})$$

The first two activities produce the finally demanded commodity  $b_1$  while the last two activities produce the finally demanded commodity  $b_2$ .

(a) Solve the above LP when  $b_1 = 1$  and  $b_2 = 1$ .

*Hint:* rewrite (i) in standard simplex algorithm format as follows:

$$\max_{x_0, x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0, s_1 \geq 0, s_2 \geq 0} \{ x_0 : \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} x_0 + \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix} x_1 + \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} x_2 + \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} x_3 + \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix} x_4 + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} s_1 + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} s_2 = \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix} \}. \quad (\text{ii})$$

Note that the  $x_1$  and  $x_4$  columns can be used in a starting basis matrix.

The dual problem to (i) is:

$$\max_{y_1 \geq 0, y_2 \geq 0} \{ y_1 b_1 + y_2 b_2 : [y_1, y_2]^T \begin{bmatrix} 1 & 1 & -2 & 0 \\ 0 & -1 & 1 & 1 \end{bmatrix} \leq [2, 1, 1, 2] \}. \quad (\text{iii})$$

(b) Find the dual prices  $y_1^*$  and  $y_2^*$  for problem (iii) when  $b_1 = 1$  and  $b_2 = 1$ .

<sup>\*19</sup> Each process can be thought of as a Leontief production function that does not allow for substitution between inputs. Leontief (1936)[289] (1941)[290] was a pioneer in setting up production models of entire economies using Leontief type production functions. His influence lives on today: many national statistical agencies produce input-output tables based on his model of production.

*Hint:* You need only list the last two components of  $B_0^{-1}$  where  $\mathbf{B}$  is the optimal basis matrix for (ii).

(c) Finally, solve the LP (i) for a general  $b_1 > 0$  and  $b_2 > 0$  and show that the corresponding dual prices do not depend on  $b_1$  and  $b_2$ .

*Comment:* This problem provides a concrete illustration of the Samuelson (1951)[341] Nonsubstitution Theorem; i.e., the optimal activities do not depend on the particular vector of final demands  $[b_1, b_2]$  that we are required to produce in this single input (labour theory of value) economy and neither do the dual prices, which can be interpreted as equilibrium prices for the outputs.

**Problem 4** Consider the following *zero sum game theory problem*, where we list the payoff matrix of the student:

		<i>Student's Strategy Space</i>	
		Go to lecture	Sleep in
<i>Professor's Strategy Space</i>	Give a good lecture	2	1
	Give a terrible lecture	0	3

We suppose that the student wishes to choose a strategy, which consists of the probability of going to a lecture,  $p_1 \geq 0$ , and the probability of sleeping in,  $p_2 \geq 0$ , where  $p_1 + p_2 = 1$ . We further suppose that the somewhat psychotic student suspects that the professor is perversely attempting to minimize the student's payoff. Thus the student would be wise to choose a strategy which would ensure him or her a minimum amount of utility irrespective of the professor's actions; this is called a *minimax strategy*. In order to find the minimax strategy, the student will want to solve the following linear program:

$$\max_{v \geq 0, p_1 \geq 0, p_2 \geq 0} \{v : v \leq 2p_1 + 1p_1; v \leq 0p_1 + 3p_1; 1 = p_1 + p_2\}. \tag{i}$$

Since the student's payoff matrix is nonnegative, we can define new variables  $x_1$  and  $x_2$  which will also be nonnegative as follows:

$$x_1 \equiv p_1/v; \quad x_2 \equiv p_2/v. \tag{ii}$$

Note that  $p_1 + p_2 = vx_1 + vx_2 = 1$  so  $v$  can be expressed in terms of  $x_1$  and  $x_2$  as follows:

$$v = 1/(x_1 + x_2). \tag{iii}$$

Thus maximizing  $v$  is equivalent to minimizing  $x_1 + x_2$  and we find that the LP (i) is equivalent to the following LP:

$$\min_{x_1 \geq 0, x_2 \geq 0} \{x_1 + x_2 : 2x_1 + 1x_2 \geq 1; 0x_1 + 3x_2 \geq 1\}. \tag{iv}$$

- (a) Solve the LP (iv) and use (ii) and (iii) to calculate the student's optimal strategy,  $p_1^*$  and  $p_2^*$ .
- (b) What probability mix of good and terrible lectures should the professor use to ensure a minimum payoff to the student?

*Hint:* We can transform the professor's minimization problem into the dual of the LP (iv).

*Comment:* This problem provides an example of the *minimax theorem* for zero sum two person games.

*Reference:* Dorfman, Samuelson and Solow (1958; chapters 15-16)[160].

**Problem 5** *Complementary Slackness in Linear Programming* ; Dantzig (1963; 136)[65]: Consider the following primal LP:

$$\max_{\mathbf{x} \geq \mathbf{0}_N} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A} \mathbf{x} \leq \mathbf{b} \} \text{ where } \mathbf{A} \text{ is an } M \times N \text{ matrix.} \tag{i}$$

Suppose a finite optimal solution  $\mathbf{x}^0 \geq \mathbf{0}_N$  exists for the above problem. Then by the *duality theorem for linear programs*, an optimal vector of dual variables  $\mathbf{y}^{0T} \equiv [y_1^0, \dots, y_M^0] \geq \mathbf{0}_M^T$  will also exist.

(a) Show that if the  $m$ th constraint in the primal LP is satisfied with a strict inequality when evaluated at the primal optimal solution  $\mathbf{x}^0$  so that  $A_m \mathbf{x}^0 < b_m$ , then the corresponding dual variable  $y_m^0$  must equal 0.

*Hint:* If the  $m$ th constraint is satisfied with a strict inequality, then the corresponding slack variable must be in the optimal basis matrix.

(b) Give an economic interpretation of the above result.

(c) Prove that  $\mathbf{y}^{0T}[\mathbf{b} - \mathbf{A}\mathbf{x}^0] = 0$  and  $[\mathbf{c}^T - \mathbf{y}^{0T}\mathbf{A}]\mathbf{x}^0 = 0$ .

**Problem 6** *Farkas (1902)[175] Lemma:* Let  $\mathbf{A}$  be a nonzero  $M \times N$  matrix and  $\mathbf{c}$  be an  $N$  dimensional vector. Then, *either* there exists an  $\mathbf{x}$  such that

$$\mathbf{A}\mathbf{x} \leq \mathbf{0}_M \text{ and } \mathbf{c}^T \mathbf{x} > 0 \quad (\text{i})$$

or there exists a  $\mathbf{y}$  such that

$$\mathbf{y}^T \mathbf{A} = \mathbf{c}^T \text{ and } \mathbf{y} \geq \mathbf{0}_M. \quad *20 \quad (\text{ii})$$

Prove the above result.

*Hint:* Consider the following primal and dual problems:

$$\max_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{0}_M \}; \quad (\text{iii})$$

$$\min_{\mathbf{y}} \{ \mathbf{y}^T \mathbf{0}_M : \mathbf{y}^T \mathbf{A} = \mathbf{c}^T; \mathbf{y} \geq \mathbf{0}_M \}. \quad (\text{iv})$$

Using the results of Problems 1 and 2 above, it can be seen that (iv) is the dual to (iii). Since  $\mathbf{x} = \mathbf{0}_N$  is feasible for (iii), there are only two cases that can occur for (iii): *Case 1:* the optimal objective function for (iii) becomes unbounded and *Case 2:* (iii) has a finite optimal solution. Show that Case 1 corresponds to (i) and Case 2 corresponds to (ii).

*Comment:* Farkas' Lemma can be used to prove Theorem 3, the duality theorem for linear programs (this is the path followed by Tucker (1956)[376]) and as the above problem shows, the duality theorem established via the simplex algorithm can be used to prove Farkas' Lemma (this is the path followed by Dantzig (1963; 145)[65]). Mangasarian (1969; chapter 2)[302] calls Farkas' Lemma a *theorem of the alternative* and he lists several other similar theorems, which we will give as problems 7-9 below. These theorems of the alternative turn out to be very useful in the theory of nonlinear programming and in the theory of economic policy. In particular, theorems of the alternative can be used to establish the existence of Pareto improving changes in economic policy instruments. For examples of these applications to the theory of economic policy, see Diewert (1978; 281)[87] (1983)[98] (1987)[100], Weymark (1979)[397] and Diewert, Turunen-Red and Woodland (1989)[151] (1991)[152].

**Problem 7** *Motzkin's (1936)[315] Transposition Theorem;* see Mangasarian (1969; 28-29)[302]: Let  $\mathbf{E}$  be an  $M_1 \times N$  matrix,  $\mathbf{F}$  be an  $M_2 \times N$  matrix and  $\mathbf{G}$  be an  $M_3 \times N$  matrix where  $M_1 > 0, N > 0, M_2 \geq 0$  and  $M_3 \geq 0$ .<sup>\*21</sup> Then *either* there exists  $\mathbf{x}$  such that

$$\mathbf{E}\mathbf{x} \gg \mathbf{0}_{M_1}; \mathbf{F}\mathbf{x} \geq \mathbf{0}_{M_2}; \mathbf{G}\mathbf{x} = \mathbf{0}_{M_3} \quad (\text{i})$$

or there exist  $\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3$  such that

$$\mathbf{y}^{1T}\mathbf{E} + \mathbf{y}^{2T}\mathbf{F} + \mathbf{y}^{3T}\mathbf{G} = \mathbf{0}_N^T; \mathbf{y}^1 > \mathbf{0}_{M_1}; \mathbf{y}^2 \geq \mathbf{0}_{M_2}; \mathbf{y}^3 \text{ unrestricted.} \quad (\text{ii})$$

<sup>\*20</sup> Both (i) and (ii) cannot hold.

<sup>\*21</sup> If  $M_2 = 0$ , then drop the matrix  $\mathbf{F}$  from the problem; if  $M_3 = 0$ , then drop the matrix  $\mathbf{G}$  from the problem.

*Hint:* Consider the following LP which is closely related to (i):

$$\max_{\mathbf{x}} \{ \mathbf{0}_N^T \mathbf{x} : -\mathbf{E}\mathbf{x} \leq -\mathbf{1}_{M_1}; -\mathbf{F}\mathbf{x} \leq \mathbf{0}_{M_2}; -\mathbf{G}\mathbf{x} = \mathbf{0}_{M_3}; \mathbf{x} \text{ unrestricted} \}. \quad (\text{iii})$$

The dual to (iii) turns out to be the following LP, which is closely related to (ii):

$$\min_{\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3} \{ -\mathbf{y}^{1T} \mathbf{1}_{M_1} : \mathbf{y}^{1T} \mathbf{E} + \mathbf{y}^{2T} \mathbf{F} + \mathbf{y}^{3T} \mathbf{G} = \mathbf{0}_N^T; \mathbf{y}^1 \geq \mathbf{0}_{M_1}; \mathbf{y}^2 \geq \mathbf{0}_{M_2}; \mathbf{y}^3 \text{ unrestricted} \}. \quad (\text{iv})$$

Note that the objective function for (iii) is bounded. Hence, there are two cases that can occur for (iii): Case 1: a feasible solution for (iii) exists and hence we have a bounded optimal solution for (iii) or Case 2: no feasible solution for (iii) exists.

**Problem 8** *Tucker's* (1956; 14)[376] *Theorem of the Alternative*; see Mangasarian (1969; 29)[302]: Let  $\mathbf{B}$  be an  $M_1 \times N$  matrix,  $\mathbf{C}$  be an  $M_2 \times N$  matrix and  $\mathbf{D}$  be an  $M_3 \times N$  matrix where  $M_1 > 0, N > 0, M_2 \geq 0$  and  $M_3 \geq 0$ . Then *either* there exists  $\mathbf{x}$  such that

$$\mathbf{B}\mathbf{x} > \mathbf{0}_{M_1}; \mathbf{C}\mathbf{x} \geq \mathbf{0}_{M_2}; \mathbf{D}\mathbf{x} = \mathbf{0}_{M_3} \quad (\text{i})$$

or there exist  $\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3$  such that

$$\mathbf{y}^{1T} \mathbf{B} + \mathbf{y}^{2T} \mathbf{C} + \mathbf{y}^{3T} \mathbf{D} = \mathbf{0}_N^T; \mathbf{y}^1 \gg \mathbf{0}_{M_1}; \mathbf{y}^2 \geq \mathbf{0}_{M_2}; \mathbf{y}^3 \text{ unrestricted}. \quad (\text{ii})$$

*Hint:* Consider the following LP which is closely related to (i):

$$\max_{\mathbf{x}} \{ \mathbf{1}_{M_1}^T \mathbf{B}\mathbf{x} : -\mathbf{B}\mathbf{x} \leq \mathbf{0}_{M_1}; -\mathbf{C}\mathbf{x} \leq \mathbf{0}_{M_2}; -\mathbf{D}\mathbf{x} = \mathbf{0}_{M_3}; \mathbf{x} \text{ unrestricted} \}. \quad (\text{iii})$$

The dual to (iii) turns out to be the following LP, which is closely related to (ii):

$$\min_{\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3} \{ \mathbf{y}^{1T} \mathbf{0}_{M_1} + \mathbf{y}^{2T} \mathbf{0}_{M_2} + \mathbf{y}^{3T} \mathbf{0}_{M_3} : \mathbf{y}^{1T} \mathbf{B} + \mathbf{y}^{2T} \mathbf{C} + \mathbf{y}^{3T} \mathbf{D} = -\mathbf{1}_{M_1}^T \mathbf{B}; \mathbf{y}^1 \geq \mathbf{0}_{M_1}; \mathbf{y}^2 \geq \mathbf{0}_{M_2}; \mathbf{y}^3 \text{ unrestricted} \}. \quad (\text{iv})$$

Note that  $\mathbf{x} = \mathbf{0}_N$  is always feasible for (iii). Hence, there are two cases that can occur for (iii): Case 1: a feasible solution for (iii) exists but the corresponding optimal objective function is 0 and hence we have a bounded optimal solution for (iii) or Case 2: a feasible solution for (iii) exists and the objective function when evaluated at this feasible solution is positive and so in this case we obtain an unbounded solution for the primal and no feasible solution for the dual (iv).

**Problem 9** *Slater's* (1951)[360] *Theorem of the Alternative*: Let  $\mathbf{A}$  be  $M_1 \times N$ ,  $\mathbf{B}$  be  $M_2 \times N$ ,  $\mathbf{C}$  be  $M_3 \times N$  and  $\mathbf{D}$  be  $M_4 \times N$  with  $M_1 > 0, M_2 > 0$  and  $N > 0$ . Then *either*

$$\mathbf{A}\mathbf{x} \gg \mathbf{0}_{M_1}; \mathbf{B}\mathbf{x} > \mathbf{0}_{M_2}; \mathbf{C}\mathbf{x} \geq \mathbf{0}_{M_3}; \mathbf{D}\mathbf{x} = \mathbf{0}_{M_4} \text{ has a solution } \mathbf{x} \text{ or} \quad (\text{i})$$

$$\mathbf{y}^{1T} \mathbf{A} + \mathbf{y}^{2T} \mathbf{B} + \mathbf{y}^{3T} \mathbf{C} + \mathbf{y}^{4T} \mathbf{D} = \mathbf{0}_N^T \quad (\text{ii})$$

with  $\mathbf{y}^1 > \mathbf{0}_{M_1}; \mathbf{y}^2 \geq \mathbf{0}_{M_2}; \mathbf{y}^3 \geq \mathbf{0}_{M_3}$  or  
 $\mathbf{y}^1 \geq \mathbf{0}_{M_1}; \mathbf{y}^2 \gg \mathbf{0}_{M_2}; \mathbf{y}^3 \geq \mathbf{0}_{M_3}$

has a solution  $\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3, \mathbf{y}^4$ .<sup>\*22</sup>

*Hint:* Consider the following LP which is closely related to (i):

$$\max_{\mathbf{x}} \{ \mathbf{1}_{M_1}^T \mathbf{B}\mathbf{x} : -\mathbf{A}\mathbf{x} \leq \mathbf{1}_{M_1}; -\mathbf{B}\mathbf{x} \leq \mathbf{0}_{M_2}; -\mathbf{C}\mathbf{x} \leq \mathbf{0}_{M_3}; -\mathbf{D}\mathbf{x} = \mathbf{0}_{M_4}; \mathbf{x} \text{ unrestricted} \}. \quad (\text{iii})$$

---

<sup>\*22</sup> Both (i) and (ii) cannot hold.

The dual to (iii) turns out to be the following LP, which is closely related to (ii):

$$\begin{aligned} \min_{\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3, \mathbf{y}^4} \{ & \mathbf{y}^{1T} \mathbf{1}_{M_1} + \mathbf{y}^{2T} \mathbf{0}_{M_2} + \mathbf{y}^{3T} \mathbf{0}_{M_3} + \mathbf{y}^{4T} \mathbf{0}_{M_4} : \mathbf{y}^{2T} \mathbf{B} + \mathbf{y}^{3T} \mathbf{B} + \mathbf{y}^{4T} \mathbf{C} + \mathbf{y}^{1T} \mathbf{D} \\ & = -\mathbf{1}_{M_1}^T \mathbf{B}; \mathbf{y}^1 \geq \mathbf{0}_{M_1}; \mathbf{y}^2 \geq \mathbf{0}_{M_2}; \mathbf{y}^3 \geq \mathbf{0}_{M_3}; \mathbf{y}^4 \text{ unrestricted} \}. \end{aligned} \quad (\text{iv})$$

For this problem, there are three cases to analyze: Case 1: a finite optimal solution for (iii) exists; Case 2: (iii) has an unbounded solution and Case 3: (iii) has no feasible solution.

**Problem 10** *Factor Price Equalization Theorem*; Samuelson (1958)[344], Diewert and Woodland (1977)[158]: Suppose that a country has  $N$  production functions of the Leontief no substitution variety where each of the  $N$  activities uses varying amounts of  $M$  primary inputs. Suppose that each production function produces a single output and suppose that all outputs are traded at the positive world price vector  $\mathbf{p} \equiv [p_1, p_2, \dots, p_N]^T$ . Suppose further that the economy has an endowment of  $M$  primary inputs,  $\mathbf{v} \equiv [v_1, v_2, \dots, v_M]^T \gg \mathbf{0}_M$ . The  $M \times N$  matrix of unit output input requirements is  $\mathbf{A} \equiv [a_{mn}]$  where each  $a_{mn} \geq 0$  for  $m = 1, \dots, M$  and  $n = 1, \dots, N$ . We assume that producers take output prices as fixed and they collectively attempt to maximize the value of their country's output subject to their country's  $M$  primary resource constraints; i.e., they attempt to solve the following LP (there are no intermediate inputs in this simple model):

$$\max_{\mathbf{x}} \{ \mathbf{p}^T \mathbf{x} : \mathbf{A} \mathbf{x} \leq \mathbf{v}; \mathbf{x} \geq \mathbf{0}_N \} \quad (\text{i})$$

where  $\mathbf{x}$  is a nonnegative vector of industry outputs. The dual to (i) is:

$$\min_{\mathbf{w}} \{ \mathbf{w}^T \mathbf{v} : \mathbf{w}^T \mathbf{A} \geq \mathbf{p}^T; \mathbf{w} \geq \mathbf{0}_M \} \quad (\text{ii})$$

where  $\mathbf{w}$  can be interpreted as vector of primary input prices or resource costs. Thus in the dual problem, we attempt to choose input prices that will minimize the aggregate cost of production subject to the constraints that nonpositive profits are made in each industry. By the complementary slackness conditions for linear programs, we know that for industries that are operated at a positive scale in a solution to (i), then 0 profits will be made in that industry in equilibrium. Industries that are operated at 0 scale will make negative or 0 profits at the equilibrium input prices. Thus if producers collectively solve the above programming problems, then they will have chosen output levels that maximize the country's value of output at international prices and they will have also determined competitive input prices for their country.

Consider the following example, which has 3 production activities ( $N = 3$ ) and two primary inputs ( $M = 2$ ):

$$\begin{aligned} \max_{x_0, x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, s_1 \geq 0, s_2 \geq 0} \{ & x_0 : \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} x_0 + \begin{bmatrix} -p_1 \\ 2 \\ 0 \end{bmatrix} x_1 + \begin{bmatrix} -p_2 \\ 0 \\ 2 \end{bmatrix} x_2 \\ & + \begin{bmatrix} -p_3 \\ 3/2 \\ 1 \end{bmatrix} x_3 + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} s_1 + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} s_2 = \begin{bmatrix} 0 \\ v_1 \\ v_2 \end{bmatrix} \}. \end{aligned} \quad (\text{iii})$$

(a) Suppose  $p_1 = p_2 = p_3 = 1$ . Calculate the optimal dual prices to the LP (iii),  $w_1^*$  and  $w_2^*$ , and show that they do not depend on the particular values that  $v_1 > 0$  and  $v_2 > 0$  take on.

*Hint:* Apply the Basis Theorem for Linear Programs.

*Comment:* If each country in the world had access to the same technology matrix and faced the same vector of world prices\*<sup>23</sup> then each country which had an endowment vector which was a nonnegative

\*<sup>23</sup> Hence we are ignoring transport costs. We are also assuming that each producer behaves competitively so there

linear combination of the same activity vectors  $A_n^*$  in an optimal basis matrix would have the same vector of factor prices  $\mathbf{w}^*$ . Thus there is a tendency for trade in commodities to equalize factor prices across countries. The above result can be generalized to deal with classical production functions and the existence of intermediate inputs; see Samuelson (1958)[344], Diewert and Woodland (1977)[158] and Woodland (1982)[405].

**Problem 11** Anderson (1958; 349)[8]: Let  $\mathbf{A}(\alpha) \equiv [a_{mn}(\alpha)]$  be an  $M \times M$  matrix whose elements depend on a scalar parameter  $\alpha$ . Suppose when  $\alpha = \alpha^0$ ,  $[\mathbf{A}(\alpha^0)]^{-1} \equiv \mathbf{B}(\alpha^0)$  exists. Let the matrix of derivatives of  $\mathbf{A}(\alpha)$  evaluated at  $\alpha^0$  be denoted as  $\mathbf{A}'(\alpha^0) \equiv [da_{mn}(\alpha^0)/d\alpha]$  and let the matrix of derivatives of  $\mathbf{B}(\alpha)$  evaluated at  $\alpha^0$  be denoted as  $\mathbf{B}'(\alpha^0) \equiv [db_{mn}(\alpha^0)/d\alpha]$ . Show that

$$\mathbf{B}'(\alpha^0) = -[\mathbf{A}(\alpha^0)]^{-1} \mathbf{A}'(\alpha^0) [\mathbf{A}(\alpha^0)]^{-1}. \quad (\text{i})$$

*Hint:* Differentiate both sides of the matrix equation  $\mathbf{A}(\alpha)[\mathbf{A}(\alpha)]^{-1} = \mathbf{I}_M$  with respect to  $\alpha$ .

**Problem 12** *The Comparative Statics of the Strict Complementary Slackness Case:* Consider the following primal LP where the elements of the  $\mathbf{c}$  and  $\mathbf{b}$  vectors depend continuously on a scalar parameter  $\alpha$  as do the elements of the  $\mathbf{A}$  matrix, which is  $M \times N$  where  $M < N$ :

$$V(\alpha) \equiv \max_{\mathbf{x}} \{ \mathbf{c}(\alpha)^T \mathbf{x} : \mathbf{A}(\alpha) \mathbf{x} = \mathbf{b}(\alpha); \mathbf{x} \geq \mathbf{0}_N \}. \quad (\text{i})$$

Define the square  $M \times M$  matrix,  $\tilde{\mathbf{A}}(\alpha^0) \equiv [A_{.1}(\alpha^0), A_{.2}(\alpha^0), \dots, A_{.M}(\alpha^0)]$  and we suppose that this matrix satisfies the following conditions:

$$[\tilde{\mathbf{A}}(\alpha^0)]^{-1} \text{ exists;} \quad (\text{ii})$$

$$\tilde{\mathbf{x}}(\alpha^0) \equiv [\tilde{\mathbf{A}}(\alpha^0)]^{-1} \mathbf{b}(\alpha^0) \gg \mathbf{0}_M; \mathbf{x}(\alpha^0)^T \equiv [\tilde{\mathbf{x}}(\alpha^0)^T, \mathbf{0}_{N-M}^T]; \quad (\text{iii})$$

$$\mathbf{y}(\alpha^0)^T \equiv [c_1(\alpha^0), c_2(\alpha^0), \dots, c_M(\alpha^0)] [\tilde{\mathbf{A}}(\alpha^0)]^{-1} \equiv \tilde{\mathbf{c}}(\alpha^0)^T [\tilde{\mathbf{A}}(\alpha^0)]^{-1}; \quad (\text{iv})$$

$$\mathbf{y}(\alpha^0)^T A_{.n}(\alpha^0) > c_n(\alpha^0) \text{ for } n = M+1, M+2, \dots, N. \quad (\text{v})$$

From our knowledge of the simplex algorithm, it can be seen that conditions (ii)-(v) are sufficient to imply that the first  $M$  columns of the  $\mathbf{A}(\alpha^0)$  matrix form an optimal basis matrix for the LP defined by (i) when  $\alpha = \alpha^0$ . Conditions (v) imply that the  $\mathbf{x}(\alpha^0)$  solution defined in (iii) is the *unique solution* to the primal when  $\alpha = \alpha^0$  and the strict inequalities in (iii) along with assumption (ii) implies that  $\mathbf{y}(\alpha^0)$  defined in (iv) is the *unique solution* to the following dual problem when  $\alpha = \alpha^0$ :

$$V(\alpha) \equiv \min_{\mathbf{y}} \{ \mathbf{y}^T \mathbf{b}(\alpha) : \mathbf{y}^T \mathbf{A}(\alpha) \geq \mathbf{c}(\alpha); \mathbf{y} \text{ unrestricted} \}. \quad (\text{vi})$$

The continuity of the function  $\mathbf{A}(\alpha)$ ,  $\mathbf{b}(\alpha)$  and  $\mathbf{c}(\alpha)$  means that for  $\alpha$  sufficiently close to  $\alpha^0$ , the primal solution  $\mathbf{x}(\alpha)$  to  $V(\alpha)$  defined by (i) will be

$$\tilde{\mathbf{x}}(\alpha) \equiv [\tilde{\mathbf{A}}(\alpha)]^{-1} \mathbf{b}(\alpha); \mathbf{x}(\alpha)^T \equiv [\tilde{\mathbf{x}}(\alpha)^T, \mathbf{0}_{N-M}^T] \quad (\text{vii})$$

and the dual solution  $\mathbf{y}(\alpha)$  to  $V(\alpha)$  defined by (vi) will be

$$\mathbf{y}(\alpha)^T \equiv [c_1(\alpha), c_2(\alpha), \dots, c_M(\alpha)] [\tilde{\mathbf{A}}(\alpha)]^{-1} \equiv \tilde{\mathbf{c}}(\alpha)^T [\tilde{\mathbf{A}}(\alpha)]^{-1}. \quad (\text{viii})$$

---

are a lot of assumptions in this model that are unlikely to be completely satisfied in the real world. Moreover, factor price equalization becomes increasingly unlikely as the number of primary inputs (and domestically produced services which are not internationally traded) increases. But the factor price equalization theorem is a "pretty" theoretical result!

(a) Show that if  $\mathbf{A}(\alpha)$  and  $\mathbf{b}(\alpha)$  are differentiable at  $\alpha = \alpha^0$ , then the optimal  $\mathbf{x}$  solution to (i) is differentiable at  $\alpha = \alpha^0$ , with

$$\mathbf{x}'(\alpha^0)^T \equiv [\tilde{\mathbf{x}}'(\alpha^0)^T, \mathbf{0}_{N-M}^T] \text{ and} \quad (\text{ix})$$

$$\tilde{\mathbf{x}}'(\alpha^0) \equiv [\tilde{\mathbf{A}}(\alpha^0)]^{-1} \mathbf{b}'(\alpha^0) - [\tilde{\mathbf{A}}(\alpha^0)]^{-1} \tilde{\mathbf{A}}'(\alpha^0) [\tilde{\mathbf{A}}(\alpha^0)]^{-1} \mathbf{b}(\alpha^0). \quad (\text{x})$$

(b) Show that if  $\mathbf{A}(\alpha)$  and  $\mathbf{c}(\alpha)$  are differentiable at  $\alpha = \alpha^0$ , then the optimal  $\mathbf{y}$  solution to (vi) is differentiable at  $\alpha = \alpha^0$ , with

$$\mathbf{y}'(\alpha)^T = \tilde{\mathbf{c}}'(\alpha)^T [\tilde{\mathbf{A}}(\alpha)]^{-1} - \tilde{\mathbf{c}}(\alpha)^T [\tilde{\mathbf{A}}(\alpha^0)]^{-1} \tilde{\mathbf{A}}'(\alpha^0) [\tilde{\mathbf{A}}(\alpha^0)]^{-1}. \quad (\text{xi})$$

(c) Show that if  $\mathbf{A}(\alpha)$ ,  $\mathbf{b}(\alpha)$  and  $\mathbf{c}(\alpha)$  are differentiable at  $\alpha = \alpha^0$ , then the optimal primal and dual objective functions  $V(\alpha)$  defined by (i) or (vi) are differentiable at  $\alpha = \alpha^0$  with

$$V'(\alpha) = \tilde{\mathbf{c}}'(\alpha)^T [\tilde{\mathbf{A}}(\alpha)]^{-1} \mathbf{b}(\alpha^0) - \tilde{\mathbf{c}}(\alpha)^T [\tilde{\mathbf{A}}(\alpha^0)]^{-1} \tilde{\mathbf{A}}'(\alpha^0) [\tilde{\mathbf{A}}(\alpha^0)]^{-1} \mathbf{b}(\alpha^0) + \tilde{\mathbf{c}}(\alpha)^T [\tilde{\mathbf{A}}(\alpha)]^{-1} \mathbf{b}'(\alpha^0). \quad (\text{xii})$$

(d) Specialize the results in (a)-(c) to the case where only  $\mathbf{A}$  depends on  $\alpha$ ; i.e., assume  $\mathbf{b}$  and  $\mathbf{c}$  are constant vectors.

(e) Specialize the results in (a)-(c) to the case where only  $\mathbf{b}$  depends on  $\alpha$ .

(f) Specialize the results in (a)-(c) to the case where only  $\mathbf{c}$  depends on  $\alpha$ .

*Comment:* For similar comparative statics results for a general nonlinear programming problem, see Diewert (1984)[99].

**Problem 13** *Some Global Comparative Statics Results for Linear Programming Problems;* Beckmann (1955-56)[31]: Consider the following two linear programs:

$$\max_{\mathbf{x}} \{ \mathbf{c}^{jT} \mathbf{x} : \mathbf{A}^j \mathbf{x} \leq \mathbf{b}^j; \mathbf{x} \geq \mathbf{0}_N \}; \quad j = 0, 1 \quad (\text{i})$$

where  $\mathbf{A}^j$  is an  $M \times N$  matrix. The dual problems that correspond to the primal problems (i) are:

$$\min_{\mathbf{y}} \{ \mathbf{y}^T \mathbf{b}^j : \mathbf{y}^T \mathbf{A}^j \geq \mathbf{c}^{jT}; \mathbf{y} \geq \mathbf{0}_M \}; \quad j = 0, 1. \quad (\text{ii})$$

Suppose that the problems in (i) and (ii) have finite optimal solutions,  $\mathbf{x}^j$  and  $\mathbf{y}^j$  respectively when  $j = 0, 1$ . The changes  $\Delta \mathbf{A}$ ,  $\Delta \mathbf{b}$ ,  $\Delta \mathbf{c}$ ,  $\Delta \mathbf{x}$  and  $\Delta \mathbf{y}$  are defined as follows:

$$\mathbf{A}^1 \equiv \mathbf{A}^0 + \Delta \mathbf{A}; \mathbf{b}^1 \equiv \mathbf{b}^0 + \Delta \mathbf{b}; \mathbf{c}^1 \equiv \mathbf{c}^0 + \Delta \mathbf{c}; \mathbf{x}^1 \equiv \mathbf{x}^0 + \Delta \mathbf{x}; \mathbf{y}^1 \equiv \mathbf{y}^0 + \Delta \mathbf{y}. \quad (\text{iii})$$

(a) Show that the changes  $\Delta \mathbf{A}$ ,  $\Delta \mathbf{b}$ ,  $\Delta \mathbf{c}$ ,  $\Delta \mathbf{x}$  and  $\Delta \mathbf{y}$  and  $\mathbf{x}^0$  and  $\mathbf{y}^0$  satisfy the following inequality:

$$[\Delta \mathbf{c}^T - \mathbf{y}^{0T} \Delta \mathbf{A}] \Delta \mathbf{x} - \Delta \mathbf{y}^T [\Delta \mathbf{b} - \Delta \mathbf{A} \mathbf{x}^0] \geq 0. \quad (\text{iv})$$

*Hint:* Use the Saddlepoint inequalities (11.118) for the LP's in (i).

(b) Specialize (iv) to the case where only  $\mathbf{b}$  changes. Further specialize (iv) to the case where only the first component of  $\mathbf{b}$  changes. Provide an economic interpretation for the resulting formula.

(c) Specialize (iv) to the case where only  $\mathbf{c}$  changes. Further specialize (iv) to the case where only the first component of  $\mathbf{c}$  changes. Provide an economic interpretation for the resulting formula.

(d) Specialize (iv) to the case where only one component of  $\mathbf{A}$  changes, say  $a_{mn}$  changes.

(e) *Bailey's (1955-56)[17] Inequalities*: Show that under the above conditions, the following two inequalities are also valid:<sup>\*24</sup>

$$\begin{aligned}
 \Delta[\mathbf{c}^T - \mathbf{y}^T \mathbf{A}] \Delta \mathbf{x} &\equiv [(\mathbf{c}^{1T} - \mathbf{y}^{1T} \mathbf{A}^1) - (\mathbf{c}^{0T} - \mathbf{y}^{0T} \mathbf{A}^0)] \Delta \mathbf{x} \\
 &= [\Delta \mathbf{c}^T - \mathbf{y}^{1T} \mathbf{A}^1 + \mathbf{y}^{0T} \mathbf{A}^0] \Delta \mathbf{x} \\
 &= [\Delta \mathbf{c}^T - \Delta \mathbf{y}^T \mathbf{A}^1 - \mathbf{y}^{0T} \Delta \mathbf{A}] \Delta \mathbf{x} \\
 &= [\Delta \mathbf{c}^T - \mathbf{y}^{1T} \Delta \mathbf{A} - \Delta \mathbf{y}^T \mathbf{A}^0] \Delta \mathbf{x} \\
 &\geq 0.
 \end{aligned} \tag{v}$$

$$\begin{aligned}
 \Delta \mathbf{y}^T \Delta [\mathbf{b} - \mathbf{A} \mathbf{x}] &\equiv \Delta \mathbf{y}^T [(\mathbf{b}^1 - \mathbf{A}^1 \mathbf{x}^1) - (\mathbf{b}^0 - \mathbf{A}^0 \mathbf{x}^0)] \\
 &= \Delta \mathbf{y}^T [\Delta \mathbf{b} - \mathbf{A}^1 \mathbf{x}^1 + \mathbf{A}^0 \mathbf{x}^0] \\
 &= \Delta \mathbf{y}^T [\Delta \mathbf{b} - \Delta \mathbf{A} \mathbf{x}^1 - \mathbf{A}^0 \Delta \mathbf{x}] \\
 &= \Delta \mathbf{y}^T [\Delta \mathbf{b} - \mathbf{A}^1 \Delta \mathbf{x}^1 - \Delta \mathbf{A} \mathbf{x}^0] \\
 &\leq 0.
 \end{aligned} \tag{vi}$$

*Hint*: Define the slack vectors for the primal problems as  $\mathbf{s}^j \equiv \mathbf{b}^j - \mathbf{A}^j \mathbf{x}^j \geq \mathbf{0}_M$  for  $j = 0, 1$ . Recall that  $\mathbf{y}^{jT} \mathbf{s}^j = 0$  for  $j = 0, 1$ . Now note that  $\Delta \mathbf{y}^T \Delta [\mathbf{b} - \mathbf{A} \mathbf{x}] = [\mathbf{y}^{1T} - \mathbf{y}^{0T}] [\mathbf{s}^1 - \mathbf{s}^0] = \mathbf{y}^{1T} \mathbf{s}^1 + \mathbf{y}^{0T} \mathbf{s}^0 - \mathbf{y}^{1T} \mathbf{s}^0 - \mathbf{y}^{0T} \mathbf{s}^1$ . Using the complementary slackness conditions for each primal LP, the inequality (vi) follows readily. A similar analysis works for (v).

**Problem 14** Define the primal and dual LP's for two  $\mathbf{c}$  vectors as follows:

$$V(\mathbf{c}^j) \equiv \max_{\mathbf{x}} \{ \mathbf{c}^{jT} \mathbf{x} : \mathbf{A} \mathbf{x} \leq \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N \}; \quad j = 1, 2; \tag{i}$$

$$V(\mathbf{c}^j) \equiv \min_{\mathbf{y}} \{ \mathbf{y}^T \mathbf{b} : \mathbf{y}^T \mathbf{A} \geq \mathbf{c}^{jT}; \mathbf{y} \geq \mathbf{0}_M \}; \quad j = 1, 2. \tag{ii}$$

Suppose that feasible solutions exist for both the primal and dual problems defined by (i) and (ii) so that each problem has a finite optimal solution. Let  $\lambda$  be a scalar such that  $0 < \lambda < 1$  and define the following LP:

$$V(\lambda \mathbf{c}^1 + (1 - \lambda) \mathbf{c}^2) \equiv \max_{\mathbf{x}} \{ [\lambda \mathbf{c}^1 + (1 - \lambda) \mathbf{c}^2]^T \mathbf{x} : \mathbf{A} \mathbf{x} \leq \mathbf{b}; \mathbf{x} \geq \mathbf{0}_N \}. \tag{iii}$$

(a) Show that a finite optimal solution for the LP defined by (iii) exists.

*Hint*: Show that a feasible solution exists for (iii) and its dual.

(b) Show that:

$$V(\lambda \mathbf{c}^1 + (1 - \lambda) \mathbf{c}^2) \leq \lambda V(\mathbf{c}^1) + (1 - \lambda) V(\mathbf{c}^2). \tag{iv}$$

*Comment*: This problem shows that the optimized objective function of a linear programming problem is a *convex function* in the  $\mathbf{c}$  vector.

**Problem 15** Define the primal and dual LP's for two  $\mathbf{b}$  vectors as follows:

$$V(\mathbf{b}^j) \equiv \max_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A} \mathbf{x} \leq \mathbf{b}^j; \mathbf{x} \geq \mathbf{0}_N \}; \quad j = 1, 2; \tag{i}$$

$$V(\mathbf{b}^j) \equiv \min_{\mathbf{y}} \{ \mathbf{y}^T \mathbf{b}^j : \mathbf{y}^T \mathbf{A} \geq \mathbf{c}^T; \mathbf{y} \geq \mathbf{0}_M \}; \quad j = 1, 2. \tag{ii}$$

<sup>\*24</sup> These two inequalities imply Beckmann's inequalities.

Suppose that feasible solutions exist for both the primal and dual problems defined by (i) and (ii) so that each problem has a finite optimal solution. Let  $\lambda$  be a scalar such that  $0 < \lambda < 1$  and define the following LP:

$$V(\lambda \mathbf{b}^1 + (1 - \lambda) \mathbf{b}^2) \equiv \max_{\mathbf{x}} \{ \mathbf{c}^T \mathbf{x} : \mathbf{A} \mathbf{x} \leq \lambda \mathbf{b}^1 + (1 - \lambda) \mathbf{b}^2; \mathbf{x} \geq \mathbf{0}_N \}. \quad (\text{iii})$$

(a) Show that a finite optimal solution for the LP defined by (iii) exists.

*Hint:* Show that a feasible solution exists for (iii) and its dual.

(b) Show that:

$$V(\lambda \mathbf{b}^1 + (1 - \lambda) \mathbf{b}^2) \geq \lambda V(\mathbf{b}^1) + (1 - \lambda) V(\mathbf{b}^2). \quad (\text{iv})$$

*Comment:* This problem shows that the optimized objective function of a linear programming problem is a *concave function* in the right hand side vector  $\mathbf{b}$  over the set of  $\mathbf{b}$ 's for which there is a feasible solution for the primal problem.

**Problem 16** Consider the LP defined by (11.99) and (11.100) and suppose that  $[b_1^*, b_2^*] = [2, 1]$  so that the  $\mathbf{b}$  vector is a multiple of  $A_{.3}$  and  $\mathbf{b}$  lies on the dashed line through  $A_{.3}$  in Figure 11.2. Now suppose  $b_1$  *increases* by a marginal unit from its initial value so that the new  $[b_1, b_2^*]$  lies in the interior of the cone 3 region. Regard the optimized objective function of the LP defined by (11.100) as a function of  $b_1$ , say  $V(b_1)$ .

(a) Show that for  $[b_1, b_2^*]$  in the interior of the cone 3 region, the derivative of the optimal objective function with respect to  $b_1$  is

$$dV(b_1)/db_1 = y_1^{(3)} = 1/2. \quad (\text{i})$$

*Hint:* Recall (11.106) and use part (xii) of Problem 12 above.

Now suppose  $b_1$  *decreases* by a marginal unit from its initial value of  $[b_1^*, b_2^*] = [2, 1]$  so that the new  $[b_1, b_2^*]$  lies in the interior of the cone 2 region. Again, regard the optimized objective function of the LP defined by (11.100) as a function of  $b_1$ , say  $V(b_1)$ .

(b) Show that for  $[b_1, b_2^*]$  in the interior of the cone 2 region, the derivative of the optimal objective function with respect to  $b_1$  is

$$dV(b_1)/db_1 = y_1^{(2)} = 2/3. \quad (\text{ii})$$

*Hint:* Recall (11.102) and use part (xii) of Problem 12 above.

(c) Use parts (a) and (b) above to show that

$$dV^+(b_1^*)/db_1 \equiv \lim_{h \rightarrow 0, h > 0} [V(b_1^* + h) - V(b_1^*)]/h = y_1^{(3)} \quad \text{and} \quad (\text{iii})$$

$$dV^-(b_1^*)/db_1 \equiv \lim_{h \rightarrow 0, h < 0} [V(b_1^* + h) - V(b_1^*)]/h = y_1^{(2)}. \quad (\text{iv})$$

*Comment:* This problem shows that for the LP defined by (11.100), the optimized objective function is not always differentiable with respect to the components of the right hand side  $\mathbf{b}$  vector. However, the *one sided derivatives* or *directional derivatives* of  $V(\mathbf{b})$  always exist. This problem also casts a bit more light on the meaning of the dual prices in the general case (as opposed to the special situation that we had in problem 12 where the dual price vector was locally unique).

## 11.12 References

Anderson, T.W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley.  
 Baily, M.J. (1955-6), "A Generalized Comparative Statics in Linear Programming", *Review of Economic Studies* 23, 236-240.

- Beale, E.M.L. (1955), "Cycling in the Dual Simplex Algorithm", *Naval Research Logistics Quarterly* 2, 269-276.
- Beckmann, M.J. (1955-6), "Comparative Statics in Linear Programming and the Giffin Paradox", *Review of Economic Studies* 23, 232-235.
- Carathéodory, C. (1911), "Über den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen", *Rendiconti Circolo Matematico Palermo* 32, 193-217.
- Dantzig, G.B. (1951), "Maximization of a Linear Function of Variables subject to Linear Inequalities", pp.339-347 in *Activity Analysis of Production and Allocation*, T.C. Koopmans (ed.), New York: John Wiley.
- Dantzig, G.B. (1963), *Linear Programming and Extensions*, Princeton N.J.: Princeton University Press.
- Diewert, W.E. (1975), "The Samuelson Nonsubstitution Theorem and the Computation of Equilibrium Prices", *Econometrica* 43, 57-64.
- Diewert, W.E. (1978), "Optimal Tax Perturbations", *Journal of Public Economics* 10, 138-177.
- Diewert, W.E. (1983), "Cost-Benefit Analysis and Project Evaluation: A Comparison of Alternative Approaches", *Journal of Public Economics* 22, 265-302.
- Diewert, W.E. (1984), "Sensitivity Analysis in Economics", *Computers and Operations Research* 11:2, 141-156.
- Diewert, W.E. (1987), "The Effects of an Innovation: A Trade Theory Approach", *The Canadian Journal of Economics* 20, 694-714.
- Diewert, W.E., A.H. Turunen-Red and A.D. Woodland (1989), "Productivity and Pareto Improving Changes in Taxes and Tariffs", *Review of Economic Studies* 56, 199-216.
- Diewert, W.E., A.H. Turunen-Red and A.D. Woodland (1991), "Tariff Reform in a Small Open Multi-Household Economy with Domestic Distortions and Nontraded Goods", *International Economic Review* 32, 937-957.
- Diewert, W.E. and A.D. Woodland (1977), "Frank Knight's Theorem in Linear Programming Revisited", *Econometrica* 45, 375-398.
- Dorfman, R., P.A. Samuelson and R.M. Solow (1958), *Linear Programming and Economic Analysis*, New York: McGraw-Hill Book Company.
- Farkas, J. (1902), "Über die Theorie der einfachen Ungleichungen", *Journal für reine und angewandte Mathematik* 124, 1-24.
- Fenchel, W. (1953), "Convex Cones, Sets and Functions", Lecture Notes at Princeton University, Department of Mathematics, Princeton, N.J.
- Goldman, A.J. and A.W. Tucker (1956), "Theory of Linear Programming", pp. 53-97 in *Linear Inequalities and Related Systems*, H.W. Kuhn and A.W. Tucker (eds.), Princeton N.J.: Princeton University Press.
- Hoffman, A.J. (1953), "Cycling in the Simplex Algorithm", National Bureau of Standards Report 2974, December 16, 7 pages.
- Hotelling, H. (1932). "Edgeworth's taxation paradox and the nature of demand and supply functions". *Journal of Political Economy* 40 (5): 577-616.
- Leontief, W.W. (1936), "Quantitative Input and Output Relations in the Economic System of the United States", *Review of Economic Statistics* 18, 105-125.
- Leontief, W.W. (1941), *The Structure of the American Economy, 1919-1929*, Cambridge, MA: Harvard University Press.
- Malinvaud, E. (1967), "Decentralized Procedures for Planning", in *Activity Analysis in the Theory of Growth and Planning*, E. Malinvaud and M.O.L. Bacharach (eds.), London: Macmillan.

- Mangasarian, O. (1969), *Nonlinear Programming*, New York: McGraw-Hill.
- Motzkin, T.S. (1936), *Beiträge zur Theorie der linearen Ungleichungen*, Doctoral Thesis, University of Basel, Jerusalem: Azriel.
- Samuelson, P.A. (1951), “Abstract of a Theorem Concerning Substitutability in Open Leontief Models”, Chapter 7 in *Activity Analysis of Production and Allocation*, T.C. Koopmans (ed.), New York: John Wiley.
- Samuelson, P.A. (1958), “Frank Knight’s Theorem in Linear Programming”, *Zeitschrift für National Ökonomie* 18, 310-317.
- Samuelson, P.A. and S. Swamy (1974), “Invariant Economic Index Numbers and Canonical Duality: Survey and Synthesis”, *American Economic Review* 64, 566-593.
- Shephard, R.W. (1953), *Theory of Cost and Production Functions*, Princeton University Press.
- Slater, M. (1951), “A Note on Motzkin’s Transposition Theorem”, *Econometrica* 19, 185-186.
- Tucker, A.W. (1956), “Dual Systems of Homogeneous Linear Relations”, pp. 3-18 in *Linear Inequalities and Related Systems*, H.W. Kuhn and A.W. Tucker (eds.), Princeton N.J.: Princeton University Press.
- Van Slyke, R.M. (1968), *Mathematical Programming and Optimal Control Theory*, Operations Research Center Report 68-21, University of California, Berkeley, July.
- von Neumann, J. (1947), “On a Maximization Problem”, manuscript, Institute for Advanced Studies, Princeton University, November.
- Weymark, J.A. (1979), “A Reconciliation of Recent Results in Optimal Taxation Theory”, *Journal of Public Economics* 12, 171-189.
- Woodland, A.D. (1982), *International Trade and Resource Allocation*, Amsterdam: North-Holland.

# Bibliography

- [1] Adcock, R. J. (1878), "A Problem in Least Squares", *Analyst [Annals of Mathematics]* 5, 53-54.
- [2] Allais, M. (1947), *Economie et Intérêt*, Paris: Imprimerie Nationale.
- [3] Allen, R.C. (1983), "Collective Invention", *Journal of Economic Behavior and Organization* 4, 1-24.
- [4] Allen, R.G.C. (1938), *Mathematical Analysis for Economists*, London: Macmillan.
- [5] Allen, R.G.D. (1939), "The Assumptions of Linear Regression", *Economica (New Series)* 6, 191-201.
- [6] Allen, R.G.D. (1949), "The Economic Theory of Index Numbers", *Economica* 16, 197-203.
- [7] Alterman, W.F., W.E. Diewert and R.C. Feenstra (1999), *International Trade Price Indexes and Seasonal Commodities*, Bureau of Labor Statistics, Washington D.C.
- [8] Anderson, T.W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley.
- [9] Anthony, R.N. (1973), "Accounting for the Cost of Equity", *Harvard Business Review* 51, 88-102.
- [10] Archibald, R.B. (1977), "On the Theory of Industrial Price Measurement: Output Price Indexes", *Annals of Economic and Social Measurement* 6, 57-72.
- [11] Arrow, K.J. (1962), "The Economic Implications of Learning by Doing", *The Review of Economic Studies* 29, 155-173.
- [12] Arrow, K.J. (1969), "Classificatory Notes on the Production and Transmission of Technological Knowledge", *American Economic Review* 59 (May), 29-35.
- [13] Arrow, K.J., H.B. Chenery, B.S. Minhas and R.M. Solow, (1961), "Capital-Labour Substitution and Economic Efficiency", *Review of Economics and Statistics* 63, 225-250.
- [14] Arrow, K.J. and A.C. Enthoven (1961), "Quasi-Concave Programming", *Econometrica* 29, 779-800.
- [15] Babbage, C. (1835), *On the Economy of Machinery and Manufactures*, Fourth Edition, reprinted by A. M. Kelley, New York, 1965.
- [16] Babbage, C. (1835), *On the Economy of Machinery and Manufactures*, Fourth Edition, London: Charles Knight.
- [17] Baily, M.J. (1955-6), "A Generalized Comparative Statics in Linear Programming", *Review of Economic Studies* 23, 236-240.
- [18] Balk, B.M. (1989), "Changing Consumer Preferences and the cost of Living Index: Theory and Nonparametric Expressions", *Journal of Economics* 50, 157-169.
- [19] Balk, B.M. (1995), "Axiomatic Price Index Theory: A Survey", *International Statistical Review* 63, 69-93.
- [20] Balk, B.M. (1998), *Industrial Price, Quantity and Productivity Indices*, Boston: Kluwer Academic Publishers.
- [21] Balk, B.M. (2008), *Price and Quantity Index Numbers*, New York: Cambridge University Press.
- [22] Bartelsman, E. J. (1995), "Of Empty Boxes: Returns to Scale Revisited," *Economics Letters* 49, 59-67.
- [23] Basu, S. and J. G. Fernald (1997), "Returns to Scale in U.S. Production: Estimates and Implications", *Journal of Political Economy* 105, 249-283.

- [24] Basu, S. and J. G. Fernald (2002), "Aggregate Productivity and Aggregate Technology", *European Economic Review* 46, 963-991.
- [25] Bates, W. (2001), *How Much Government? The effects of high government spending on economic performance*, New Zealand Business Roundtable, Wellington, August 2001.
- [26] Baumol, W.J. (1952), "The Transactions Demand For Cash: An Inventory Theoretic Approach", *Quarterly Journal of Economics* 66, 545-556.
- [27] Baxter, W.T. (1971), *Depreciation*, London: Sweet and Maxwell.
- [28] Baxter, W.T. (1975), *Accounting Values and Inflation*, London: McGraw-Hill.
- [29] Baxter, W.T. (1984), *Inflation Accounting*, Oxford: Philip Allen Publishers.
- [30] Beale, E.M.L. (1955), "Cycling in the Dual Simplex Algorithm", *Naval Research Logistics Quarterly* 2, 269-276.
- [31] Beckmann, M.J. (1955-6), "Comparative Statics in Linear Programming and the Giffin Paradox", *Review of Economic Studies* 23, 232-235.
- [32] Beidelman, C. (1973), *Valuation of Used Capital Assets*, Sarasota Florida: American Accounting Association.
- [33] Beidelman, C.R. (1976), "Economic Depreciation in a Capital Goods Industry", *National Tax Journal* 29, 379-390.
- [34] Bell, A.L. (1953), "Fixed Assets and Current Costs", *The Accounting Review* 28, 44-53.
- [35] Berge, C. (1963), *Topological Spaces*, New York: MacMillan.
- [36] Blackorby, C. (1975), "Degrees of Cardinality and Aggregate Partial Orderings", *Econometrica* 43, 845-852.
- [37] Blackorby, C., R. Davidson and D. Donaldson (1977), "A Homiletic Exposition of the Expected Utility Theorem", *Economica* 44, 351-358.
- [38] Blackorby, C. and W.E. Diewert (1979), "Expenditure Functions, Local Duality and Second Order Approximations", *Econometrica* 47, 579-601.
- [39] Blackorby, C., D. Primont and R.R. Russell (1978), *Duality, Separability and Functional Structure: Theory and Economic Applications*, New York: North-Holland.
- [40] Böhm-Bawerk, E. V. (1891), *The Positive Theory of Capital*, W. Smart (translator of the original German book published in 1888), New York: G.E. Stechert.
- [41] Bowley, A.L. (1901), *Elements of Statistics*, Westminster: P.S. King and Son.
- [42] Bowley, A.L. (1919), "The Measurement of Changes in the Cost of Living", *Journal of the Royal Statistical Society* 82, 343-372.
- [43] Burgess, D. F. (1974), "A Cost Minimization Approach to Import Demand Equations," *Review of Economics and Statistics* 56 (2): 224-234.
- [44] Canning, J.B. (1929), *The Economics of Accountancy*, New York: The Ronald Press Co.
- [45] Carathéodory, C. (1911), "Über den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen", *Rendiconti Circolo Matematico Palermo* 32, 193-217.
- [46] Carli, Gian-Rinaldo, (1804), "Del valore e della proporzione de' metalli monetati", pp. 297-366 in *Scrittori classici italiani di economia politica*, Volume 13, Milano: G.G. Destefanis (originally published in 1764).
- [47] Carsberg, B. (1982), "The Case for Financial Capital Maintenance", pp. 59-74 in *Maintenance of Capital: Financial versus Physical*, R.R. Sterling and K.W. Lemke (eds.), Houston: Scholars Book Co.
- [48] Cauchy, A.L. (1821), *Cours d'analyse de l'École Polytechnique*, Volume 1, *Analyse algébrique*, Paris.
- [49] Caves, D.W., L.R. Christensen and W.E. Diewert (1982), "The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity", *Econometrica* 50, 1392-1414.
- [50] Chipman, J.S. (1966), "A Survey of the Theory of International Trade: Part 3: The Modern Theory", *Econometrica* 34, 18-76.

- 
- [51] Chipman, J.S. (1972), "The Theory of Exploitative Trade and Investment Policies: A Reformulation and Synthesis", in *International Economics and Development: Essays in Honor of Raul Prebisch*, L.D. de Marco (ed.), New York: Academic Press.
- [52] Christensen, L.R. and D.W. Jorgenson (1969), "The Measurement of U.S. Real Capital Input, 1929-1967", *Review of Income and Wealth* 15, 293-320.
- [53] Christensen, L.R., D.W. Jorgenson and L.J. Lau (1971), "Conjugate Duality and the Transcendental Logarithmic Production Function," *Econometrica* 39, 255-256.
- [54] Christensen, L.R. and D.W. Jorgenson (1973), "Measuring the Performance of the Private Sector of the U.S. Economy, 1929-1969", pp. 233-351 in *Measuring Economic and Social Performance*, M. Moss (ed.), New York: Columbia University Press.
- [55] Christensen, L.R., D.W. Jorgenson and L.J. Lau (1975), "Transcendental Logarithmic Utility Functions", *American Economic Review* 65, 367-383.
- [56] Church, A.H. (1901), "The Proper Distribution of Establishment Charges, Parts I, II, and III", *The Engineering Magazine* 21, 508-517; 725-734; 904-912.
- [57] Clark, D. (1940), *The Conditions of Economic Progress*, London: Macmillan.
- [58] Clements, K.W., H.Y. Izan and E.A. Selvanathan (2006), "Stochastic Index Numbers: A Review", *International Statistical Review* 74, 235-270.
- [59] Cobb, C. and P.H. Douglas (1928), "A Theory of Production", *American Economic Review*, Supplement, 18, 139-165.
- [60] Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton, New Jersey: Princeton University Press.
- [61] Crandell, W.T. (1935), "Income and its Measurement", *The Accounting Review* 10, 380-400.
- [62] Daines, H.C. (1929), "The Changing Objectives of Accounting", *The Accounting Review* 4, 94-110.
- [63] Daniels, M.B. (1933), "The Valuation of Fixed Assets", *The Accounting Review* 8, 302-316.
- [64] Dantzig, G.B. (1951), "Maximization of a Linear Function of Variables subject to Linear Inequalities", pp.339-347 in *Activity Analysis of Production and Allocation*, T.C. Koopmans (ed.), New York: John Wiley.
- [65] Dantzig, G.B. (1963), *Linear Programming and Extensions*, Princeton N.J.: Princeton University Press.
- [66] Davies, G.R. (1924), "The Problem of a Standard Index Number Formula", *Journal of the American Statistical Association* 19, 180-188.
- [67] Davies, G.R. (1924), "The Problem of a Standard Index Number Formula", *Journal of the American Statistical Association* 27, 180-188.
- [68] Davies, G.R. (1932), "Index Numbers in Mathematical Economics", *Journal of the American Statistical Association* 27, 58-64.
- [69] de Haan, J. and H.A. van der Grient (2011), "Eliminating Chain drift in Price Indexes Based on Scanner Data", *Journal of Econometrics* 161, 36-46.
- [70] de Haan, J. and F. Krsinich (2012), "The Treatment of Unmatched Items in Rolling Year GEKS Price Indexes: Evidence from New Zealand Scanner Data", paper presented at the Meeting of Groups of Experts on Consumer Price Indices Organized jointly by UNECE and ILO at the United Nations Palais des Nations, Geneva Switzerland, May 30-June 1, 2012.
- [71] Debreu, G. (1959), *Theory of Value*, New York: John Wiley and Sons.
- [72] Denny, M. (1974), "The Relationship Between Functional Forms for the Production System", *Canadian Journal of Economics* 7, 21-31.
- [73] Diewert, W.E. (1971), "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function", *Journal of Political Economy* 79, 481-507.
- [74] Diewert, W.E. (1973), "Functional Forms for Profit and Transformation Functions", *Journal of Economic Theory* 6, 284-316.
- [75] Diewert, W.E. (1974), "Intertemporal Consumer Theory and the Demand for Durables", *Econometrica* 42, 497-516.

- [76] Diewert, W.E. (1974), "Applications of Duality Theory", pp. 106-171 in M.D. Intriligator and D.A. Kendrick (ed.), *Frontiers of Quantitative Economics*, Vol. II, Amsterdam: North-Holland.
- [77] Diewert, W.E. (1974a), "Applications of Duality Theory", pp. 106-171 in *Frontiers of Quantitative Economics*, Volume 2, M.D. Intriligator and D.A. Kendrick (eds.), Amsterdam: North-Holland.
- [78] Diewert, W.E. (1974b), "Functional Forms for Revenue and Factor Requirements Functions", *International Economic Review* 15, 119-130.
- [79] Diewert, W.E. (1974c), "Functional Forms for Revenue and Factor Requirements Functions", *International Economic Review* 15, 119-130.
- [80] Diewert, W.E. (1974), "Intertemporal Consumer Theory and the Demand for Durables", *Econometrica* 42, 497-516.
- [81] Diewert, W.E. (1975), "The Samuelson Nonsubstitution Theorem and the Computation of Equilibrium Prices", *Econometrica* 43, 57-64.
- [82] Diewert, W.E. (1976), "Exact and Superlative Index Numbers", *Journal of Econometrics* 4, 114-145.
- [83] Diewert, W.E. (1977), "Walras' Theory of Capital Formation and the Existence of a Temporary Equilibrium", pp. 73-126 in *Equilibrium and Disequilibrium in Economic Theory*, E. Schwödiauer (ed.), Reidel Publishing Co.
- [84] Diewert, W.E. (1977), "Walras' Theory of Capital Formation and the Existence of a Temporary Equilibrium", pp. 73-126 in *Equilibrium and Disequilibrium in Economic Theory*, G. Schwödiauer (ed.), Dordrecht: D. Reidel.
- [85] Diewert, W.E. (1978), "Superlative Index Numbers and Consistency in Aggregation", *Econometrica* 46, 883-900.
- [86] Diewert, W.E. (1978), "Hicks' Aggregation Theorem and the Existence of a Real Value Added Function", pp. 17-51, Vol. 2, in *Production Economics: A Dual Approach to Theory and Applications*, M. Fuss and D. McFadden, editors, North-Holland, Amsterdam.
- [87] Diewert, W.E. (1978), "Optimal Tax Perturbations", *Journal of Public Economics* 10, 138-177.
- [88] Diewert, W.E. (1980), "Symmetry Conditions for Market Demand Functions", *Review of Economic Studies* 47, 595-601.
- [89] Diewert, W.E. (1980), "Aggregation Problems in the Measurement of Capital", pp. 433-528 in *The Measurement of Capital*, D. Usher (ed.), Chicago: The University of Chicago Press.
- [90] Diewert, W.E. (1980), "Aggregation Problems in the Measurement of Capital", pp.433-528 in *The Measurement of Capital*, edited by D. Usher, Studies in Income and Wealth, Vol. 45, National Bureau of Economic Research, University of Chicago Press, Chicago.
- [91] Diewert, W.E. (1981), "The Economic Theory of Index Numbers: A Survey", pp. 163-208 in *Essays in the Theory and Measurement of Consumer Behavior in Honour of sir Richard Stone*. A. Deaton (ed.), London: Cambridge University Press.
- [92] Diewert, W.E. (1981), "The Comparative Statics of Industry Long Run Equilibrium", *The Canadian Journal of Economics* 14, 78-92.
- [93] Diewert, W.E. (1982), "Duality Approaches to Microeconomic Theory", pp. 535-599 in *Handbook of Mathematical Economics*, Volume 2, K.J. Arrow and M.D. Intriligator (eds.), Amsterdam: North-Holland.
- [94] Diewert, W.E. (1983), "The Measurement of Waste within the Production Sector of an Open Economy", *Scandinavian Journal of Economics* 85, 159-179.
- [95] Diewert, W.E. (1983a), "The Theory of the Cost of Living Index and the Measurement of Welfare Change", pp. 163-233 in *Price Level Measurement*, W.E. Diewert and C. Montmarquette (eds.), Ottawa: Statistics Canada, reprinted as pp. 79-147 in *Price Level Measurement*, W.E. Diewert (ed.), Amsterdam: North-Holland, 1990.
- [96] Diewert, W.E. (1983b), "The Theory of the Output Price Index and the Measurement of Real Output Change", pp. 1049-1113 in *Price Level Measurement*, W.E. Diewert and C. Montmarquette (eds.), Ottawa: Statistics Canada.

- 
- [97] Diewert, W.E. (1983), "The Theory of the Output Price Index and the Measurement of Real Output Change", pp. 1049-1113 in *Price Level Measurement*, editors W.E. Diewert and C. Montmarquette, Ottawa: Statistics Canada.
- [98] Diewert, W.E. (1983), "Cost-Benefit Analysis and Project Evaluation: A Comparison of Alternative Approaches", *Journal of Public Economics* 22, 265-302.
- [99] Diewert, W.E. (1984), "Sensitivity Analysis in Economics", *Computers and Operations Research* 11:2, 141-156.
- [100] Diewert, W.E. (1987), "The Effects of an Innovation: A Trade Theory Approach", *The Canadian Journal of Economics* 20, 694-714.
- [101] Diewert, W.E. (1992), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *Journal of Productivity Analysis* 3, 211-248.
- [102] Diewert, W.E. (1992a), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *Journal of Productivity Analysis* 3, 211-248.
- [103] Diewert, W.E. (1992b), "Exact and Superlative Welfare Change Indicators", *Economic Inquiry* 30, 565-582.
- [104] Diewert, W.E. (1992a), "The Measurement of Productivity", *Bulletin of Economic Research* 44:3, 163-198.
- [105] Diewert, W.E. (1992b), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *Journal of Productivity Analysis* 3, 211-248.
- [106] Diewert, W.E. (1993), "Symmetric Means and Choice under Uncertainty", pp. 355-433 in *Essays in Index Number Theory*, Volume 1 (W.E. Diewert and A.O. Nakamura editors), Amsterdam: North-Holland.
- [107] Diewert, W.E. (1993), "Symmetric Means and Choice Under Uncertainty", pp. 355-433 in *Essays in Index Number Theory, Volume I*, Contributions to Economic Analysis 217, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North Holland.
- [108] Diewert, W.E. (1993), "Duality Approaches to Microeconomic Theory", pp. 105-175 in *Essays in Index Number Theory*, Volume 1, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North-Holland. This paper is a rewrite of Diewert (1982) but it also includes proofs.
- [109] Diewert, W.E. (1993), "The Early History of Price Index Research", pp. 33-65 in *Essays in Index Number Theory*, Volume 1, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North-Holland.
- [110] Diewert, W.E. (1993a), "The Early History of Price Index Research", pp. 33-65 in *Essays in Index Number Theory*, Volume 1, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North-Holland.
- [111] Diewert, W.E. (1993b), "Duality Approaches To Microeconomic Theory", in *Essays in Index Number Theory*, pp. 105-175 in Volume I, Contributions to Economic Analysis 217, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North Holland.
- [112] Diewert, W.E. (1993c), "Symmetric Means and Choice under Uncertainty", pp. 355-433 in *Essays in Index Number Theory*, Volume 1, W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North-Holland.
- [113] Diewert, W.E. (1995), "On the Stochastic Approach to Index Numbers", Discussion Paper 95-31, Department of Economics, University of British Columbia, Vancouver, Canada.
- [114] Diewert, W.E. (1996), "Seasonal Commodities, High Inflation and Index Number Theory". Discussion Paper No. 96-06, Department of Economics, University of British Columbia, Vancouver, Canada, V6T 1Z1, January, available on the web at: <http://web.arts.ubc.ca/econ/diewert/Disc.htm>
- [115] Diewert, W.E. (1997), "Commentary on Mathew D. Shapiro and David W. Wilcox: Alternative Strategies for Aggregating Price in the CPI", *The Federal Reserve Bank of St. Louis Review*, Vol. 79:3, (May/June), 127-137.
- [116] Diewert, W.E. (1998), "Index Number Issues in the Consumer Price Index", *Journal of Economic Perspectives* 12:1 (Winter), 47-58.

- [117] Diewert, W.E. (1998), "High Inflation, Seasonal Commodities and Annual Index Numbers", *Macroeconomic Dynamics* 2, 456-471.
- [118] Diewert, W.E. (1999), "Index Number Approaches to Seasonal Adjustment", *Macroeconomic Dynamics* 3, 48-68.
- [119] Diewert, W.E. (2001), "The Consumer Price Index and Index Number Purpose", *Journal of Economic and Social Measurement* 27, 167-248.
- [120] Diewert, W.E. (2001), "Productivity Growth and the Role of Government", Discussion Paper No. 01-13, Department of Economics, The University of British Columbia, Vancouver, Canada, V6T 1Z1. <http://www.econ.ubc.ca/discpapers/dp0113.pdf>
- [121] Diewert, W.E. (2001), "Measuring the Price and Quantity of Capital Services Under Alternative Assumptions", Discussion Paper 01-24, Department of Economics, University of British Columbia, Vancouver, Canada, June.
- [122] Diewert, W.E. (2002), "The Quadratic Approximation Lemma and Decompositions of Superlative Indexes", *Journal of Economic and Social Measurement* 28, 63-88.
- [123] Diewert, W.E. (2004), "On the Stochastic Approach to Linking the Regions in the ICP", Department of Economics, Discussion Paper 04-16, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z1.
- [124] Diewert, W.E. (2004), "A New Axiomatic Approach to Index Number Theory", Discussion Paper 04-05, Department of Economics, University of British Columbia, Vancouver, Canada, V6T 1Z1.
- [125] Diewert, W.E. (2004a), "Measuring Capital", Discussion Paper 04-10, Department of Economics, University of British Columbia, Vancouver, Canada, July.
- [126] Diewert, W.E. (2004b), "Index Number Problems in the Measurement of Real Net Exports and Real Net Changes in Inventories", paper presented at the Bureau of Economic Analysis, Washington D.C., September 21.
- [127] Diewert, W.E. (2005), "Weighted Country Product Dummy Variable Regressions and Index Number Formulae", *The Review of Income and Wealth* 51:4, 561-571.
- [128] Diewert, W.E. (2005a), "Issues in the Measurement of Capital Services, Depreciation, Asset Price Changes and Interest Rates", pp. 479-542 in *Measuring Capital in the New Economy*, C. Corrado, J. Haltiwanger and D. Sichel (eds.), Chicago: University of Chicago Press.
- [129] Diewert, W.E. (2005b), "On Measuring Inventory Change in Current and Constant Dollars", Discussion Paper 05-12, Department of Economics, University of British Columbia, Vancouver, Canada, August.
- [130] Diewert, W.E. (2005b), "Accounting Theory and Alternative Methods of Asset Valuation", Chapter 3 of a *Tutorial The Measurement of Business Capital, Income and Performance* presented at the University Autònoma of Barcelona, Spain, September 21-22, 2005; revised December 2005.
- [131] Diewert, W.E. (2006a), "Capital and Accounting Theory: The Early History", Chapter 2 of a *Tutorial The Measurement of Business Capital, Income and Performance* presented at the University Autònoma of Barcelona, Spain, September 21-22, 2005; revised February 2006.
- [132] Diewert, W.E. (2006b), "The Measurement of Income", Chapter 7 of a *Tutorial The Measurement of Business Capital, Income and Performance* presented at the University Autònoma of Barcelona, Spain, September 21-22, 2005; revised May 2006.
- [133] Diewert, W.E. (2009), "Similarity Indexes and Criteria for Spatial Linking", pp. 183-216 in *Purchasing Power Parities of Currencies: Recent Advances in Methods and Applications*, D.S. Prasada Rao (ed.), Cheltenham UK: Edward Elgar.
- [134] Diewert, W.E. (2012), *Consumer Price Statistics in the UK*, Government Buildings, Cardiff Road, Newport, UK, NP10 8XG: Office for National Statistics. <http://www.ons.gov.uk/ons/guide-method/userguidance/prices/cpi-and-rpi/index.html>
- [135] Diewert, W.E. (2013), "An Empirical Illustration of Index Construction using Israeli Data on Vegetables", paper presented at the 13th Meeting of the Ottawa Group On Prices at Copen-

- hagen, Denmark. May 2.
- [136] Diewert, W.E., M. Avriel and I. Zang (1981), "Nine Kinds of Quasiconcavity and Concavity", *Journal of Economic Theory* 25:3, 397-420.
- [137] Diewert, W.E. and K.J. Fox (1999), "Can Measurement Error Explain the Productivity Paradox?", *Canadian Journal of Economics* 32, 251-280. Also available at: <http://web.arts.ubc.ca/econ/diewert/hmpgdie.htm>
- [138] Diewert, W.E. and K.J. Fox (2004), "On the Estimation of Returns to Scale, Technical Progress and Monopolistic Markups", Discussion Paper 04-09, Department of Economics, University of British Columbia, July.
- [139] Diewert, W.E. and K.J. Fox (2005), "The New Economy and an Old Problem: Net Versus Gross Output", Center for Applied Economic Research Working Paper 2005/02, University of New South Wales, January.
- [140] Diewert, W.E. and R.J. Hill (2009), "Comment on Different Approaches to Index Number Theory", Discussion Paper 09-05, Department of Economics, University of British Columbia, Vancouver, Canada, V6T 1Z1
- [141] Diewert, W.E. and D. Lawrence, (1994), *The Marginal Costs of Taxation in New Zealand*, Report prepared for the New Zealand Business Roundtable by Swan Consultants, Canberra.
- [142] Diewert, W.E. and D.A. Lawrence (2000), "Progress in Measuring the Price and Quantity of Capital", pp. 273-326 in *Econometrics and the Cost of Capital: Essays in Honor of Dale W. Jorgenson*, L.J. Lau (ed.), Cambridge MA: The MIT Press.
- [143] Diewert, W.E. and D. Lawrence (2002), "The Deadweight Costs of Capital Taxation in Australia", pp. 103-167 in *Efficiency in the Public Sector*, Kevin J. Fox (ed.), Boston: Kluwer Academic Publishers.
- [144] Diewert, W.E. and D. Lawrence (2006), *Measuring the Contributions of Productivity and Terms of Trade to Australia's Economic Welfare*, Report by Meyrick and Associates to the Productivity Commission, Canberra, Australia.
- [145] Diewert, W.E. and D. Lawrence (2006), *Measuring the Contributions of Productivity and Terms of Trade to Australia's Economic Welfare*, Consultancy Report to the Productivity Commission, Australian Government, Canberra, March.
- [146] Diewert, W.E. and C.J. Morrison (1986), "Adjusting Output and Productivity Indexes for Changes in the Terms of Trade", *Economic Journal* 96, 659-679.
- [147] Diewert, W.E. H Mizobuchi, K Nomura (2005), "On Measuring Japan's Productivity, 1955-2003", Discussion Paper 05-05, Department of Economics, University of British Columbia, Vancouver, Canada.
- [148] Diewert, W.E. and A.O. Nakamura (1999), "Benchmarking and the Measurement of Best Practice Efficiency: An Electricity Generation Application", *Canadian Journal of Economics* 32, 570-588.
- [149] Diewert, W.E. and A.O. Nakamura (2003), "Index Number Concepts, Measures and Decompositions of Productivity Growth", *Journal of Productivity Analysis* 19, 127-159.
- [150] Diewert, W.E. and A.M. Smith (1994), "Productivity Measurement for a Distribution Firm", *The Journal of Productivity Analysis* 5, 335-347.
- [151] Diewert, W.E., A.H. Turunen-Red and A.D. Woodland (1989), "Productivity and Pareto Improving Changes in Taxes and Tariffs", *Review of Economic Studies* 56, 199-216.
- [152] Diewert, W.E., A.H. Turunen-Red and A.D. Woodland (1991), "Tariff Reform in a Small Open Multi-Household Economy with Domestic Distortions and Nontraded Goods", *International Economic Review* 32, 937-957.
- [153] Diewert, W.E. and T.J. Wales (1987), "Flexible Functional Forms and Global Curvature Conditions", *Econometrica* 55, 43-68.
- [154] Diewert, W.E. and T.J. Wales (1988a), "Normalized Quadratic Systems of Consumer Demand Functions", *Journal of Business and Economic Statistics* 6, 303-12.
- [155] Diewert, W.E. and T.J. Wales (1988b), "A Normalized Quadratic Semiflexible Functional

- Form", *Journal of Econometrics* 37, 327-42.
- [156] Diewert, W.E. and T.J. Wales (1992), "Quadratic Spline Models for Producer's Supply and Demand Functions", *International Economic Review* 33, 705-722.
- [157] Diewert, W.E. and T.J. Wales (1993), "Linear and Quadratic Spline Models for Consumer Demand Functions", *Canadian Journal of Economics* 26, 77-106.
- [158] Diewert, W.E. and A.D. Woodland (1977), "Frank Knight's Theorem in Linear Programming Revisited", *Econometrica* 45, 375-398.
- [159] Doms, M.E. (1996), "Estimating Capital Efficiency Schedules within Production Functions", *Economic Inquiry* 34, 78-92.
- [160] Dorfman, R., P.A. Samuelson and R.M. Solow (1958), *Linear Programming and Economic Analysis*, New York: McGraw-Hill Book Company.
- [161] Driffill, E.J. and H.S. Rosen (1983), "Taxation and Excess Burden: A Life Cycle Perspective", *International Economic Review* 24, 671-683.
- [162] Drobisch, M. W. (1871), "Ueber die Berechnung der Veränderungen der Waarenpreise und des Geldwerths", *Jahrbücher für Nationalökonomie und Statistik* 16, 143-156.
- [163] Dupor, B., L. Lochner, C. Taber, and M.B. Wittekind (1996), "Some Effects of Taxes on Schooling and Training", *American Economic Review* 86 (May), 340-346.
- [164] Edgeworth, F.Y. (1888), "The Mathematical Theory of Banking", *Journal of the Royal Statistical Society* 51, 113-127.
- [165] Edgeworth, F.Y. (1888), "Some New Methods of Measuring Variation in General Prices", *Journal of the Royal Statistical Society* 51, 346-368.
- [166] Edgeworth, F.Y. (1896), "A Defense of Index Numbers", *Economic Journal* 6, 132-142.
- [167] Edgeworth, F.Y. (1901), "Mr. Walsh on the Measurement of General Exchange Value", *Economic Journal* 11, 404-416.
- [168] Edwards, E.O. and P.W. Bell (1961), *The Theory and Measurement of Business Income*, Berkeley, California: University of California Press.
- [169] Eichhorn, W. (1978), *Functional Equations in Economics*, Reading, MA: Addison-Wesley Publishing Company.
- [170] Eichhorn, W. (1978), *Functional Equations in Economics*, London: Addison-Wesley.
- [171] Eichhorn, W. and J. Voeller (1976), *Theory of the Price Index*, Lecture Notes in Economics and Mathematical Systems, Vol. 140, Berlin: Springer-Verlag.
- [172] Epstein, L.G. (1977), *Essays in the Economics of Uncertainty*, unpublished Ph. D thesis, Vancouver: The University of British Columbia.
- [173] Eurostat (1993), *System of National Accounts 1993*, Brussels, Washington, Paris, New York and Washington: Eurostat, IMF, OECD, UN and World Bank.
- [174] Eurostat, International Monetary Fund, OECD, United Nations and World Bank (1993), *System of National Accounts 1993*, Luxembourg, New York, Paris, Washington DC.
- [175] Farkas, J. (1902), "Über die Theorie der einfachen Ungleichungen", *Journal für reine und angewandte Mathematik* 124, 1-24.
- [176] Feenstra, R.C. (2004), *Advanced International Trade: Theory and Evidence*, Princeton N.J.: Princeton University Press.
- [177] Feenstra, Robert C. and Matthew D. Shapiro (2003), "High-Frequency Substitution and the Measurement of Price Indexes", pp. 123-146 in *Scanner Data and Price Indexes*, Robert C. Feenstra and Matthew D. Shapiro (eds.), Studies in Income and Wealth Volume 64, Chicago: The University of Chicago Press.
- [178] Feldstein, M. (1996), "How Big Should Government Be?", Working Paper 5868, National Bureau of Economic Research, Cambridge, Massachusetts.
- [179] Fenchel, W. (1953), "Convex Cones, Sets and Functions", Lecture Notes at Princeton University, Department of Mathematics, Princeton, N.J.
- [180] Ferger, W.F. (1946), "Historical Note on the Purchasing Power Concept and Index Numbers",

- Journal of the American Statistical Association* 41, 53-57.
- [181] Fisher, F.M. and K. Shell (1972), "The Pure Theory of the National Output Deflator", pp. 49-113 in *The Economic Theory of Price Indexes*, New York: Academic Press.
- [182] Fisher, I. (1896), *Appreciation and Interest*, New York: Macmillan.
- [183] Fisher, I. (1897), "The Role of Capital in Economic Theory", *The Economic Journal* 7, 341-367.
- [184] Fisher, I. (1908), "Are Savings Income?", *Publications of the American Economic Association*, Third Series 9, 21-47.
- [185] Fisher, I. (1911), *The Purchasing Power of Money*, London: Macmillan.
- [186] Fisher, I. (1921), "The Best Form of Index Number", *Quarterly Publication of the American Statistical Association* 17, 533-537.
- [187] Fisher, I. (1922), *The Making of Index Numbers*, Houghton-Mifflin, Boston.
- [188] Fisher, I. (1922), *The Making of Index Numbers*, London: Macmillan and Company.
- [189] Fox, K.J. and U. Kohli (1998), "GDP Growth, Terms of Trade Effects and Total Factor Productivity", *The Journal of International Trade and Economic Development* 7:1, 87-110.
- [190] Fox, Kevin J. and Ulrich Kohli (1998) "GDP Growth, Terms-of-trade Effects, and Total Factor Productivity", *Journal of International Trade and Economic Development* 7, 87-110.
- [191] Frisch, R. (1930), "Necessary and Sufficient Conditions Regarding the Form of an Index Number which Shall Meet Certain of Fisher's Tests", *American Statistical Association Journal* 25, 397-406.
- [192] Frisch, R. (1936), "Annual Survey of General Economic Theory: The Problem of Index Numbers", *Econometrica* 4, 1-39.
- [193] Funke, H., G. Hacker and J. Voeller (1979), "Fisher's Circular Test Reconsidered", *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 115, 677-687.
- [194] Funke, H. and J. Voeller (1978), "A Note on the Characterization of Fisher's Ideal Index," pp. 177-181 in *Theory and Applications of Economic Indices*, W. Eichhorn, R. Henn, O. Opitz and R.W. Shephard (eds.), Würzburg: Physica-Verlag.
- [195] Funke, H., and J. Voeller (1979), "Characterization of Fisher's Ideal Index by Three Reversal Tests", *Statistische Hefte* 20, 54-60.
- [196] Gale, D, V.L. Klee and R.T. Rockafellar (1968), "Convex Functions on Convex Polytopes", *Proceedings of the American Mathematical Society* 19, 867-873.
- [197] Garcke, E. and J.M. Fells (1893), *Factory Accounts: Their Principles and Practice*, Fourth Edition, (First Edition 1887), London: Crosby, Lockwood and Son.
- [198] Gilman, S. (1939), *Accounting Concepts of Profit*, New York: The Rolland Press Co.
- [199] Goldman, A.J. and A.W. Tucker (1956), "Theory of Linear Programming", pp. 53-97 in *Linear Inequalities and Related Systems*, H.W. Kuhn and A.W. Tucker (eds.), Princeton N.J.: Princeton University Press.
- [200] Golub, G. H. and C. F. Van Loan (1980), "An Analysis of the Total Least Squares Problem", *Siam Journal of Numerical Analysis* 17, 883-893.
- [201] Gorman, W.M. (1968), "Measuring the Quantities of Fixed Factors", pp. 141-172 in *Value, Capital and Growth: Papers in Honour of Sir John Hicks*, J.N. Wolfe (ed.), Chicago: Aldine.
- [202] Green, J.B. (1915), "The Perpetual Inventory in Practical Stores Operation", *The Engineering Magazine* 48, 879-888.
- [203] Griliches, Z. (1963), "Capital Stock in Investment Functions: Some Problems of Concept and Measurement", pp. 115-137 in *Measurement in Economics*, C. Christ and others (eds.), Stanford California: Stanford University Press; reprinted as pp. 123-143 in *Technology, Education and Productivity*, Z. Griliches (ed.), (1988), Oxford: Basil Blackwell.
- [204] Griliches, Z. (1980), "Returns to Research and Development Expenditures in the Private Sector", pp. 419-454 in *New Developments in Productivity Measurement*, J.W. Kendrick and B. Vaccara (eds.), Studies in Income and Wealth, Volume 44, Chicago: University of Chicago Press.

- [205] Hadar, E. and S. Peleg (1998), "Efforts to Present Useful National Accounts under High Inflation", paper presented at the 25th General Conference of The International Association for Research in Income and Wealth, Cambridge, UK, August 23-29.
- [206] Hadley, G. and T.M. Whitin (1963), *Analysis of Inventory Systems*, Englewood Cliffs, N.J.: Prentice-Hall.
- [207] Haig, R.M. (1959), "The Concept of Income: Economic and Legal Aspects", pp. 54-76 in *Readings in the Economics of Taxation*, R.A. Musgrave and C.S. Shoup (eds.), Homewood, Illinois: Richard D. Irwin (Haig's chapter was originally published in 1921).
- [208] Hall, R.E. (1971), "the Measurement of Quality Change from Vintage Price Data", pp. 240-271 in *Price Indexes and Quality Change*, Z. Griliches (ed.), Cambridge Massachusetts: Harvard University Press.
- [209] Hall, R.E. (1988), "The Relationship between Price and Marginal Cost in U. S. Industry", *Journal of Political Economy* 96, 921-947.
- [210] Hall, R.E. (1990), "Invariance Properties of Solow's Productivity Residual", in *Growth, Productivity, Employment*, P. Diamond (ed.), Cambridge MA: MIT Press.
- [211] Haltiwanger, J. (2000), "Aggregate Growth: What Have we Learned from Microeconomic Evidence?" Economics Department Working Paper No. 267, Paris: OECD.
- [212] Harberger, Arnold (1998), "A Vision of the Growth Process," *American Economic Review* 88, 1-32.
- [213] Hardy, G.H., J.E. Littlewood and G. Polya, (1934), *Inequalities*, Cambridge, England: Cambridge University Press.
- [214] Harper, M.J., E.R. Berndt and D.O. Wood (1989), "Rates of Return and Capital Aggregation Using Alternative Rental Prices", pp. 331-372 in *Technology and Capital Formation*, D.W. Jorgenson and R. Landau (eds.), Cambridge MA: The MIT Press.
- [215] Harris, F.W. (1915), *Operations and Cost*, Chicago: A.W. Shaw Company.
- [216] Harris, R.G. (1999), "Making a Case for Tax Cuts", paper prepared for the Business Council on National Issues *Global Agenda Initiative*.
- [217] Harris, R.G. (2001), "Determinants of Canadian Productivity Growth: Issues and Prospects", Forthcoming in *Productivity Issues in a Canadian Context*, A. Sharpe and S. Rao (eds.), Montreal: McGill-Queen's Press.
- [218] Hayek, F.A. v. (1941), "Maintaining Capital Intact: A Reply", *Economica* 8, 276-280.
- [219] Hicks, J.R. (1939), *Value and Capital*, Oxford: The Clarendon Press.
- [220] Hicks, J.R. (1941-42), "Consumers' Surplus and Index Numbers", *The Review of Economic Studies* 9, 126-137.
- [221] Hicks, J.R. (1942), "Maintaining Capital Intact: a Further Suggestion", *Economica* 9, 174-179.
- [222] Hicks, J.R. (1946), *Value and Capital*, Second Edition, Oxford: Clarendon Press.
- [223] Hicks, J.R. (1961), "The Measurement of Capital in Relation to the Measurement of Other Economic Aggregates", pp. 18-31 in *The Theory of Capital*, F.A. Lutz and D.C. Hague (eds.), London: Macmillan.
- [224] Hicks, J. (1969), *A Theory of Economic History*, London: Oxford university Press.
- [225] Hicks, J. (1973), *Capital and Time: A Neo-Austrian Theory*, London: Oxford University Press.
- [226] Hicks, J. (1973), *Capital and Time*, Oxford: The Clarendon Press.
- [227] Hicks, J. (1973), *Capital and Time: A Neo-Austrian Theory*, Oxford: Clarendon Press.
- [228] Hill, P. (1996), *Inflation Accounting: A Manual on National Accounting under Conditions of High Inflation*, Paris: OECD.
- [229] Hill, P. (1999); "Capital Stocks, Capital Services and Depreciation"; paper presented at the third meeting of the Canberra Group on Capital Stock Statistics, Washington, D.C..
- [230] Hill, P. (2000); "Economic Depreciation and the SNA"; paper presented at the 26th Conference of the International Association for Research on Income and Wealth; Cracow, Poland.
- [231] Hill, R.J. (1999a), "Comparing Price Levels across Countries Using Minimum Spanning Trees", *The Review of Economics and Statistics* 81, 135-142.

- 
- [232] Hill, R.J. (1999b), "International Comparisons using Spanning Trees", pp. 109-120 in *International and Interarea Comparisons of Income, Output and Prices*, A. Heston and R.E. Lipsey (eds.), Studies in Income and Wealth Volume 61, NBER, Chicago: The University of Chicago Press.
- [233] Hill, R.J. (2001), "Measuring Inflation and Growth Using Spanning Trees", *International Economic Review* 42, 167-185.
- [234] Hill, R.J. (2004), "Constructing Price Indexes Across Space and Time: The Case of the European Union", *American Economic Review* 94, 1379-1410.
- [235] Hill, R.J. (2006), "Superlative Indexes: Not All of Them are Super", *Journal of Econometrics* 130, 25-43.
- [236] Hill, R.J. (2009), "Comparing Per Capita Income Levels Across Countries Using Spanning Trees: Robustness, Prior Restrictions, Hybrids and Hierarchies", pp. 217-244 in *Purchasing Power Parities of Currencies: Recent Advances in Methods and Applications*, D.S. Prasada Rao (ed.), Cheltenham UK: Edward Elgar.
- [237] Hill, R.J. and T.P. Hill (2003), "Expectations, Capital Gains and Income", *Economic Inquiry* 41, 607-619.
- [238] Hill, T.P. (1988), "Recent Developments in Index Number Theory and Practice", *OECD Economic Studies* 10, 123-148.
- [239] Hill, T.P. (1993), "Price and Volume Measures", pp. 379-406 in *System of National Accounts 1993*, Eurostat, IMF, OECD, UN and World Bank, Luxembourg, Washington, D.C., Paris, New York, and Washington, D.C.
- [240] Hoffman, A.J. (1953), "Cycling in the Simplex Algorithm", National Bureau of Standards Report 2974, December 16, 7 pages.
- [241] Hotelling, H. (1925), "A General Mathematical Theory of Depreciation", *Journal of the American Statistical Association* 20, 340-353.
- [242] Hotelling, H. (1932), "Edgeworth's Taxation Paradox and the Nature of Demand and Supply Functions", *Journal of Political Economy* 40, 577-616.
- [243] Hotelling, H. (1935), "Demand Functions with Limited Budgets", *Econometrica* 3, 66-78.
- [244] Hulten, C.R. (1990), "The Measurement of Capital", pp. 119-152 in *Fifty Years of Economic Measurement*, E.R. Berndt and J.E. Triplett (eds.), Studies in Income and Wealth, Volume 54, The National Bureau of Economic Research, Chicago: The University of Chicago Press.
- [245] Hulten, C.R. (1996), "Capital and Wealth in the Revised SNA", pp. 149-181 in *The New System of National Accounts*, J.W. Kendrick (ed.), New York: Kluwer Academic Publishers.
- [246] Hulten, C.R. and F.C. Wykoff (1981a), "The Estimation of Economic Depreciation using Vintage Asset Prices", *Journal of Econometrics* 15, 367-396.
- [247] Hulten, C.R. and F.C. Wykoff (1981b), "The Measurement of Economic Depreciation", pp. 81-125 in *Depreciation, Inflation and the Taxation of Income from Capital*, C.R. Hulten (ed.), Washington D.C.: The Urban Institute Press.
- [248] Hulten, C.R. and F.C. Wykoff (1996), "Issues in the Measurement of Economic Depreciation: Introductory Remarks", *Economic Inquiry* 34, 10-23.
- [249] ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004), *Consumer Price Index Manual: Theory and Practice*, Peter Hill (ed.), Geneva: International Labour Office.
- [250] Ivancic, L., W.E. Diewert and K.J. Fox (2009), "Scanner Data, Time Aggregation and the Construction of Price Indexes", Discussion Paper 09-09, Department of Economics, University of British Columbia, Vancouver, Canada.
- [251] Ivancic, L., W.E. Diewert and K.J. Fox (2011), "Scanner Data, Time Aggregation and the Construction of Price Indexes", *Journal of Econometrics* 161, 24-35.
- [252] Jevons, W.S., (1865), "The Variation of Prices and the Value of the Currency since 1782", *Journal of the Statistical Society of London* 28, 294-320; reprinted in *Investigations in Currency and Finance* (1884), London: Macmillan and Co., 119-150.
- [253] Jevons, W.S., (1884), "A Serious Fall in the Value of Gold Ascertained and its Social Effects

- Set Forth (1863)", pp. 13-118 in *Investigations in Currency and Finance*, London: Macmillan and Co.
- [254] Jorgenson, D.W. (1963), "Capital Theory and Investment Behaviour", *American Economic Review* 53:2, 247-259.
- [255] Jorgenson, D.W. (1989), "Capital as a Factor of Production", pp. 1-35 in *Technology and Capital Formation*, D.W. Jorgenson and R. Landau (eds.), Cambridge MA: The MIT Press.
- [256] Jorgenson, D.W. (1996a), "Empirical Studies of Depreciation", *Economic Inquiry* 34, 24-42.
- [257] Jorgenson, D.W. (1996b), *Investment: Volume 2; Tax Policy and the Cost of Capital*, Cambridge, Massachusetts: The MIT Press.
- [258] Jorgenson, D.W. and Z. Griliches (1967). "The Explanation of Productivity Change", *Review of Economic Studies* 34, 249-283.
- [259] Jorgenson, D.W., and Z. Griliches (1972), "Issues of Growth Accounting: A Reply to Edward F. Denison", *Survey of Current Business* 55(5), part II, 65-94.
- [260] Jorgenson, D.W. and Z. Griliches (1972), "Issues in Growth Accounting: A Reply to Edward F. Denison", *Survey of Current Business* 52:4, Part II (May), 65-94.
- [261] Jorgenson, D.W. and M. Nishimizu (1978), "U.S. and Japanese Economic Growth, 1952-1974", *Economic Journal* 88, 707-726.
- [262] Jorgenson, D.W. and K.-Y. Yun (1986), "Tax Policy and Capital Allocation", *Scandinavian Journal of Economics* 88, 355-377.
- [263] Jorgenson, D.W. and K.-Y. Yun (1990), "Tax Reform and US Economic Growth", *Journal of Political Economy* 98(5), S151-S193.
- [264] Jorgenson, D.W. and K.-Y. Yun (1991), *Tax Reform and the Cost of Capital*, Oxford: Clarendon Press.
- [265] Kaldor, N. (1972), "The Irrelevance of Equilibrium Economics", *The Economic Journal* 82, 1237-1255.
- [266] Kendall, M.G. and A.S. Stuart (1967), *The Advanced Theory of Statistics: Volume 2: Inference and Relationship*, Second Edition, New York: Hafner Publishing Co.
- [267] Kesselman, J.R. (1997), *General Payroll Taxes: Economics, Politics and Design*, Canadian Tax Paper No. 101, Toronto: The Canadian Tax Foundation.
- [268] Kesselman, J.R. (2000), "Flat Taxes, Dual Taxes, Smart Taxes: Making the Best Choices", *Policy Matters*, Volume 1, no. 7, Montreal: The Institute for Research On Public Policy. Email: irpp@irpp.org
- [269] Keynes, J.M. (1930), *Treatise on Money*, Vol. 1, London: Macmillan.
- [270] Kneller, R., M.F. Bleaney and N. Gemmell (1999), "Fiscal Policy and Growth: Evidence from OECD Countries", *Journal of Public Economics* 74:2, 171-190.
- [271] Knibbs, Sir G.H. (1924), "The Nature of an Unequivocal Price Index and Quantity Index", *Journal of the American Statistical Association* 19, 42-60 and 196-205.
- [272] Kohli, U.R.J. (1978), "A Gross National Product Function and the Derived Demand for Imports and Supply of Exports", *Canadian Journal of Economics* 11, 167-182.
- [273] Kohli, Ulrich (1982) "Production Theory, Technological Change, and the Demand for Imports: Switzerland, 1948-1974", *European Economic Review* 18, 369-386.
- [274] Kohli, U. (1990), "Growth Accounting in the Open Economy: Parametric and Nonparametric Estimates", *Journal of Economic and Social Measurement* 16, 125-136.
- [275] Kohli, U. (1991), *Technology, Duality and Foreign Trade: The GNP Function Approach to Modelling Imports and Exports*, Ann Arbor, MI: University of Michigan Press.
- [276] Kohli, U. (2003), "Growth Accounting in the Open Economy: International Comparisons", *International Review of Economics and Finance* 12, 417-435.
- [277] Kohli, U. (2004a), "An Implicit Törnqvist Index of Real GDP", *Journal of Productivity Analysis* 21, 337-353.
- [278] Kohli, U. (2004b), "Real GDP, Real Domestic Income and Terms of Trade Changes", *Journal of International Economics* 62, 83-106.

- [279] Kolmogoroff, A. (1930), "Sur la notion de la moyenne", *Atti della Reale Accademia nazionale dei Lincei* 12(6), 388-391.
- [280] Konüs, A.A. (1924), "The Problem of the True Index of the Cost of Living", translated in *Econometrica* 7, (1939), 10-29.
- [281] Konüs, A.A. (1939), "The Problem of the True Index of the Cost of Living", *Econometrica* 7, 10-29. [Originally published in 1924]
- [282] Konüs, A.A. and S.S. Byushgens (1926), "K probleme pokupatelnoi cili deneg", *Voprosi Konyunkturi* 2, 151-172.
- [283] Krugman, P. (1991), *Geography and Trade*, Cambridge, MA: The MIT Press.
- [284] Lachman, L.M. (1941), "On the Measurement of Capital", *Economica* 8, 361-377.
- [285] Laspeyres, E. (1871), "Die Berechnung einer mittleren Waarenpreissteigerung", *Jahrbücher für Nationalökonomie und Statistik* 16, 296-314.
- [286] Lau, L. (1976), "A Characterization of the Normalized Restricted Profit Function", *Journal of Economic Theory*, 12:1, 131-163.
- [287] Lau, L.J. (1979), "On Exact Index Numbers", *Review of Economics and Statistics* 61, 73-82.
- [288] Lehr, J. (1885), *Beitrage zur Statistik der Preise*, Frankfurt: J.D. Sauerlander.
- [289] Leontief, W.W. (1936), "Quantitative Input and Output Relations in the Economic System of the United States", *Review of Economic Statistics* 18, 105-125.
- [290] Leontief, W.W. (1941), *The Structure of the American Economy 1919-1929*, Cambridge, MA: Harvard University Press.
- [291] L'Hospital (1696), *L'analyse des infiniment petits pour l'intelligence des lignes courbes*, Paris.
- [292] Lipsey, R.G. (2000), "Economies of Scale in Theory and Practice", unpublished paper available at: <http://www.sfu.ca/~rlipsey/res.html>
- [293] Lipsey, R.G. and K. Carlaw (2000), "What does Total Factor Productivity Measure?", unpublished paper available at: <http://www.sfu.ca/~rlipsey/res.html>
- [294] Littleton, A.C. (1933), "Socialized Accounts", *The Accounting Review* 8, 267-271.
- [295] Lowe, J. (1823), *The Present State of England in Regard to Agriculture, Trade and Finance*, second edition, London: Longman, Hurst, Rees, Orme and Brown.
- [296] Luenberger, D.G. (1968), "Quasi-Convex Programming", *SIAM Journal of Applied Mathematics* 16, 1090-1095.
- [297] Madansky, A. (1959), "The Fitting of Straight Lines when both Variables are Subject to Error", *Journal of the American Statistical Association* 54, 173-206.
- [298] Maddison, A. (1993), "Standardized Estimates of Fixed Capital Stock: A Six Country Comparison", *Innovazione e materie prime*, April, 1-29.
- [299] Malinvaud, E. (1953), "Capital Accumulation and the Efficient Allocation of Resources", *Econometrica* 21, 233-268.
- [300] Malinvaud, E. (1967), "Decentralized Procedures for Planning", in *Activity Analysis in the Theory of Growth and Planning*, E. Malinvaud and M.O.L. Bacharach (eds.), London: Macmillan.
- [301] Malmquist, S. (1953) "Index Numbers and Indifference Surfaces", *Trabajos de Estadística* 4, 209-242.
- [302] Mangasarian, O. (1969), *Nonlinear Programming*, New York: McGraw-Hill.
- [303] Marshall, A. (1887), "Remedies for Fluctuations of General Prices", *Contemporary Review* 51, 355-375.
- [304] Marshall, A. (1890), *Principles of Economics*, London: Macmillan.
- [305] Marshall, A. (1898), *Principles of Economics*, Fourth Edition (first edition 1890, eighth edition 1920), London: The Macmillan Co.
- [306] Matheson, E. (1910), *Depreciation of Factories, Mines and Industrial Undertakings and their Valuations*, Fourth Edition, (First Edition 1884), London: Spon.
- [307] McFadden, D. (1966), "Cost, Revenue and Profit Functions: A Cursory Review", IBER Working Paper No. 86, University of California, Berkeley.
- [308] McFadden, D. (1978), "Cost, Revenue and Profit Functions", pp. 3-109 in *Production Eco-*

- nomics: A Dual Approach*, Volume 1, M. Fuss and D. McFadden (eds.), Amsterdam: North-Holland.
- [309] McFadden, D. (1978), "Cost, Revenue and Profit Functions", pp. 3-109 in *Production Economics: A Dual Approach to Theory and Applications*. Volume 1, M. Fuss and D. McFadden (eds.), Amsterdam: North-Holland.
- [310] McKenzie, L.W. (1956-7), "Demand Theory without a Utility Index", *Review of Economic Studies* 24, 184-189.
- [311] Middleditch, L. (1918), "Should Accounts Reflect the Changing Value of the Dollar?", *The Journal of Accountancy* 25. 114-120.
- [312] Minkowski, H. (1911), *Theorie der konvexen Körper*, Gesammelte Abhandlungen, Zweiter Band, Berlin.
- [313] Mintz, J.M. (1999), *Why Canada Must Undertake Business Tax Reform Soon*, Backgrounder, Toronto: C.D. Howe Institute.
- [314] Morrison, C. and W.E. Diewert (1990), "Productivity Growth and Changes in the Terms of Trade in Japan and the United States", pp. 201-227 in *Productivity Growth in Japan and the United States*, C.R. Hulten (ed.), University of Chicago Press, Chicago.
- [315] Motzkin, T.S. (1936), *Beiträge zur Theorie der linearen Ungleichungen*, Doctoral Thesis, University of Basel, Jerusalem: Azriel.
- [316] Nagumo, M. (1930), "Über eine Klasse der Mittelwerte", *Japanese Journal of Mathematics* 7, 71-79.
- [317] Nakajima, T., A. Nakamura and M. Nakamura (2002), "Japanese TFP Growth before and after the Financial Bubble: Japanese Manufacturing Industries", paper presented at the NBER, Cambridge MA, July 26, 2002.
- [318] Nakajima, T., M. Nakamura and K. Yoshioka (1998), "An Index Number Method for Estimating Scale Economies and Technical Progress Using Time-Series of Cross-Section Data: Sources of Total Factor Productivity Growth for Japanese Manufacturing, 1964-1988", *Japanese Economic Review* 49, 310-334.
- [319] Nakamura, A.O. and W.E. Diewert (2000), "Insurance for the Unemployed: Canadian Reforms and their Relevance for the United States", pp. 217-247 in *Long-Term Unemployment and Reemployment Policies*, L.J. Bassi and S.A. Woodbury (eds.), Stamford Connecticut: JAI Press.
- [320] Nakamura, A.O. and P. Lawrence (1994), "Education, Training and Prosperity", John Deutsch Institute for the Study of Economic Policy (March), 235-279.
- [321] Nordhaus, W.D. (1969), "Theory of Innovations: An Economic Theory of Technological Change", *American Economic Review* 59 (May), 18-28.
- [322] Norman, R.G. and S. Bahiri (1972), *Productivity Measurement and Incentives*, Oxford: Butterworth-Heinemann.
- [323] OECD (1993), *Methods Used by OECD Countries to Measure Stocks of Fixed Capital. National Accounts: Sources and Methods No. 2*, Paris: Organisation of Economic Co-operation and Development.
- [324] Oliner, S.D. (1996), "New Evidence on the Retirement and Depreciation of Machine Tools", *Economic Inquiry* 34, 57-77.
- [325] Paasche, H. (1874), "Über die Preisentwicklung der letzten Jahre nach den Hamburger Borsnotirungen", *Jahrbücher für Nationalökonomie und Statistik* 12, 168-178.
- [326] Pierson, N.G. (1896), "Further Considerations on Index-Numbers," *Economic Journal* 6, 127-131.
- [327] Pigou, A.C. (1924), *The Economics of Welfare*, Second Edition, London: Macmillan.
- [328] Pigou, A.C. (1935), "Net Income and Capital Depletion", *The Economic Journal* 45, 235-241.
- [329] Pigou, A.C. (1941), "Maintaining Capital Intact", *Economica* 8, 271-275.
- [330] Pollak, R.A. (1969), "Conditional Demand Functions and Consumption Theory", *Quarterly*

- Journal of Economics* 83, 60-78.
- [331] Pollak, R.A. (1983), "The Theory of the Cost-of-Living Index", pp. 87-161 in *Price Level Measurement*, W.E. Diewert and C. Montmarquette (eds.), Ottawa: Statistics Canada; reprinted as pp. 3-52 in R.A. Pollak, *The Theory of the Cost-of-Living Index*, Oxford: Oxford University Press, 1989.
- [332] Pollak, R.A. (1989), *The Theory of the Cost-of-Living Index*, Oxford: Oxford University Press.
- [333] Ponstein, J. (1967), "Seven Kinds of Convexity", *SIAM Review* 9, 115-119.
- [334] Roberts, A.W. and D.E. Varberg (1973), *Convex Functions*, New York: Academic Press.
- [335] Rockafellar, R.T. (1970), *Convex Analysis*, Princeton, N.J.: Princeton University Press.
- [336] Romer, P. (1994), "New Goods, Old Theory and the Welfare Costs of Trade Restrictions", *Journal of Development Economics* 43, 5-38.
- [337] Rudin, W., (1953), *Principles of Mathematical Analysis*, New York: McGraw-Hill Book Co.
- [338] Rymes, T.K. (1968), "Professor Read and the Measurement of Total Factor Productivity", *The Canadian Journal of Economics* 1, 359-367.
- [339] Rymes, T.K. (1983), "More on the Measurement of Total Factor Productivity", *The Review of Income and Wealth* 29 (September), 297-316.
- [340] Samuelson, P.A. (1947), *Foundations of Economic Analysis*, Cambridge, MA: Harvard University Press.
- [341] Samuelson, P.A. (1951), "Abstract of a Theorem Concerning Substitutability in Open Leontief Models", Chapter 7 in *Activity Analysis of Production and Allocation*, T.C. Koopmans (ed.), New York: John Wiley.
- [342] Samuelson, P.A. (1953), "Prices of Factors and Goods in General Equilibrium", *Review of Economic Studies* 21, 1-20.
- [343] Samuelson, P.A. (1953-54), "Prices of Factors and Goods in General Equilibrium", *Review of Economic Studies* 21, 1-20.
- [344] Samuelson, P.A. (1958), "Frank Knight's Theorem in Linear Programming", *Zeitschrift für National Ökonomie* 18, 310-317.
- [345] Samuelson, P.A. (1961), "The Evaluation of 'Social Income': Capital Formation and Wealth", pp. 32-57 in *The Theory of Capital*, F.A. Lutz and D.C. Hague (eds.), London: Macmillan.
- [346] Samuelson, P.A. (1967), "The Monopolistic Competition Revolution", In *Monopolistic Competition Theory: Studies in Impact*, R.E. Kuenne (ed.), New York: John Wiley.
- [347] Samuelson, P.A. (1974), "Complementarity—An Essay on the 40th Anniversary of the Hicks-Allen Revolution in Demand Theory", *The Journal of Economic Literature* 12, 1255-1289.
- [348] Samuelson, P.A. and S. Swamy (1974), "Invariant Economic Index Numbers and Canonical Duality: Survey and Synthesis", *American Economic Review* 64, 566-593.
- [349] Sato, K. (1976), "The Meaning and Measurement of the Real Value Added Index", *Review of Economics and Statistics* 58, 434-442.
- [350] Schlömilch, O., (1858), "Über Mittelgrößen verschiedener Ordnungen", *Zeitschrift für Mathematik und Physik* 3, 308-310.
- [351] Schmalenbach, E. (1959), *Dynamic Accounting*, translated from the German 12th edition of 1955 (first edition 1919) by G.W. Murphy and K.S. Most, London: Gee and Company.
- [352] Schreyer, P. (2001), *OECD Productivity Manual: A Guide to the Measurement of Industry-Level and Aggregate Productivity Growth*, Paris: OECD.
- [353] Schwarz, H.A., (1885), "Über ein die Flächen Kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung", *Acta Societatis scientiarum Fennicae* 15, 315-362.
- [354] Selvanathan, E.A. and D.S. Prasada Rao (1994), *Index Numbers: A Stochastic Approach*, Ann Arbor: The University of Michigan Press.
- [355] Shephard, R.W. (1953), *Cost and Production Functions*, Princeton N.J.: Princeton University Press.
- [356] Shephard, R.W. (1953), *Theory of Cost and Production Functions*, Princeton University Press.

- [357] Shephard, R.W. (1967), "The Notion of a Production Function", *Unternehmensforschung* 11, 209-232.
- [358] Shephard, R.W. (1970), *Theory of Cost and Production Functions*, Princeton N.J.: Princeton University Press.
- [359] Sidgwick, H. (1883), *The Principles of Political Economy*, London: Macmillan.
- [360] Slater, M. (1951), "A Note on Motzkin's Transposition Theorem", *Econometrica* 19, 185-186.
- [361] Smith, A. (1963), *The Wealth of Nations*, Volume 1 (first published in 1776), Homewood, Illinois: Richard D. Irwin.
- [362] Solomons, D. (1961), "Economic and Accounting Concepts of Income", *The Accounting Review* 36, 374-383.
- [363] Solomons, D. (1968), "The Historical Development of Costing", pp. 3-49 in *Studies in Cost Analysis*, D. Solomons (ed.), Homewood Illinois: Richard D. Irwin.
- [364] Sterling, R.R. (1975), "Relevant Financial Reporting in an Age of Price Changes", *The Journal of Accountancy* 139 (February), 42-51.
- [365] Sweeney, H.W. (1934), "Approximations of Appraisal Values by Index Numbers", *Harvard Business Review* 13, 108-115.
- [366] Sweeney, H.W. (1935), "The Technique of Stabilized Accounting", *The Accounting Review* 10, 185-205.
- [367] Sweeney, H.W. (1964), *Stabilized Accounting*, New York: Holt, Rinehart and Winston (reissue of the 1936 original with a new foreword).
- [368] Szulc, B.J. (1983), "Linking Price Index Numbers," pp. 537-566 in *Price Level Measurement*, W.E. Diewert and C. Montmarquette (eds.), Ottawa: Statistics Canada.
- [369] Szulc, B.J. (1987), "Price Indices below the Basic Aggregation Level", *Bulletin of Labour Statistics* 2, 9-16.
- [370] *The Economist* (2000), "New Zealand's Economy", London, December 2.
- [371] Theil, H. (1967), *Economics and Information Theory*, Amsterdam: North-Holland Publishing.
- [372] Tobin, J. (1956), "The Interest Elasticity of Transactions Demand for Cash", *The Review of Economics and Statistics* 38, 241-247.
- [373] Törnqvist, L. (1936), "The Bank of Finland's Consumption Price Index", *Bank of Finland Monthly Bulletin* 10, 1-8.
- [374] Törnqvist, L. and E. Törnqvist (1937), "Vilket är förhållandet mellan finska markens och svenska kronans köpkraft?", *Ekonomiska Samfundets Tidskrift* 39, 1-39 reprinted as pp. 121-160 in *Collected Scientific Papers of Leo Törnqvist*, Helsinki: The Research Institute of the Finnish Economy, 1981.
- [375] Triplett, J.E. (1996), "Depreciation in Production Analysis and in Income and Wealth Accounts: Resolution of an Old Debate", *Economic Inquiry* 34, 93-115.
- [376] Tucker, A.W. (1956), "Dual Systems of Homogeneous Linear Relations", pp. 3-18 in *Linear Inequalities and Related Systems*, H.W. Kuhn and A.W. Tucker (eds.), Princeton N.J.: Princeton University Press.
- [377] Tweedie, D. and G. Whittington (1984), *The Debate on Inflation Accounting*, London: Cambridge University Press.
- [378] Uzawa, H. (1962), "Production Functions with Constant Elasticities of Substitution", *Review of Economic Studies* 29, 291-299.
- [379] Uzawa, H. (1964), "Duality Principles in the Theory of Cost and Production", *International Economic Review* 5, 291-299.
- [380] Vanoli, A. (1998), "Interest and Inflation Accounting", paper presented at the 25th General Conference of The International Association for Research in Income and Wealth, Cambridge, UK, August 23-29.
- [381] Van Slyke, R.M. (1968), *Mathematical Programming and Optimal Control Theory*, Operations Research Center Report 68-21, University of California, Berkeley, July.
- [382] Vartia, Y.O. (1976), "Ideal Log-Change Index Numbers", *Scandinavian Journal of Statistics*

- 3, 121-126.
- [383] Vogt, A. (1980), "Der Zeit und der Faktorkehrtest als 'Finders of Tests'", *Statistische Hefte* 21, 66-71.
- [384] Vogt, A. and J. Barta (1997), *The Making of Tests for Index Numbers*, Heidelberg: Physica-Verlag.
- [385] von Neumann, J. (1937), "Über ein Ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes", *Ergebnisse eines Mathematische Kolloquiums* 8, 73-83; translated as "A Model of General Economic Equilibrium", *Review of Economic Studies* (1945-6) 12, 1-9.
- [386] von Neumann, J. (1947), "On a Maximization Problem", manuscript, Institute for Advanced Studies, Princeton University, November.
- [387] Walras, L. (1954), *Elements of Pure Economics*, a translation by W. Jaffé of the Edition Définitive (1926) of the *Eléments d'économie pure*, first edition published in 1874, Homewood, Illinois: Richard D. Irwin.
- [388] Walsh, B. (2000), "The Role of Tax Policy in Ireland's Economic Renaissance", *Canadian Tax Journal* 48:3, 658-673.
- [389] Walsh, C.M. (1901), *The Measurement of General Exchange Value*, New York: Macmillan and Co.
- [390] Walsh, C.M. (1921), *The Problem of Estimation*, London: P.S. King & Son.
- [391] Walsh, C.M. (1921a), *The Problem of Estimation*, London: P.S. King & Son.
- [392] Walsh, C. M. (1921b), "Discussion", *Journal of the American Statistical Association* 17, 537-544.
- [393] Walsh, C.M. (1924), "Professor Edgeworth's Views on Index Numbers", *Quarterly Journal of Economics* 38, 500-519.
- [394] Walters, A.A. (1961), "Production and Cost Functions: An Econometric Survey", *Econometrica* 31, 1-66.
- [395] Watson, W. (1999), "Labour Day, From the Front Porch", *Financial Post*, Toronto, Canada, September 8.
- [396] Westergaard, H. (1890), *Die Grundzüge der Theorie der Statistik*, Jena: Fischer.
- [397] Weymark, J.A. (1979), "A Reconciliation of Recent Results in Optimal Taxation Theory", *Journal of Public Economics* 12, 171-189.
- [398] Whitin, T.M. (1952), "Inventory Control in Theory and Practice", *Quarterly Journal of Economics* 66, 502-521.
- [399] Whitin, T.M. (1957), *The Theory of Inventory Management*, Second edition, Princeton, N.J.: Princeton University Press.
- [400] Whittington, G. (1980), "Pioneers of Income Measurement and Price Level Accounting: A Review Article", *Accounting and Business Research* 10 (Spring), 232-240.
- [401] Whittington, G. (1992), "Inflation Accounting", pp. 400-402 in *The New Palgrave Dictionary of Money and Finance*, Volume 2, P. Newman, M. Milgate and J. Eatwell (eds.), London: Macmillan.
- [402] Wiley, D.E., W.H. Schmidt and W.J. Bramble (1973), "Studies of a Class of Covariance Structure Models", *Journal of the American Statistical Association* 68, 317-323.
- [403] Wold, H. (1944), "A Synthesis of Pure Demand Analysis; Part 3", *Skandinavisk Aktuarietidskrift* 27, 69-120.
- [404] Wold, H. (1953), *Demand Analysis*, New York: John Wiley.
- [405] Woodland, A.D. (1982), *International Trade and Resource Allocation*. Amsterdam: North Holland.
- [406] Wynne, M.A. (1997), "Commentary on Measuring short Run Inflation for Central Bankers", *Federal Reserve Bank of St. Louis Review* 79:3, 161-167.
- [407] Yaari, M.E. (1977), "A Note on Separability and Quasiconcavity", *Econometrica* 45, 1183-1186.

- [408] Young, A.A. (1928), "Increasing Returns and Economic Progress", *Economic Journal* 38, 527-542.
- [409] Zeitsch, J. and D. Lawrence (1996), "Decomposing Economic Inefficiency in Base-Load Power Plants", *The Journal of Productivity Analysis* 7, 359-378.