AN INTRODUCTION TO EFFICIENCY AND PRODUCTIVITY ANALYSIS

Second Edition

AN INTRODUCTION TO EFFICIENCY AND PRODUCTIVITY ANALYSIS

Second Edition

by

Timothy J. Coelli
D.S. Prasada Rao
Christopher J. O'Donnell
George E. Battese

Springer

Tim Coelli
University of Queensland
Australia

D.S. Prasada Rao
University of Queensland
Australia

Christopher J. O'Donnell
University of Queensland
Australia

George E. Battese
University of Queensland
Australia

Library of Congress Cataloging-in-Publication Data

An introduction to efficiency and productivity analysis / by Timothy Coelli ... [et al].—
2nd ed.
p. cm.
Rev. ed. of: An introduction to efficiency and productivity analysis / by Tim Coelli.
c1998.
Includes bibliographical references and index.
ISBN-10: 0-387-24265-1 ISBN-13: 978-0387-24265-1
ISBN-10: 0-387-24266-X ISBN-13: 978-0387-24266-8 (softcover)
e-ISBN-10: 0-387-25895-7
1. Production (Economic theory) 2. Production functions (Economic theory) 3. Industrial productivity. I. Coelli, Tim. II. Coelli, Tim. Introduction to efficiency and productivity analysis.

HB241.C64 2005
338'.06—dc22

2005042642

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 11053217

springeronline.com

To

Michelle, Visala, Adrienne and Marilyn

TABLE OF CONTENTS

FIGURES

TABLES

PREFACE

The second edition of this book has been written for the same audience as the first edition. It is designed to be a "first port of call" for people wishing to study efficiency and productivity analysis. The book provides an accessible introduction to the four principal methods involved: econometric estimation of average response models; index numbers; data envelopment analysis (DEA); and stochastic frontier analysis (SFA). For each method, we provide a detailed introduction to the basic concepts, give some simple numerical examples, discuss some of the more important extensions to the basic methods, and provide references for further reading. In addition, we provide a number of detailed empirical applications using real-world data.

The book can be used as a textbook or as a reference text. As a textbook, it probably contains too much material to cover in a single semester, so most instructors will want to design a course around a subset of chapters. For example, Chapter 2 is devoted to a review of production economics and could probably be skipped in a course for graduate economics majors. However, it should prove useful to undergraduate students and those doing a major in another field, such as business management or health studies.

There have been several excellent books written on performance measurement in recent years, including Färe, Grosskopf and Lovell (1985, 1994), Fried, Lovell and Schmidt (1993), Charnes et al (1995), Färe, Grosskopf and Russell (1998) and Kumbhakar and Lovell (2000). The present book is not designed to compete with these advanced-level books, but to provide a lower-level bridge to the material contained within them, as well as to many other books and journal articles written on this topic.

We believe this second edition remains a unique book in this field insofar as:

1. it is an introductory text;
2. it contains detailed discussion and comparison of the four principal methods for efficiency and productivity analysis; and

3. it provides detailed advice on computer programs that can be used to implement these methods. The book contains computer instructions and output listings for the SHAZAM, LIMDEP, TFPIP, DEAP and FRONTIER computer programs. More extensive listings of data and computer instruction files are available on the book website (www.uq.edu.au/economics/cepa/crob2005).

The first edition of this book was published in 1998. It grew out of a set of notes that were written for a series of short courses that the Centre for Efficiency and Productivity Analysis (CEPA) had designed for a number of government agencies in Australia in the mid 1990's. The success of the first edition was largely due to its focus on the provision of information for practitioners (rather than academic theorists), and also due to the valuable feedback and suggestions provided by those people who attended these early short courses.

In the subsequent years we have continued to present CEPA short courses to people in business and government, using the first edition as a set of course notes. However, in recent years we have noted that we have been supplying increasing quantities of "extra materials" at these courses, reflecting the number of significant advances that have occurred in this field since 1998. Hence, when the publisher approached us to write a second edition, we were keen to take the opportunity to update the book with this new material. We also took the opportunity to freshen some of the original material to reflect our maturing understanding of various topics, and to incorporate some of the excellent suggestions provided by many readers and short course participants over the past seven years.

Readers familiar with the first edition will notice a number of changes in this second edition. Structurally, the material included in various chapters has been re-organised to provide a more logical ordering of economic theory and empirical methods. A number of new empirical examples have also been provided. Separate chapters have now been devoted to data measurement issues (Chapter 5) and the econometric estimation of average response functions (Chapter 8).

Many other changes and additions have also been incorporated. For example, the parametric methods section has been updated to cover confidence intervals; testing and imposing regularity conditions; and Bayesian methods. The DEA section has been updated to cover weights restrictions; super efficiency; bootstrapping; short-run cost minimisation; and profit maximisation. Furthermore, the productivity growth section has been updated to cover the issues of shadow prices and scale effects.

We wish to thank the many people whose comments, feedback and discussions have contributed to improving our understanding of the material within this book. In particular we wish to thank our recent CEPA visitors: Knox Lovell, Bert Balk, Erwin Diewert, Rolf Färe and Shawna Grosskopf. Rolf and Shawna were visiting

during the final few weeks of writing, and were very generous with their time, reading a number of draft chapters and providing valuable comments.

Finally, we hope that you, the readers, continue to find this book useful in your studies and research, and we look forward to receiving your comments and feedback on this second edition.

Timothy J. Coelli
D.S. Prasada Rao
Christopher J. O'Donnell
George E. Battese

Centre for Efficiency and
Productivity Analysis
University of Queensland
Brisbane, Australia.

1. INTRODUCTION

1.1 Introduction

This book is concerned with measuring the performance of firms, which convert inputs into outputs. An example of a firm is a shirt factory that uses materials, labour and capital (inputs) to produce shirts (output). The performance of this factory can be defined in many ways. A natural measure of performance is a productivity ratio: the ratio of outputs to inputs, where larger values of this ratio are associated with better performance. Performance is a relative concept. For example, the performance of the factory in 2004 could be measured relative to its 2003 performance or it could be measured relative to the performance of another factory in 2004, etc.

The methods of performance measurement that are discussed in this book can be applied to a variety of "firms".[1] They can be applied to private sector firms producing goods, such as the factory discussed above, or to service industries, such as travel agencies or restaurants. The methods may also be used by a particular firm to analyse the relative performance of units within the firm (e.g., bank branches or chains of fast food outlets or retail stores). Performance measurement can also be applied to non-profit organisations, such as schools or hospitals.

---

[1]In some of the literature on productivity and efficiency analysis the rather ungainly term "decision making unit" (DMU) is used to describe a productive entity in instances when the term "firm" may not be entirely appropriate. For example, when comparing the performance of power plants in a multi-plant utility, or when comparing bank branches in a large banking organisation, the units under consideration are really *parts* of a firm rather than firms themselves. In this book we have decided to use the term "firm" to describe any type of decision making unit, and ask that readers keep this more general definition in mind as they read the remainder of this book.

All of the above examples involve micro-level data. The methods we consider can also be used for making performance comparisons at higher levels of aggregation. For example, one may wish to compare the performance of an industry over time or across geographical regions (e.g., shires, counties, cities, states, countries, etc.).

We discuss the use and the relative merits of a number of different performance measurement methods in this book. These methods differ according to the type of measures they produce; the data they require; and the assumptions they make regarding the structure of the production technology and the economic behaviour of decision makers. Some methods only require data on quantities of inputs and outputs while other methods also require price data and various behavioural assumptions, such as cost minimisation, profit maximisation, etc.

But before we discuss these methods any further, it is necessary for us to provide some informal definitions of a few terms. These definitions are not very precise, but they are sufficient to provide readers, new to this field, some insight into the sea of jargon in which we swim. Following this we provide an outline of the contents of the book and a brief summary of the principal performance measurement methods that we consider.

1.2 Some Informal Definitions

In this section we provide a few informal definitions of some of the terms that are frequently used in this book. More precise definitions will be provided later in the book. The terms are:

- productivity;
- technical efficiency;
- allocative efficiency;
- technical change;
- scale economies;
- total factor productivity (TFP);
- production frontier; and
- feasible production set.

We begin by defining the **productivity** of a firm as the ratio of the output(s) that it produces to the input(s) that it uses.

$$\text{productivity} = \text{outputs/inputs} \tag{1.1}$$

When the production process involves a single input and a single output, this calculation is a trivial matter. However, when there is more than one input (which is often the case) then a method for aggregating these inputs into a single index of inputs must be used to obtain a ratio measure of productivity.[2] In this book, we discuss some of the methods that are used to aggregate inputs (and/or outputs) for the construction of productivity measures.

When we refer to productivity, we are referring to **total factor productivity**, which is a productivity measure involving all factors of production.[3] Other traditional measures of productivity, such as labour productivity in a factory, fuel productivity in power stations, and land productivity (yield) in farming, are often called *partial* measures of productivity. These partial productivity measures can provide a misleading indication of overall productivity when considered in isolation.

The terms, **productivity** and **efficiency,** have been used frequently in the media over the last ten years by a variety of commentators. They are often used interchangeably, but this is unfortunate because they are not precisely the same things. To illustrate the distinction between the terms, it is useful to consider a simple production process in which a single input (x) is used to produce a single output (y). The line 0F′ in Figure 1.1 represents a **production frontier** that may be used to define the relationship between the input and the output. The production frontier represents the maximum output attainable from each input level. Hence it reflects the current state of technology in the industry. More is stated about its properties in later sections. Firms in this industry operate either on that frontier, if they are **technically efficient**, or beneath the frontier if they are not technically efficient. Point A represents an inefficient point whereas points B and C represent efficient points. A firm operating at point A is inefficient because technically it could increase output to the level associated with the point B without requiring more input.[4]

We also use Figure 1.1 to illustrate the concept of a **feasible production set** which is the set of all input-output combinations that are feasible. This set consists of all points between the production frontier, OF', and the x-axis (inclusive of these bounds).[5] The points along the production frontier define the efficient subset of this feasible production set. The primary advantage of the *set representation* of a production technology is made clear when we discuss multi-input/multi-output production and the use of distance functions in later chapters.

---

[2] The same problem occurs with multiple outputs.

[3] It also includes all outputs in a multiple-output setting.

[4] Or alternatively, it could produce the same level of output using less input (i.e., produce at point C on the frontier).

[5] Note that this definition of the production set assumes free disposability of inputs and outputs. These issues will be discussed further in subsequent chapters.

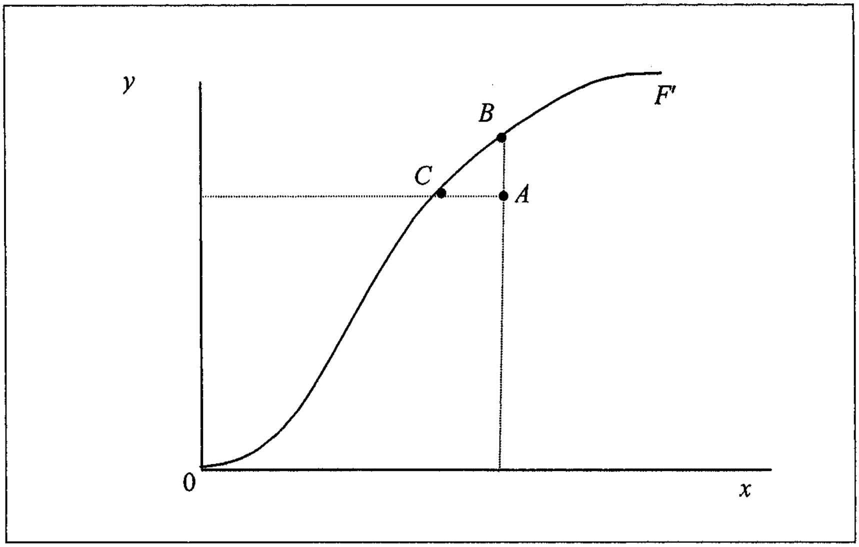

Figure 1.1 Production Frontiers and Technical Efficiency

To illustrate the distinction between technical efficiency and productivity we utilise Figure 1.2. In this figure, we use a ray through the origin to measure productivity at a particular data point. The slope of this ray is y/x and hence provides a measure of productivity. If the firm operating at point A were to move to the technically efficient point B, the slope of the ray would be greater, implying higher productivity at point B. However, by moving to the point C, the ray from the origin is at a tangent to the production frontier and hence defines the point of maximum possible productivity. This latter movement is an example of exploiting **scale economies**. The point C is the point of (technically) optimal scale. Operation at any other point on the production frontier results in lower productivity.

From this discussion, we conclude that a firm may be technically efficient but may still be able to improve its productivity by exploiting scale economies. Given that changing the scale of operations of a firm can often be difficult to achieve quickly, technical efficiency and productivity can in some cases be given short-run and long-run interpretations.

The discussion above does not include a time component. When one considers productivity comparisons through time, an additional source of productivity change, called **technical change,** is possible. This involves advances in technology that may be represented by an upward shift in the production frontier. This is depicted in Figure 1.3 by the movement of the production frontier from $0F_0'$ in period 0 to $0F_1'$ in period 1. In period 1, all firms can technically produce more output for each level of input, relative to what was possible in period 0. An example of technical change is the installation of a new boiler for a coal-fired power plant that extends the plant productivity potential beyond previous limits.[6]

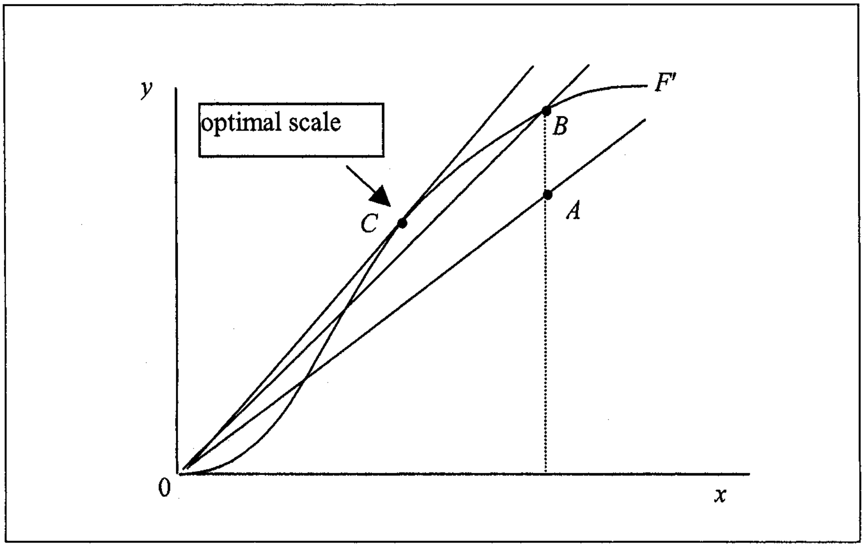

Figure 1.2 Productivity, Technical Efficiency and Scale Economies

When we observe that a firm has increased its productivity from one year to the next, the improvement need not have been from efficiency improvements alone, but may have been due to technical change or the exploitation of scale economies or from some combination of these three factors.

Up to this point, all discussion has involved physical quantities and technical relationships. We have not discussed issues such as costs or profits. If information on prices is available, and a behavioural assumption, such as cost minimisation or profit maximisation, is appropriate, then performance measures can be devised which incorporate this information. In such cases it is possible to consider **allocative efficiency,** in addition to technical efficiency. Allocative efficiency in input selection involves selecting that mix of inputs (e.g., labour and capital) that produces a given quantity of output at minimum cost (given the input prices which prevail). Allocative and technical efficiency combine to provide an overall economic efficiency measure.[7]

---

[6] This is an example of embodied technical change, where the technical change is embodied in the capital input. Disembodied technical change is also possible. One such example, is that of the introduction of legume/wheat crop rotations in agriculture in recent decades.

[7] In the case of a multiple-output industry, allocative efficiency in output mix may also be considered.

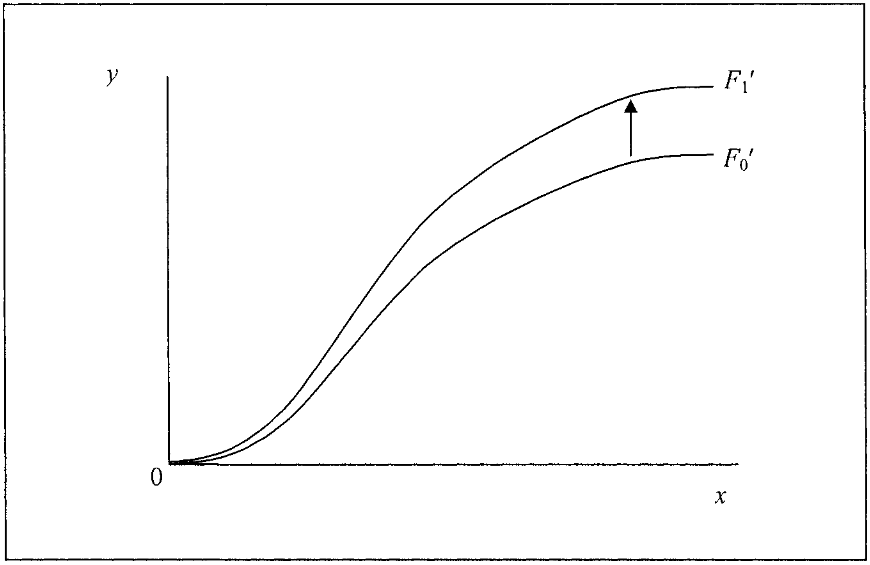

Figure 1.3 Technical Change Between Two Periods

Now that we are armed with this handful of informal definitions we briefly describe the layout of the book and the principal methods that we consider in subsequent chapters.

1.3 Overview of Methods

There are essentially four major methods discussed in this book:

1. least-squares econometric production models;
2. total factor productivity (TFP) indices;
3. data envelopment analysis (DEA); and
4. stochastic frontiers.

The first two methods are most often applied to aggregate time-series data and provide measures of technical change and/or TFP. Both of these methods assume all firms are technically efficient. Methods 3 and 4, on the other hand, are most often applied to data on a sample of firms (at one point in time) and provide measures of relative efficiency among those firms. Hence these latter two methods do not assume that all firms are technically efficient. However, multilateral TFP indices can also be used to compare the relative productivity of a group of firms at one point in time. Also DEA and stochastic frontiers can be used to measure both technical change and efficiency change, if panel data are available.

Thus we see that the above four methods can be grouped according to whether they recognise inefficiency or not. An alternative way of grouping the methods is to note that methods 1 and 4 involve the econometric estimation of parametric functions, while methods 2 and 3 do not. These two groups may therefore be termed "parametric" and "non-parametric" methods, respectively. These methods may also be distinguished in several other ways, such as by their data requirements, their behavioural assumptions and by whether or not they recognise random errors in the data (i.e. noise). These differences are discussed in later chapters.

1.4 Outline of Chapters

Summaries of the contents of the remaining 11 chapters are provided below.

Chapter 2. Review of Production Economics: This is a review of production economics at the level of an upper-undergraduate microeconomics course. It includes a discussion of the various ways in which one can provide a functional representation of a production technology, such as production, cost, revenue and profit functions, including information on their properties and dual relationships. We also review a variety of production economics concepts such as elasticities of substitution and returns to scale.

Chapter 3. Productivity and Efficiency Measurement Concepts: Here we describe how one can alternatively use set constructs to define production technologies analogous to those described using functions in Chapter 2. This is done because it provides a more natural way of dealing with multiple output production technologies, and allows us to introduce the concept of a distance function, which helps us define a number of our efficiency measurement concepts, such as technical efficiency. We also provide formal definitions of concepts such as technical efficiency, allocative efficiency, scale efficiency, technical change and total factor productivity (TFP) change.

Chapter 4. Index Numbers and Productivity Measurement: In this chapter we describe the familiar Laspeyres and Paasche index numbers, which are often used for price index calculations (such as a consumer price index). We also describe Tornqvist and Fisher indices and discuss why they may be preferred when calculating indices of input and output quantities and TFP. This involves a discussion of the economic theory that underlies various index number methods, plus a description of the various axioms that index numbers should ideally possess. We also cover a number of related issues such as chaining in time series comparisons and methods for dealing with transitivity violations in spatial comparisons.

Chapter 5. Data and Measurement Issues: In this chapter we discuss the very important topic of data set construction. We discuss a range of issues relating to the collection of data on inputs and outputs, covering topics such as quality variations; capital measurement; cross-sectional and time-series data; constructing implicit quantity measures using price deflated value aggregates; aggregation issues, international comparisons; environmental differences; overheads allocation; plus many more. The index number concepts introduced in Chapter 4 are used regularly in this discussion.

Chapter 6. Data Envelopment Analysis: In this chapter we provide an introduction to DEA, the mathematical programming approach to the estimation of frontier functions and the calculation of efficiency measures. We discuss the basic DEA models (input- and output- orientated models under the assumptions of constant returns to scale and variable returns to scale) and illustrate these methods using simple numerical examples.

Chapter 7. Additional Topics on Data Envelopment Analysis: Here we extend our discussion of DEA models to include the issues of allocative efficiency; short run models; environmental variables; the treatment of slacks; super-efficiency measures; weights restrictions; and so on. The chapter concludes with a detailed empirical application.

Chapter 8. Econometric Estimation of Production Technologies: In this chapter we provide an overview of the main econometric methods that are used for estimating economic relationships, with an emphasis on production and cost functions. Topics covered include selection of functional form; alternative estimation methods (ordinary least squares, maximum likelihood, nonlinear least squares and Bayesian techniques); testing and imposing restrictions from economic theory; and estimating systems of equations. Even though the econometric models in this chapter implicitly assume no technical inefficiency, much of the discussion here is also useful background for the stochastic frontier methods discussed in the following two chapters. Data on rice farmers in the Philippines is used to illustrate a number of models.

Chapter 9. Stochastic Frontier Analysis: This is an alternative approach to the estimation of frontier functions using econometric techniques. It has advantages over DEA when data noise is a problem. The basic stochastic frontier model is introduced and illustrated using a simple example. Topics covered include maximum likelihood estimation, efficiency prediction and hypothesis testing. The rice farmer data from Chapter 8 is used to illustrate a number of models.

Chapter 10. Additional Topics on Stochastic Frontier Analysis: In this chapter we extend the discussion of stochastic frontiers to cover topics such as allocative efficiency, panel data models, the inclusion of environmental and management variables, risk modeling and Bayesian methods. The rice farmer data from Chapter 8 is used to illustrate a number of models.

Chapter 11. The Calculation and Decomposition of Productivity Change using Frontier Methods: In this chapter we discuss how one may use frontier methods (such as DEA and stochastic frontiers) in the analysis of panel data for the purpose of measuring TFP growth. We discuss how the TFP measures may be decomposed into technical efficiency change and technical change. The chapter concludes with a detailed empirical application using the rice farmer data from Chapter 8, which raises various topics including the effects of data noise, shadow prices and aggregation.

Chapter 12. Conclusions.

1.5 What is Your Economics Background?

When writing this book we had two groups of readers in mind. The first group contains postgraduate economics majors who have recently completed a graduate course on microeconomics, while the second group contains people with less knowledge of microeconomics. This second group might include undergraduate students, MBA students and researchers in industry and government who do not have a strong economics background (or who did their economics training a number of years ago). The first group may quickly review Chapters 2 and 3. The second group of readers should read Chapters 2 and 3 carefully. Depending on your background, you may also need to supplement your reading with some of the reference texts that are suggested in these chapters.

2. REVIEW OF PRODUCTION ECONOMICS

2.1 Introduction

This chapter reviews key economic concepts needed for a proper understanding of efficiency and productivity measurement. To make the chapter accessible we have chosen to use functions and graphs, rather than sets,[1] to describe the technological possibilities faced by firms. To further simplify matters, we assume i) the production activities of the firm take place in a single period, ii) the prices of all inputs and outputs are known with certainty, and iii) the firm is technically efficient in the sense that it uses its inputs to produce the maximum outputs that are technologically feasible (this last assumption is relaxed in Chapter 3). In all these respects, our review of production economics is similar to that found in most undergraduate economics textbooks.

We begin, in Section 2.2, by showing how the production possibilities of single-output firms can be represented using production functions. We explain some of the properties of these functions (eg., monotonicity) and define associated quantities of economic interest (eg., elasticities of substitution). In Section 2.3, we show how the production possibilities of *multiple*-output firms can be represented using transformation functions. However, this section is kept brief, not least because transformation functions can be viewed as special cases of the distance functions discussed in detail in Chapter 3. In Section 2.4, we show how multiple-output technologies can also be represented using cost functions. We discuss the properties of these functions and show how they can be used to quickly and easily

[1] Set representations of production technologies are discussed in Chapter 3.

derive input demand functions (using Shephard's Lemma). In Section 2.5, we briefly consider an alternative but less common representation of the production technology, the revenue function. Finally, in Section 2.6, we discuss the profit function. Among other things, we show that profit maximisation implies both cost minimisation and revenue maximisation.

Much of the material presented in this chapter is drawn from the microeconomics textbooks by Call and Holahan (1983), Chambers, (1988), Beattie and Taylor (1985), Varian (1992) and Henderson and Quandt (1980). More details are available in these textbooks, and almost any other microeconomics textbooks used in undergraduate economics classes.

2.2 Production Functions

Consider a firm that uses amounts of N inputs (eg., labour, machinery, raw materials) to produce a single output. The technological possibilities of such a firm can be summarised using the production function[2]

$$q = f(\mathbf{x}) \tag{2.1}$$

where q represents output and $\mathbf{x} = (x_1, x_2, \ldots, x_N)'$ is an $N \times 1$ vector of inputs. Throughout this chapter we assume these inputs are within the effective control of the decision maker. Other inputs that are outside the control of the decision maker (eg., rainfall) are also important, but, for the time being, it is convenient to subsume them into the general structure of the function $f(.)$. A more explicit treatment of these variables is provided in Section 10.6.

2.2.1 Properties

Associated with the production function 2.1 are several properties that underpin much of the economic analysis in the remainder of the book. Principal among these are (eg., Chambers, 1988):

F.1 *Nonnegativity:* The value of $f(\mathbf{x})$ is a finite, non-negative, real number.

F.2 *Weak Essentiality:* The production of positive output is impossible without the use of at least one input.

F.3 *Nondecreasing in* $\mathbf{x}$: (or *monotonicity*) Additional units of an input will not decrease output. More formally, if $\mathbf{x}^0 \geq \mathbf{x}^1$ then

[2] Most economics textbooks refer to the technical relationship between inputs and output as a production *function* rather than a production *frontier*. The two terms can be used interchangeably. The efficiency measurement literature tends to use the term *frontier* to emphasise the fact that the function gives the *maximum* output that is technologically feasible.

$f(\mathbf{x}^0) \geq f(\mathbf{x}^1)$. If the production function is continuously differentiable, monotonicity implies all marginal products are non-negative.

F.4 *Concave in* $\mathbf{x}$: Any linear combination of the vectors $\mathbf{x}^0$ and $\mathbf{x}^1$ will produce an output that is no less than the same linear combination of $f(\mathbf{x}^0)$ and $f(\mathbf{x}^1)$. Formally[3], $f(\theta\mathbf{x}^0 + (1-\theta)\mathbf{x}^1) \geq \theta f(\mathbf{x}^0) + (1-\theta) f(\mathbf{x}^1)$ for all $0 \leq \theta \leq 1$. If the production function is continuously differentiable, concavity implies all marginal products are non-increasing (i.e., the well-known law of diminishing marginal productivity).

These properties are not exhaustive, nor are they universally maintained. For example, the monotonicity assumption is relaxed in cases where heavy input usage leads to *input congestion* (eg., when labour is hired to the point where "too many cooks spoil the broth"), and the weak essentiality assumption is usually replaced by a stronger assumption in situations where *every* input is essential for production.

To illustrate some of these ideas, Figure 2.1 depicts a production function defined over a single input, x. Notice that

- for the values of x represented on the horixontal axis, the values of q are all non-negative and finite real numbers. Thus, the function satisfies the non-negativity property F.1.
- the function passes through the origin, so it satisfies property F.2.
- the marginal product[4] of x is positive at all points between the origin and point G, implying the monotonicity property F.3 is satisfied at these points. However, monotonicity is violated at all points on the curved segment GR.
- as we move along the production function from the origin to point D, the marginal product of x increases. Thus, the concavity property F.4 is violated at these points. However, concavity is satisfied at all points on the curved segment DR.

In summary, the production function depicted in Figure 2.1 violates the concavity property in the region 0D and violates the monotonicity property in the region GR. However, it is consistent with all properties along the curved segment between points D and G – we refer to this as the *economically-feasible region* of production. Within this region, the point E is the point at which the average product[5] is maximised. We refer to this point as the *point of optimal scale* (of operations).

---

[3] For readers who are unfamiliar with vector algebra, when we pre-multiply a vector by a scalar we simply multiply every element of the vector by the scalar. For example, if $\mathbf{x} = (x_1, x_2, \ldots, x_N)'$ then $\theta\mathbf{x} = (\theta x_1, \theta x_2, \ldots, \theta x_N)'$.

[4] Graphically, the marginal product at a point is the slope of the production function at that point.

[5] In the case of a single-input production function the average product is AP = q/x. Graphically, the average product at a point is given by the slope of the ray that passes through the origin and that point.

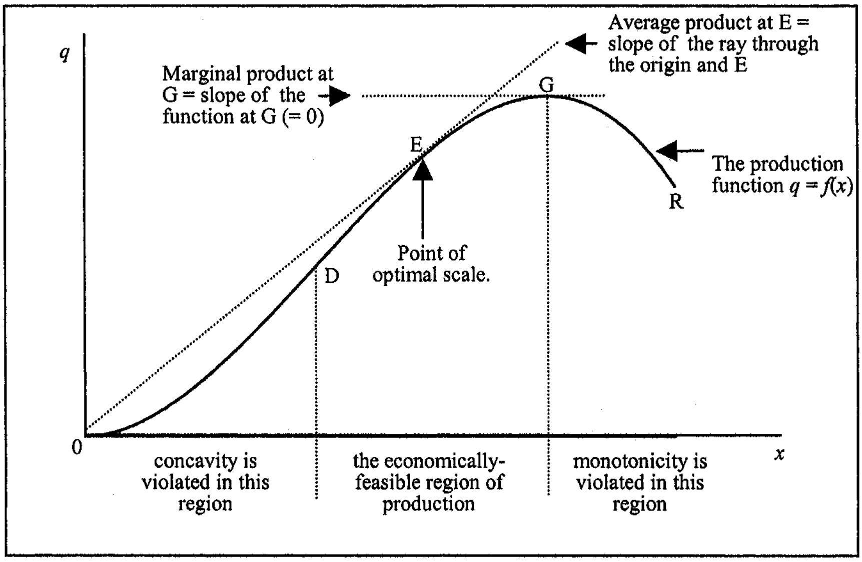

Figure 2.1 Single-Input Production Function

Extending this type of graphical analysis to the multiple-input case is difficult, not least because it is difficult to draw diagrams in more than two dimensions[6]. In such cases, it is common practice to plot the relationship between two of the variables while holding all others fixed. For example, in Figure 2.2 we consider a two-input production function and plot the relationship between the inputs x_1 and x_2 while holding output fixed at the value q^0. We also plot the relationship between the two inputs when output is fixed at the values q^1 and q^2, where $q^2 > q^1 > q^0$. The curves in this figure are known as output *isoquants*. If properties F.1 to F.4 are satisfied, these isoquants are non-intersecting functions that are convex to the origin, as depicted in Figure 2.2. The slope of the isoquant is known as the *marginal rate of technical substitution (MRTS)* – it measures the rate at which x_1 must be substituted for x_2 in order to keep output at its fixed level.

An alternative representation of a two-input production function is provided in Figure 2.3. In this figure, the lowest of the four functions, $q = f(x_1 | x_2 = x_2^0)$, plots the relationship between q and x_1 while holding x_2 fixed at the value x_2^0. The other functions plot the relationship between q and x_1 when x_2 is fixed at the values x_2^1, x_2^2, x_2^3 and x_2^4, where $x_2^4 > x_2^3 > x_2^2 > x_2^1 > x_2^0$.

[6] Some 3D representations of production functions can be found in Beattie and Taylor (1985).

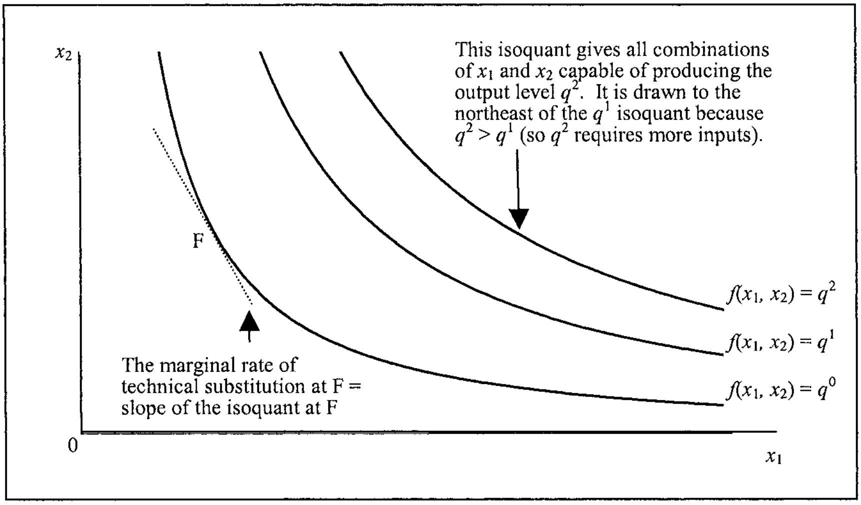

Figure 2.2 Output Isoquants

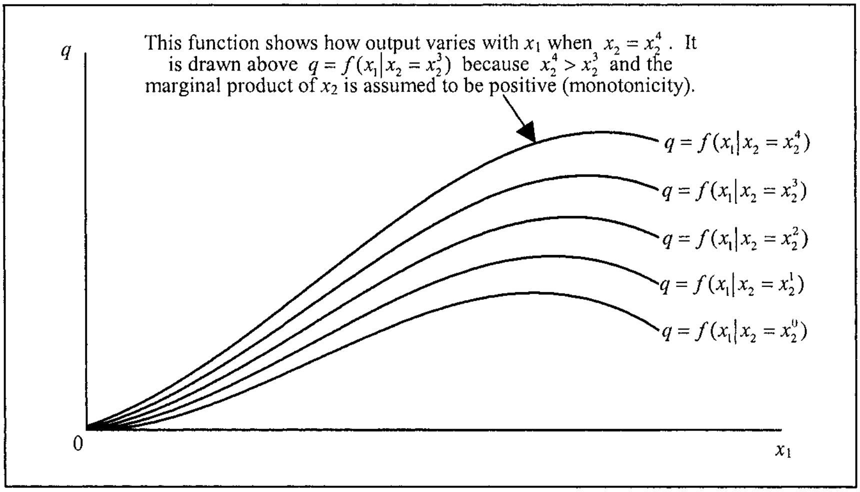

Figure 2.3 A Family of Production Functions

2.2.2 Quantities of Interest

If the production function 2.1 is twice-continuously differentiable we can use calculus to define a number of economic quantities of interest. For example, two quantities we have already encountered are the *marginal product,*

$$\mathrm{MP}_n = \frac{\partial f(\mathbf{x})}{\partial x_n}, \tag{2.2}$$

and the *marginal rate of technical substitution*:

$$\mathrm{MRTS}_{nm} = \frac{\partial x_n(x_1,\ldots,x_{n-1},x_{n+1},\ldots,x_N)}{\partial x_m} = -\frac{\mathrm{MP}_m}{\mathrm{MP}_n}. \tag{2.3}$$

In equation 2.3, $x_n(x_1, \ldots, x_{n-1}, x_n, \ldots, x_N)$ is an implicit function telling us how much of x_n is required to produce a fixed output when we use amounts $x_1, \ldots, x_{n-1}, x_n, \ldots, x_N$ of the other inputs[7]. Related concepts that do not depend on units of measurement are the *output elasticity*,

$$\mathrm{E}_n = \frac{\partial f(\mathbf{x})}{\partial x_n}\frac{x_n}{f(\mathbf{x})}, \tag{2.4}$$

and the *direct elasticity of substitution*:

$$\mathrm{DES}_{nm} = \frac{d(x_m/x_n)}{d(\mathrm{MP}_n/\mathrm{MP}_m)}\frac{\mathrm{MP}_n/\mathrm{MP}_m}{x_m/x_n}. \tag{2.5}$$

In the two-input case the DES is usually denoted σ.

Recall from Figure 2.2 that the MRTS measures the *slope* of an isoquant. The DES measures the percentage change in the input ratio relative to the percentage change in the MRTS, and is a measure of the *curvature* of the isoquant. To see this, consider movements along the isoquants depicted in Figure 2.4. In panel (a), an infinitesimal movement from one side of point A to the other results in an infinitesimal change in the input ratio but an infinitely large change in the MRTS, implying $\sigma \equiv \mathrm{DES}_{12} = 0$. Thus, in the case of a right-angled isoquant, an efficient firm must use its inputs in fixed proportions (i.e., no substitution is possible). In panel (c), a movement from D to E results in a large percentage change in the input ratio but leaves the MRTS unchanged, implying $\sigma = \infty$. In this case, the isoquant is a straight line and inputs are perfect substitutes. An intermediate (and more common) case is depicted in panel (b).

---

[7] For example, in the two-input case the implicit function $x_2(x_1)$ must satisfy $q^0 = f(x_1, x_2(x_1))$ where q^0 is a fixed value. Incidentally, differentiating both sides of this expression with respect to x_1 (and rearranging) we can show that MRTS is the negative of the ratio of the two marginal products (i.e., equation 2.3).

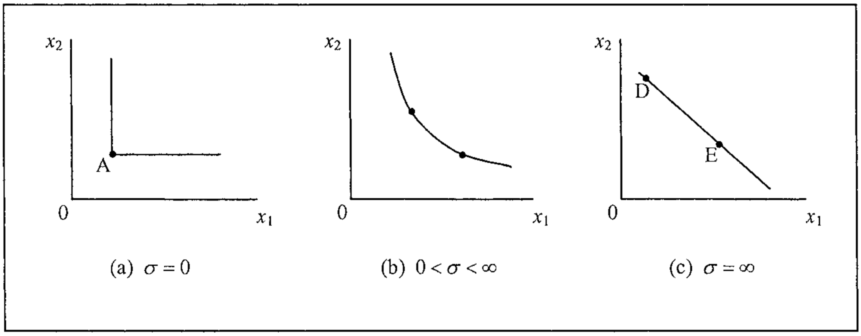

Figure 2.4 Elasticities of Substitution

In the multiple-input case it is possible to define at least two other elasticities of substitution – the *Allen partial elasticity of substitution* (AES) and the *Morishima elasticity of substitution* (MES). The DES is sometimes regarded as a short-run elasticity because it measures substitutability between x_n and x_m while holding all other inputs fixed (economists use the term "short-run" to refer to time horizons so short that at least one input is fixed). The AES and MES are long-run elasticities because they allow all inputs to vary. When there are only two inputs DES = AES. For more details see Chambers (1988, pp. 27-36).

The marginal product given by equation 2.2 measures the output response when one input is varied and all other inputs are held fixed. However, we are often interested in measuring output response when all inputs are varied simultaneously. If a proportionate increase in all inputs results in a *less than proportionate* increase in output (eg., doubling all inputs results in less than twice as much output) then we say the production function exhibits *decreasing returns to scale* (DRS). If a proportionate increase in inputs results in the *same* proportionate increase in output (eg., doubling all inputs results in exactly twice as much output) the production function is said to exhibit constant returns to scale (CRS). Finally, if a proportionate increase inputs leads to a *more than proportionate* increase in output the production function exhibits increasing returns to scale (IRS). Mathematically, if we scale all inputs by an amount $k > 1$ then

$$f(k\mathbf{x}) < kf(\mathbf{x}) \quad \Leftrightarrow \quad \text{DRS}, \tag{2.6}$$

$$f(k\mathbf{x}) = kf(\mathbf{x}) \quad \Leftrightarrow \quad \text{CRS}, \tag{2.7}$$

and

$$f(k\mathbf{x}) > kf(\mathbf{x}) \quad \Leftrightarrow \quad \text{IRS}. \tag{2.8}$$

There are many reasons why firms may experience different returns to scale. For example, a firm may exhibit IRS if the hiring of more staff permits some specialisation of labour, but may eventually exhibit DRS if it becomes so large that management is no longer able to exercise effective control over the production