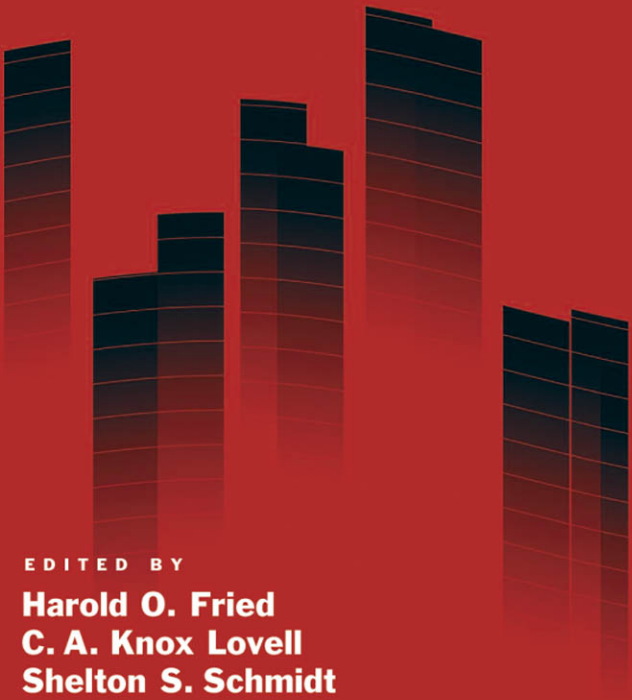
The Measurement of Productive Efficiency and Productivity Growth



EDITED BY

Harold O. Fried
C. A. Knox Lovell
Shelton S. Schmidt

The Measurement of Productive Efficiency and Productivity Growth

This page intentionally left blank

The Measurement of Productive Efficiency and Productivity Growth

Edited by

Harold O. Fried
C. A. Knox Lovell
Shelton S. Schmidt

OXFORD
UNIVERSITY PRESS
2008

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Library of Congress Cataloging-in-Publication Data

The measurement of productive efficiency and productivity growth/edited by
Harold O. Fried, C.A. Knox Lovell, and Shelton S. Schmidt.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-19-518352-8
1. Industrial efficiency—Measurement. 2. Industrial productivity—Measurement.
I. Fried, Harold O. II. Lovell, C. A. Knox.
III. Schmidt, Shelton S.
HD56.25.M4356 2007
338.5—dc22 2007015061

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

We extend thanks to our many friends and colleagues around the world; your research and stimulating conversations have inspired the material in this book.

This page intentionally left blank

Preface

Some individuals are more productive than others; some small businesses find and exploit a lucrative market niche that others miss; some large corporations are more profitable than others; and some public agencies provide more efficient service than others. In each case, performance, both absolute and relative to the competition, can improve through time or lag behind. Success in the short run can be associated with failure in the long run; failure in the short run can lead to death, or it may be the precursor of success in the long run.

What do we mean by business performance? Surely it is multidimensional, but for most producers, the ultimate yardstick is profit. However, we take the view that profit, or any other financial indicator, is a reflection, rather than a measure, of business performance. Performance itself means doing the right things right. This involves solving the purely technical problem of avoiding waste, by producing maximum outputs from available inputs or by using minimum inputs required to produce desired outputs. It also involves solving the allocative problems of using inputs in the right proportion and producing outputs in the right proportion, where "right" generally respects prevailing input prices and output prices. Technical and allocative efficiency in the use of inputs leads to cost efficiency. Technical and allocative efficiency in the production of outputs leads to revenue efficiency. Overall technical and allocative efficiency leads to profit efficiency, the generation of maximum possible profit under the circumstances.

However, circumstances change through time, and so changes in business performance involve changes in technical and allocative efficiency. But they also involve changes in productivity arising from development or adoption of

new technologies that bring improvements in efficiency of both types. Thus, business performance has both static and dynamic aspects.

Why does business performance vary? Ultimate responsibility for performance rests with management. We believe that inefficiency arises from the varying abilities of managers and that firms with varying degrees of inefficiency that operate in overlapping markets can coexist for some period of time. Our belief comes from observing the real world, in which reported quarterly earnings vary across similar firms and through time, and in which these same firms provide employment opportunities for myriad business consulting gurus armed with their buzzwords and their remedies for a host of economic ailments.

Managerial ability is unobservable. Consequently, we infer it from technical efficiency relative to the competition, which involves construction of a best practice production frontier and measurement of distance to it for each entity. We also infer it from cost, revenue, or profit efficiency relative to the competition, which involves construction of a best practice cost, revenue, or profit frontier and measurement of distance to it for each entity. In a dynamic context, we are interested in the progressive entities that push the envelope and in which entities keep pace or fall behind. In all circumstances, we attempt to level the playing field by distinguishing variation in business performance from variation in the operating environment.

This is a relatively new, and to some a heretical, approach to business performance evaluation. The notions of best practice and benchmarking are firmly entrenched in the business world, but they are inconsistent with the economics textbook world in which all firms optimize all the time. The objective of this book is to merge the two strands of thought by developing analytically rigorous models of failure to optimize and of failure to be progressive. We do not dispense with the time-honored paradigm of optimizing behavior. Rather, we build on the paradigm by allowing for varying degrees of success in the pursuit of conventional economic objectives. The objective is to bring economic analysis closer to a framework useful for the evaluation of business performance. Call it economically and analytically rigorous benchmarking.

Two approaches to business performance evaluation have been developed, one in economics and the other in management science. The former uses parametric econometric techniques, and the latter uses nonparametric mathematical programming techniques. They share a common objective, that of benchmarking the performance of the rest against that of the best. We take an eclectic approach by reporting both, and mixtures and extensions of the two, as well. For the novice reader, chapter 1 provides a broad-ranging introduction to the field that provides preparation and motivation for continuing. Distinguished experts who have developed and extended the field of efficiency and productivity analysis contribute subsequent chapters. For open-minded and agnostic readers, chapters 2 and 3 provide in-depth expositions of the two approaches. For readers already committed to an approach, chapters 2 and 3 provide the opportunity to broaden their perspective by an exposure to

an alternative approach. For all readers, chapter 4 advances both approaches by providing a nonparametric statistical approach to performance evaluation. Again for all readers, chapter 5 extends and combines the two approaches to develop a dynamic approach to performance evaluation in which the measurement of productivity growth and the identification of its sources occupy center stage.

The contributors to this book work hard and play hard. In fact, we believe there is no clear distinction between work and play. Although we are all getting older (some more advanced than others), there remains a freshness and excitement toward the material, much of which is new. The field is young; many challenging problems have been solved, but many others remain. It is our hope that this book will inspire the jaded and recruit the novices. There are discoveries to be made, at the office and at the bar. This was all too apparent when we gathered in Athens, Georgia, to share early drafts of the chapters. Enjoy reading what we have enjoyed writing.

Harold O. Fried
C. A. Knox Lovell
Shelton S. Schmidt

This page intentionally left blank

Contents

This page intentionally left blank

Contributors

Ozren Despić
Aston Business School
Aston University
Birmingham, UK
o.despic@aston.ac.uk

Rolf Färe
Department of Agricultural
& Resource Economics
& Department of Economics
Oregon State University
Corvallis, OR
rolf.fare@orst.edu

Harold O. Fried
Department of Economics
Union College
Schenectady, NY
friedh@union.edu

William H. Greene
Department of Economics
Stern School of Business
New York University
New York, NY
wgreene@stern.nyu.edu

Shawna Grosskopf
Department of Economics
Oregon State University
Corvallis, OR
shawna.grosskopf@orst.edu

C. A. Knox Lovell
Emeritus Professor
Department of Economics
University of Georgia
Athens, GA
knox@terry.uga.edu
and
Honorary Professor
School of Economics
University of Queensland
Brisbane, Australia

Dimitri Margaritis
Finance Department
Auckland University of Technology Auckland, New Zealand
dimitri.margaritis@aut.ac.nz

Maria C. S. Portela
Faculdade de Economia e Gestão
Universidade Católica Portuguesa
Porto, Portugal
csilva@porto.ucp.pt

Shelton S. Schmidt
Department of Economics
Union College
Schenectady, NY
schmidts@union.edu

Léopold Simar
Institut de Statistique
Université Catholique de Louvain
Louvain-la-Neuve, Belgium
simar@stat.ucl.ac.be

Emmanuel Thanassoulis
Aston Business School
Aston University
Birmingham, UK
e.thanassoulis@aston.ac.uk

Paul W. Wilson
The John E. Walker Department of Economics
Clemson University
Clemson, SC
pww@clemson.edu

The Measurement of Productive Efficiency and Productivity Growth

This page intentionally left blank

1

Efficiency and Productivity

Harold O. Fried, C. A. Knox Lovell,
and Shelton S. Schmidt

1.1 Introduction

Airlines in the United States have encountered difficulties since September 11, 2001, particularly on domestic routes. Figure 1.1 plots quarterly operating profit margins (profit from domestic operations as a percentage of operating revenue) for three segments of the industry: the small regional airlines; the medium-size, low-cost airlines; and the large network airlines. The regional airlines have performed relatively well, earning more than a 10% profit margin, and the low-cost airlines have performed adequately, earning a considerably smaller but nonetheless positive profit margin. However, the network airlines have performed poorly, earning a large negative profit margin. Some have sought bankruptcy protection.

When we ask why airline performance has varied so much, we naturally think of revenues and costs. Figure 1.2 plots quarterly operating revenue per available seat-mile (a measure of average revenue), and figure 1.3 plots quarterly operating expense per available seat-mile (a measure of average cost). On the revenue side, the regional airlines earned the highest operating revenue per available seat-mile, trailed in order by the network airlines and the low-cost airlines. On the cost side, the low-cost airlines incurred the lowest operating cost per available seat-mile, appropriately enough, trailed by the network airlines and the regional airlines.

It appears that the regional airlines have been the most profitable segment of the domestic airline industry despite having had the highest unit costs. The low-cost airlines have been marginally profitable because their low unit

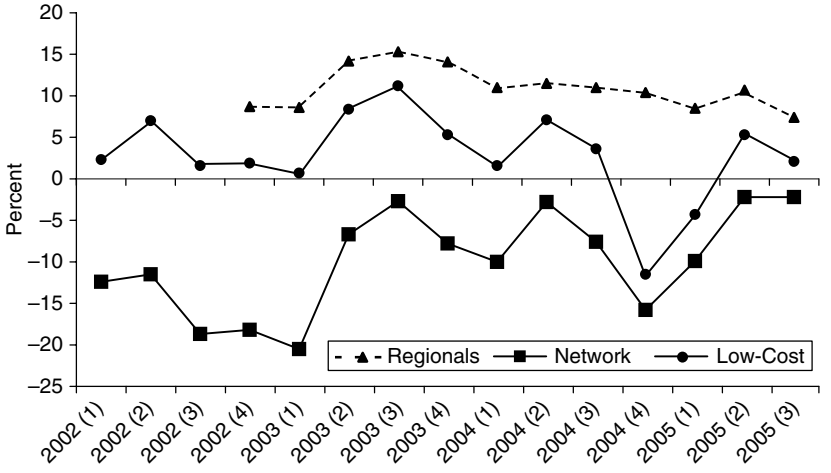


Figure 1.1. Airline Domestic Operating Profit Margin (BLS, 2005b)

revenues have been offset by even lower unit costs. Finally, the network airlines have lost money primarily because of their high unit costs.

On the cost side, three hypotheses spring quickly to mind, each inspired by conventional economic theory. First, the pattern of unit operating costs may reflect a pattern of scale economies that generates a U-shaped minimum average cost function favoring the medium-size low-cost airlines. Second, it may reflect higher input prices paid by the regional and network airlines. This hypothesis rings true for the older network airlines, which at the time were

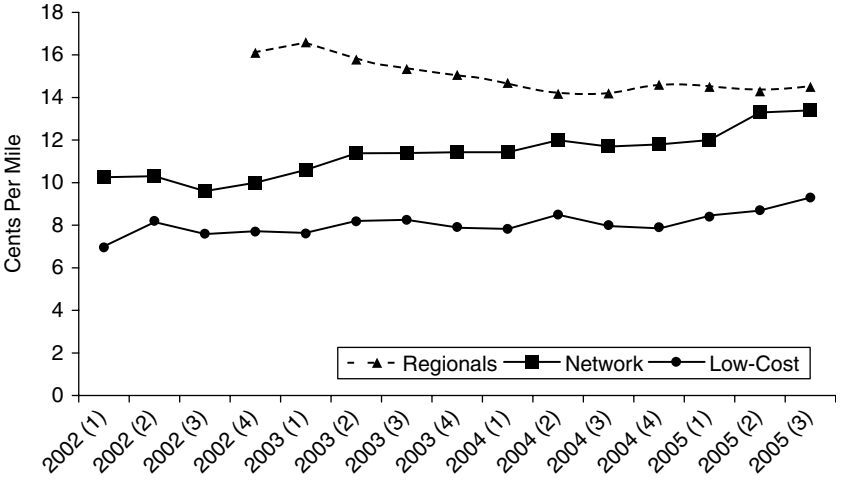


Figure 1.2. Airline Domestic Unit Revenue (BLS, 2005b)

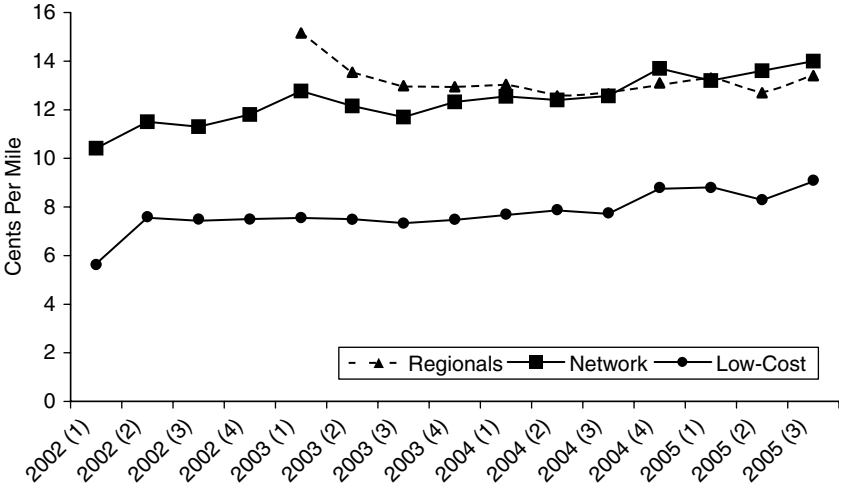


Figure 1.3. Airline Domestic Unit Cost (BLS, 2005b)

burdened by high labor costs attributable in large part to onerous pension obligations. Third, it may reflect different technologies embedded in a "low-cost business model" employed by the low-cost airlines and an inefficient "hub-and-spoke" system employed by the network airlines. Support for this hypothesis comes from the network airlines themselves, which predict efficiency gains and cost savings as they gradually abandon the system they adopted three decades ago.

On the revenue side, differential pricing power is a possible explanation, although it is not clear why the small regional airlines would have such an advantage. A more likely explanation is variation in rates of capacity utilization as measured by load factors (the percentage of available seats actually sold), which might have favored the regional airlines and penalized the low-cost airlines.

Each of these hypotheses is suggested by economic theory and may or may not be refuted by the evidence. We now put forth an additional pair of refutable hypotheses that, although not suggested by conventional economic theory, should not be dismissed a priori.

One hypothesis concerns the cost side and posits that part of the observed pattern of unit operating cost may be a consequence of cost inefficiency at the regional and network airlines. Cost inefficiency can be "technical," arising from excessive resource use given the amount of traffic, or "allocative," arising from resources being employed in the wrong mix, given their prices. Perhaps the low-cost airlines had relatively low unit costs because they utilized part-time labor and because they leased, rather than purchased, aircraft. Either strategy would reduce idleness and down time. More generally, perhaps the low-cost airlines had relatively low unit costs because their resources, human

and physical, were well managed. This would place them on the minimum average cost function, whereas cost inefficiency at the regional and network airlines would place them above the minimum average cost function.

The second hypothesis concerns the revenue side and posits that part of the observed pattern of unit operating revenue may be a consequence of revenue inefficiency at the network and low-cost airlines. Revenue inefficiency can be "technical," arising from a failure to provide maximum service from the available resources, or "allocative," arising from the provision of services in the wrong mix, given their prices. Perhaps the regional airlines were nimble enough to adjust their route structures to respond quickly to fluctuations in passenger demand. Perhaps the regional airlines have faster gate turnaround times than the network airlines, whose hub-and-spoke technology leaves aircraft and crew idle and sacrifices revenue. This would place the regional airlines on the maximum average revenue function, whereas revenue inefficiency at the network and low-cost airlines would place them beneath the maximum average revenue function.

The point of the foregoing discussion is not to engage in a deep exploration into airline economics, about which we are blissfully ignorant. We are merely frequent fliers who happen to be curious economists wondering what might explain the observed variation in the recent domestic performance of U.S. airlines. The point is to suggest that variation in productive efficiency, in both the management of resources and the management of services, may be a potentially significant source of variation in financial performance. Inefficient behavior is assumed away in conventional economic theory, in which first-order and second-order optimizing conditions are satisfied. But it exists in the real world, as a perusal of almost any trade publication will verify, and as the hordes of consultants armed with their buzzwords will testify.

Productive inefficiency exists, and it deserves to be included in our analytical toolkit because it can generate refutable hypotheses concerning the sources of variation in business performance. This book is devoted to the study of inefficiency in production and its impact on economic and financial performance. The study ranges from the underlying theory to the analytical foundations, and then to the quantitative techniques and the empirical evidence.

Chapter 1 sets the stage. Section 1.2 provides background material and focuses on hypotheses that have been proposed in the literature that would explain variation in producer performance. This section also provides a glimpse at the empirical literature and demonstrates that the search for variation in producer performance has been conducted in a wide variety of settings. Section 1.3 lays the theoretical foundation for the measurement of productive efficiency. It provides definitions of alternative notions of productive efficiency, and it provides corresponding measures of efficiency. Section 1.4 offers a brief introduction to alternative techniques that have been developed to quantify inefficiency empirically. Section 1.5 introduces various econometric approaches to efficiency estimation, while section 1.6 introduces variants of the mathematical programming approach to efficiency estimation. Section 1.7

introduces the Malmquist productivity index and shows how to decompose it into various sources of productivity change, including variation in productive efficiency. Section 1.8 describes three ways of approximating a Malmquist productivity index: the use of superlative index numbers, the use of econometric techniques, and the use of mathematical programming techniques. Section 1.9 offers some concluding observations.

Chapter 2 extends section 1.5 by providing a detailed survey of the econometric approach to efficiency estimation. Chapter 3 extends section 1.6 by providing a detailed survey of the mathematical programming approach to efficiency estimation. Chapter 4 recasts the parametric and statistical approach of chapter 2, and the nonparametric and deterministic approach of chapter 3, into a nonparametric and statistical approach. Chapter 5 extends sections 1.7 and 1.8 by discussing alternative approaches to the measurement of productivity change, with special emphasis on efficiency change as a source of productivity change.

1.2 Background

When discussing the economic performance of producers, it is common to describe them as being more or less "efficient" or more or less "productive." In this section, we discuss the relationship between these two concepts. We consider some hypotheses concerning the *determinants* of producer performance, and we consider some hypotheses concerning the financial *consequences* of producer performance.

By the *productivity* of a producer, we mean the ratio of its output to its input. This ratio is easy to calculate if the producer uses a single input to produce a single output. In the more likely event that the producer uses several inputs to produce several outputs, the outputs in the numerator must be aggregated in some economically sensible fashion, as must the inputs in the denominator, so that productivity remains the ratio of two scalars. Productivity growth then becomes the difference between output growth and input growth, and the aggregation requirement applies here, as well.

Variation in productivity, either across producers or through time, is thus a residual, which Abramovitz (1956) famously characterized as "a measure of our ignorance." Beginning perhaps with Solow (1957), much effort has been devoted to dispelling our ignorance by "whittling away at the residual" (Stone, 1980). Much of the whittling has involved minimizing measurement error in the construction of output and input quantity indexes. The conversion of raw data into variables consistent with economic theory is a complex undertaking. Griliches (1996) surveys the economic history of the residual, and state-of-the-art procedures for whittling away at it are outlined by the Organisation for Economic Co-operation and Development (OECD, 2001). When the whittling is finished, we have a residual suitable for analysis.

In principle, the residual can be attributed to differences in production technology, differences in the scale of operation, differences in operating efficiency, and differences in the operating environment in which production occurs. The U.S. Department of Labor's Bureau of Labor Statistics (BLS 2005) and the OECD (2001) attribute variation in productivity through time to these same sources. Proper attribution is important for the adoption of private managerial practices and the design of public policies intended to improve productivity performance. We are naturally interested in isolating the first three components, which are under the control of management, from the fourth, which is not. Among the three endogenous components, our interest centers on the efficiency component and on measuring both its cross-sectional contribution to variation in productivity and its intertemporal contribution to productivity change.

By the *efficiency* of a producer, we have in mind a comparison between observed and optimal values of its output and input. The exercise can involve comparing observed output to maximum potential output obtainable from the input, or comparing observed input to minimum potential input required to produce the output, or some combination of the two. In these two comparisons, the optimum is defined in terms of production possibilities, and efficiency is technical. It is also possible to define the optimum in terms of the behavioral goal of the producer. In this event, efficiency is measured by comparing observed and optimum cost, revenue, profit, or whatever goal the producer is assumed to pursue, subject, of course, to any appropriate constraints on quantities and prices. In these comparisons, the optimum is expressed in value terms, and efficiency is economic.

Even at this early stage, three problems arise, and much of this section is devoted to exploring ways each has been addressed. First, which outputs and inputs are to be included in the comparison? Second, how are multiple outputs and multiple inputs to be weighted in the comparison? And third, how is the technical or economic potential of the producer to be determined?

Many years ago, Knight (1933/1965) addressed the first question by noting that if all outputs and all inputs are included, then since neither matter nor energy can be created or destroyed, all producers would achieve the same unitary productivity evaluation. In this circumstance, Knight proposed to redefine productivity as the ratio of useful output to input. Extending Knight's redefinition to the ratio of useful output to useful input, and representing usefulness with weights incorporating market prices, generates a modern economic productivity index. As a practical matter, however, the first problem is not how to proceed when all outputs and all inputs are included, but rather how to proceed when not enough outputs and inputs are included.

As Stigler (1976) has observed, measured inefficiency may be a reflection of the analyst's failure to incorporate all relevant variables and, complicating the first problem, to specify the right economic objectives and the right constraints. Stigler was criticizing the work of Leibenstein (1966, 1976), who focused on inadequate motivation, information asymmetries, incomplete

contracts, agency problems, and the attendant monitoring difficulties within the firm, and who lumped all these features together and called the mix "X-inefficiency." When the agents' actions are not aligned with the principal's objective, potential output is sacrificed. Thus, what appears as inefficiency to Leibenstein is evidence of an incomplete model to Stigler (1976), who called it waste and concluded that "waste is not a useful economic concept. Waste is error within the framework of modern economic analysis" (p. 216). The practical significance of this exchange is that if Stigler's wish is not granted, and not all variables reflecting the objectives and constraints of the principal and the agents are incorporated into the model, agency and related problems become potential sources of measured (if not actual) inefficiency.

Leibenstein was not writing in a vacuum. His approach fits nicely into the agency literature. The recognition of agency problems goes back at least as far as the pioneering Berle and Means (1932) study of the consequences of the separation of ownership from control, in which owners are the principals and managers are the agents. Leibenstein's notion of X-inefficiency also has much in common with Simon's (1955) belief that in a world of limited information processing ability, managers exhibit "bounded rationality" and engage in "satisficing" behavior. Along similar lines, Williamson (1975, 1985) viewed firms as seeking to economize on transaction costs, which in his view boiled down to economizing on bounded rationality. Bounded rationality and the costs of transacting also become potential sources of measured inefficiency.

It would be desirable, if extraordinarily difficult, to construct and implement Stigler's complete model involving all the complexities mentioned above. We have not seen such a model. What we have seen are simplified (if not simple) models of the firm in which measured performance differentials presumably reflect variation in the ability to deal with the complexities of the real world. Indeed, performance measures based on simplified models of the firm are often useful, and sometimes necessary. They are useful when the objectives of producers, or the constraints facing them, are either unknown or unconventional or subject to debate. In this case, a popular research strategy has been to model producers as unconstrained optimizers of some conventional objective and to test the hypothesis that inefficiency in this environment is consistent with efficiency in the constrained environment. The use of such incomplete measures has proved necessary in a number of contexts for lack of relevant data. One example of considerable policy import occurs when the production of desirable (and measured and priced) outputs is recorded, but the generation of undesirable (and frequently unmeasured and more frequently unpriced) byproducts is not. Another occurs when the use of public infrastructure enhances private performance, but its use goes unrecorded. In each case, the measure of efficiency or productivity that is obtained may be very different from the measure one would like to have.

Even when all relevant outputs and inputs are included, there remains the formidable second problem of assigning weights to variables. Market prices provide a natural set of weights, but two types of question arise. First,

suppose market prices exist. If market prices change through time, or vary across producers, is it possible to disentangle the effects of price changes and quantity changes in a relative performance evaluation? Alternatively, if market prices reflect monopoly or monopsony power, or cross-subsidy, or the determination of a regulator, do they still provide appropriate weights in a relative performance evaluation? Second, suppose some market prices do not exist. In the cases of environmental impacts and public infrastructure mentioned above, the unpriced variables are externalities either generated by or received by market sector producers. How do we value these externalities? However, the weighting problem is more pervasive than the case of externalities. The nonmarket sector is growing relative to the market sector in most advanced economies, and by definition, the outputs in this sector are not sold on markets. How, then, do we value outputs such as law enforcement and fire protection services, or even public education services, each of which is publicly funded rather than privately purchased? Is it possible to develop proxies for missing prices that would provide appropriate weights in a performance evaluation? The presence of distorted or missing prices complicates the problem of determining what is meant by "relevant."

The third problem makes the first two seem easy. It is as difficult for the analyst to determine a producer's potential as it is for the producer to achieve that potential. It is perhaps for this reason that for many years the productivity literature ignored the efficiency component identified by the BLS and the OECD. Only recently, with the development of a separate literature devoted to the study of efficiency in production, has the problem of determining productive potential been seriously addressed. Resolution of this problem makes it possible to integrate the two literatures. Integration is important for policy purposes, since action taken to enhance productivity performance requires an accurate attribution of observed performance to its components.

By way of analogy, we do not know, and cannot know, how fast a human can run 100 meters. But we do observe best practice and its improvement through time, and we do observe variation in actual performance among runners. The world of sport is full of statistics, and we have all-star teams whose members are judged to be the best at what they do. Away from the world of sports, we use multiple criteria to rank cities on the basis of quality-of-life indicators (Zurich and Geneva are at the top). At the macro level, we use multiple criteria to rank countries on the basis of economic freedom (Norway, Sweden, and Australia are at the top), environmental sustainability (Finland and Norway are at the top), business risk (Iraq and Zimbabwe pose the most risk), and corruption (Finland and New Zealand are the least corrupt), among many others. The United Nation's Human Development Index is perhaps the best-known and most widely studied macroeconomic performance indicator (Norway and Sweden are at the top). In each of these cases, we face the three problems mentioned at the outset of this section: what indicators to include, how to weight them, and how to define potential. The selection and weighting of indicators are controversial by our standards, although comparisons

are appropriately made relative to best practice rather than to some ideal standard.

The same reasoning applies to the evaluation of business performance. We cannot know "true" potential, whatever the economic objective. But we do observe best practice and its change through time, and we also observe variation in performance among producers operating beneath best practice. This leads to the association of "efficient" performance with undominated performance, or operation on a best-practice "frontier," and of inefficient performance with dominated performance, or operation on the wrong side of a best-practice frontier. Interest naturally focuses on the identification of best-practice producers and on benchmarking the performance of the rest against that of the best. Businesses themselves routinely benchmark their performance against that of their peers, and academic interest in benchmarking is widespread, although potential synergies between the approaches adopted by the two communities have yet to be fully exploited. Davies and Kochhar (2002) offer an interesting academic critique of business benchmarking.

Why the interest in measuring efficiency and productivity? We can think of three reasons. First, only by measuring efficiency and productivity, and by separating their effects from those of the operating environment so as to level the playing field, can we explore hypotheses concerning the sources of efficiency or productivity differentials. Identification and separation of controllable and uncontrollable sources of performance variation are essential to the institution of private practices and public policies designed to improve performance. Zeitsch et al. (1994) provide an empirical application showing how important it is to disentangle variation in the operating environment (in this case, customer density) from variation in controllable sources of productivity growth in Australian electricity distribution.

Second, macro performance depends on micro performance, and so the same reasoning applies to the study of the growth of nations. Lewis (2004) provides a compelling summary of McKinsey Global Institute (MGI) productivity studies of 13 nations over 12 years, the main findings being that micro performance drives macro performance, and that a host of institutional impediments to strong micro performance can be identified. This book, and the studies on which it is based, make it clear that there are potential synergies, as yet sadly unexploited, between the MGI approach and the academic approach to performance evaluation.

Third, efficiency and productivity measures are success indicators, performance metrics, by which producers are evaluated. However, for most producers the ultimate success indicator is financial performance, and the ultimate metric is the bottom line. Miller's (1984) clever title, "Profitability = Productivity + Price Recovery," encapsulates the relationship between productivity and financial performance. It follows that productivity growth leads to improved financial performance, provided it is not offset by declining price recovery attributable to falling product prices and/or rising input prices. Grifell-Tatjé and Lovell (1999) examine the relationship for Spanish

banks facing increasing competition as a consequence of European monetary union. Salerian (2003) explores the relationship for Australian railroads, for which increasing intermodal competition has contributed to declining price recovery that has swamped the financial benefits of impressive productivity gains. This study also demonstrates that, although the bottom line may be paramount in the private sector, it is not irrelevant in the public sector; indeed, many governments monitor the financial performance as well as the nonfinancial performance of their public service providers.

Many other studies, primarily in the business literature, adopt alternative notions of financial performance, such as return on assets or return on equity. These studies typically begin with the "DuPont triangle," which decomposes return on assets as $\pi/A = (\pi/R)(R/A)$ = (return on sales)(investment turnover), where π = profit, A = assets, and R = revenue. The next step is to decompose the first leg of the DuPont triangle as $(\pi/R) = [(R - C)/R] = [1 - (R/C)^{-1}]$, where C is cost and R/C is profitability. The final step is to decompose profitability into productivity and price recovery, a multiplicative alternative to Miller's additive relationship. The objective is to trace the contribution of productivity change up the triangle to change in financial performance. Horrigan (1968) provides a short history of the DuPont triangle as an integral part of financial ratio analysis, and Eilon (1984) offers an accessible survey of alternative decomposition strategies. Banker et al. (1993) illustrate the decomposition technique with an application to the U.S. telecommunications industry, in which deregulation led to productivity gains that were offset by deteriorating price recovery brought on by increased competition.

In some cases, measurement enables us to quantify performance differentials that are predicted qualitatively by economic theory. An example is provided by the effect of market structure on performance. There is a common belief that productive efficiency is a survival condition in a competitive environment, and that its importance diminishes as competitive pressure subsides. Hicks (1935) gave eloquent expression to this belief by asserting that producers possessing market power "are likely to exploit their advantage much more by not bothering to get very near the position of maximum profit, than by straining themselves to get very close to it. The best of all monopoly profits is a quiet life" (p. 8). Berger and Hannan (1998) provide a test of the quiet life hypothesis in U.S. banking and find evidence that banks in relatively concentrated markets exhibit relatively low cost efficiency.

Continuing the line of reasoning that firms with market power might not be "pure" profit maximizers, Alchian and Kessel (1962) replaced the narrow profit maximization hypothesis with a broader utility maximization hypothesis, in which case monopolists and competitors might be expected to be equally proficient in the pursuit of utility. The ostensible efficiency differential is then explained by the selection of more (observed) profit by the competitor and more (unobserved) leisure by the monopolist, which of course recalls the analyst's problem of determining the relevant outputs and inputs of the

production process. Alchian and Kessel offer an alternative explanation for the apparent superior performance of competitive producers. This is that monopolies are either regulated, and thereby constrained in their pursuit of efficiency, or unregulated but threatened by regulation (or by antitrust action) and consequently similarly constrained. If these producers are capable of earning more than the regulated profit, and if their property rights to the profit are attenuated by the regulatory or antitrust environment, then inefficiency becomes a free good to producers subject to, or threatened by, regulation or antitrust action. As Alchian and Kessel put it, "The cardinal sin of a monopolist … is to be too profitable" (p. 166).

Baumol (1959), Gordon (1961), and Williamson (1964) argued along similar lines. An operating environment characterized by market power and separation of ownership from control leaves room for "managerial discretion." Given the freedom to choose, managers would seek to maximize a utility function in which profit was either one of several arguments or, more likely, a constraint on the pursuit of alternative objectives. This idea, and variants of it, recurs frequently in the agency literature.

Thus, competition is expected to enhance performance either because it forces producers to concentrate on "observable" profit-generating activities at the expense of Hicks's quiet life, or because it frees producers from the actual or potential constraints imposed by the regulatory and antitrust processes. One interesting illustration of the market structure hypothesis is the measurement of the impact of international trade barriers on domestic industrial performance. Many years ago, Carlsson (1972) used primitive frontier techniques to uncover a statistically significant inverse relationship between the performance of Swedish industries and various measures of their protection from international competition. More recently, Tybout and Westbrook (1995), Pavcnik (2002), and Schor (2004) have applied modern frontier techniques to longitudinal micro data in an effort to shed light on the linkage between openness and productivity in Mexico, Chile, and Brazil. Specific findings vary, but a general theme emerges. Trade liberalization brings aggregate productivity gains attributable among other factors to improvements in productivity among continuing firms, and to entry of relatively productive firms and exit of relatively unproductive firms.

A second situation in which measurement enables the quantification of efficiency or productivity differentials predicted fairly consistently by theory is in the area of economic regulation. The most commonly cited example is rate-of-return regulation, to which many utilities have been subjected for many years, and for which there exists a familiar and tractable analytical paradigm developed by Averch and Johnson (1962). Access to a tractable model and to data supplied by regulatory agencies has spawned numerous empirical studies, virtually all of which have found rate-of-return regulation to have led to overcapitalization that has had an adverse impact on utility performance and therefore on consumer prices. These findings have motivated a movement toward incentive regulation in which utilities are reimbursed on the

basis of a price cap or revenue cap formula RPI – X, with X being a productivity (or efficiency) offset to movements in an appropriate price index RPI. The reimbursement formula allows utilities to pass along any cost increases incorporated in RPI, less any expected performance improvements embodied in the offset X. Since X is a performance indicator, this trend has spawned a huge theoretical and empirical literature using efficiency and productivity measurement techniques to benchmark the performance of regulated utilities. Bogetoft (2000, and references cited therein) has developed the theory within a frontier context, in which X can be interpreted as the outcome of a game played between a principal (the regulator) and multiple agents (the utilities). The Netherlands Bureau for Economic Policy Analysis (2000) provides a detailed exposition of the techniques. Kinnunen (2005) reports either declining or stable trends in customer electricity prices in Finland, Norway, and Sweden, where variants of incentive regulation have been in place for some time. Since enormous amounts of money are involved, the specification and weighting of relevant variables and the sample selection criteria become important, and frequently contentious, issues in regulatory proceedings.

Another regulatory context in which theoretical predictions have been quantified by empirical investigation is the impact of environmental controls on producer performance. In this context, however, the private cost of reduced efficiency or productivity must be balanced against the social benefits of environmental protection. Of course, the standard paradigm that hypothesizes private costs of environmental constraints may be wrong; Porter (1991) has argued that well-designed environmental regulations can stimulate innovation, enhance productivity, and thus be privately profitable. Ambec and Barla (2002) develop a theory that predicts the Porter hypothesis. In any event, the problem of specifying and measuring the relevant variables crops up once again. Färe et al. (1989, 1993) have developed the theory within a frontier context. Reinhard et al. (1999) examined a panel of Dutch dairy farms that generate surplus manure, the nitrogen content of which contaminates groundwater and surface water and contributes to acid rain. They calculated a mean shadow price of the nitrogen surplus of just greater than 3 Netherlands guilders (NLG) per kilogram, slightly higher than a politically constrained levy actually imposed of NLG1.5 per kilogram of surplus. Ball et al. (2004) calculated exclusive and inclusive productivity indexes for U.S. agriculture, in which pesticide use causes water pollution. They found that inclusive productivity growth initially lagged behind exclusive productivity growth. However, when the U.S. Environmental Protection Agency began regulating the manufacture of pesticides, inclusive productivity growth caught up with, and eventually surpassed, exclusive productivity growth, as would be expected. Consistent with these findings, Ball et al. found an inverted U-shaped pattern of shadow prices, reflecting a period of lax regulation followed by tightened regulation that eventually led to the discovery and use of relatively benign and more effective pesticides.

A third situation in which measurement can quantify theoretical propositions is the effect of ownership on performance. Alchian (1965) noted that the inability of public-sector owners to influence performance by trading shares in public-sector producers means that public-sector managers worry less about bearing the costs of their decisions than do their private-sector counterparts. Hence, they are contractually constrained in their decision-making latitude, given less freedom to choose, so to speak. "Because of these extra constraints—or because of the 'costs' of them—the public arrangement becomes a higher cost (in the sense of 'less efficient') than that for private property agencies" (p. 828). A literature has developed based on the supposition that public managers have greater freedom to pursue their own objectives, at the expense of conventional objectives. Niskanen (1971) viewed public managers as budget maximizers, de Alessi (1974) viewed public managers as preferring capital-intensive budgets, and Lindsay (1976) viewed public managers as preferring "visible" variables. Each of these hypotheses suggests that measured performance is lower in the public sector than in the private sector. Holmstrom and Tirole (1989) survey much of the theoretical literature, as does Hansmann (1988), who introduces private not-for-profit producers as a third category. Empirical tests of the public/private performance differential hypothesis are numerous. Many of the comparisons have been conducted using regulated utility data, because public and private firms frequently compete in these industries, because of the global trend toward privatization of public utilities, and because regulatory agencies collect and provide data. Jamash and Pollitt (2001) survey the empirical evidence for electricity distribution. Education and health care are two additional areas in which numerous public/private performance comparisons have been conducted.

In any public/private performance comparison, one confronts the problem of how to measure their performance. Pestieau and Tulkens (1993) offer a spirited defense of a narrow focus on technical efficiency, so as to level the playing field. They argue that public enterprises have objectives and constraints (e.g., fiscal balance and universal service, uniform price requirements, but at the same time a soft budget constraint) different from those of private enterprises, and the only common ground on which to compare their performance is on the basis of their technical efficiency.

In some cases, theory gives no guidance, or provides conflicting signals, concerning the impact on performance of some phenomenon. In such cases, empirical measurement provides qualitative, as well as quantitative, evidence. Four examples illustrate the point. Are profit-maximizing firms more efficient than cooperatives? Is one form of sharecropping more efficient than another? Is slavery an efficient way of organizing production? Is organized crime efficiently organized? The answer to each question seems to be "it depends," and so empirical measurement is called for. Theory and evidence are offered by Pencavel (2001) for cooperatives, by Otsuka et al. (1992) and Garrett and Xu (2003) for sharecropping, by Fogel and Engerman (1974) for slavery, and by Fiorentini and Peltzman (1995) for organized crime.

Table 1.1
Empirical Applications of Efficiency and Productivity Analysis

Application	Analysis
Accounting, advertising, auditing, and law firms	Banker et al. (2005) Luo and Donthu (2005) Dopuch et al. (2003) Wang (2000)
Airports	Oum and Yu (2004) Sarkis and Talluri (2004) Yoshida and Fujimoto (2004) Yu (2004)
Air transport	Coelli et al. (2002) Sickles et al. (2002) Scheraga (2004) Duke and Torres (2005)
Bank branches	Davis and Albright (2004) Camanho and Dyson (2005) Porembski et al. (2005) Silva Portela and Thanassoulis (2005)
Bankruptcy prediction	Wheelock and Wilson (2000) Becchetti and Sierra (2003) Cielen et al. (2004)
Benefit–cost analysis	Goldar and Misra (2001) Hofler and List (2004) Chien et al. (2005)
Community and rural health care	Birman et al. (2003) Dervaux et al. (2003) Jiménez et al. (2003) Kirigia et al. (2004)
Correctional facilities	Gyimah-Brempong (2000) Nyhan (2002)
Credit risk evaluation	Emel et al. (2003) Paradi et al. (2004)
Dentistry	Buck (2000) Grytten and Rongen (2000) Linna et al. (2003) Widstrom et al. (2004)
Discrimination	Croppenstedt and Meschi (2000) Bowlin et al. (2003) Mohan and Ruggiero (2003)
Education: primary and secondary	Dolton et al. (2003) Mayston (2003) Ammar et al. (2004) Dodson and Garrett (2004)
Education: tertiary	Bonaccorsi and Daraio (2003) Mensah and Werner (2003) Guan and Wang (2004) Warning (2004)

Application	Analysis
Elections	Obata and Ishii (2003) Foroughi et al. (2005)
Electricity distribution	Agrell et al. (2005) Delmas and Tokat (2005) Pollitt (2005) Edvardsen et al. (2006)
Electricity generation	Arocena and Waddams Price (2003) Korhonen and Luptacik (2004) Atkinson and Halabi (2005) Cook and Green (2005)
Environment: macro applications	Jeon and Sickles (2004) Zaim (2004) Arcelus and Arocena (2005) Henderson and Millimet (2005)
Environment: micro applications	Gang and Felmingham (2004) Banzhaf (2005) Shadbegian and Gray (2005) Wagner (2005)
Financial statement analysis	Chen and Zhu (2003) Feroz et al. (2003)
Fishing	Chiang et al. (2004) Herrero (2004) Martinez-Cordero and Leung (2004) Kompas and Che (2005)
Forestry	Otsuki et al. (2002) Bi (2004) Hof et al. (2004) Liu and Yin (2004)
Gas distribution	Rossi (2001) Carrington et al. (2002) Hammond et al. (2002) Hawdon (2003)
Hospitals	Chang et al. (2004) Stanford (2004) Ventura et al. (2004) Gao et al. (2006)
Hotels	Hwang and Chang (2003) Chiang et al. (2004) Barros (2005) Sigala et al. (2005)
Inequality and Poverty Insurance	Deutsch and Silber (2005) Greene and Segal (2004) Cummins et al. (2005) Jeng and Lai (2005) Tone and Sahoo (2005)

(Continued)

Table 1.1
(Continued)

Application	Analysis
Internet commerce	Wen et al. (2003) Barua et al. (2004) Chen et al. (2004) Serrano-Cinca et al. (2005)
Labor markets	Sheldon (2003) Ibourk et al. (2004) Lang (2005) Millimet (2005)
Libraries	Hammond (2002) Shim (2003) Kao and Lin (2004) Reichmann and Sommersguter-Reichmann (2006)
Location	Thomas et al. (2002) Cook and Green (2003) Takamura and Tone (2003)
Macroeconomics	Cherchye et al. (2004) Despotis (2005) Ravallion (2005) Yörük and Zaim (2005)
Mergers	Cuesta and Orea (2002) Ferrier and Valdmanis (2004) Bogetoft and Wang (2005) Sherman and Rupert (2006)
Military	Barros (2002) Bowlin (2004) Brockett et al. (2004) Sun (2004)
Municipal services	Hughes and Edwards (2000) Moore et al. (2001) Prieto and Zofio (2001) Southwick (2005)
Museums	Mairesse and Vanden Eeckaut (2002) Bishop and Brand (2003) Basso and Funari (2004)
Nursing homes	Farsi and Filippini (2004) Hougaard et al. (2004) Laine et al. (2005) Dervaux et al. (2006)
Physicians and physician practices	Wagner et al. (2003) Rosenman and Friesner (2004)
Police	Spottiswoode (2000) Wisniewski and Dickson (2001) Stone (2002) Drake and Simper (2004)
Ports	Clark et al. (2004) Lawrence and Richards (2004)

Application	Analysis
	Park and De (2004) Cullinane et al. (2005)
Postal services	Pimenta et al. (2000) Maruyama and Nakajima (2002) Borenstein et al. (2004)
Public infrastructure	Mamatzakis (2003) Martıın et al. (2004) Paul et al. (2004) Salinas-Jiminez (2004)
Rail transport	Kennedy and Smith (2004) Loizides and Tsionas (2004) Farsi et al. (2005) Smith (2005)
Real estate investment trusts	Lewis et al. (2003) Anderson et al. (2004)
Refuse collection and recycling	Bosch et al. (2000) Worthington and Dollery (2001) Lozano et al. (2004)
Sports	Haas (2003) Lins et al. (2003) Fried et al. (2004) Amos et al. (2005)
Stocks, mutual funds, and hedge funds	Basso and Funari (2003) Abad et al. (2004) Chang (2004) Troutt et al. (2005)
Tax administration	Serra (2003)
Telecommunications	Guedes de Avellar et al. (2002) Uri (2004) Lam and Lam (2005) Resende and Façanha (2005)
Urban transit	De Borger et al. (2002) Dalen and Gómez-Lobo (2003) Jörss et al. (2004) Odeck (2006)
Water distribution	Corton (2003) Tupper and Resende (2004) Aubert and Reynaud (2005) Cubbin (2005)
World Health Organization	Hollingsworth and Wildman (2003) Richardson et al. (2003) Greene (2004) Lauer et al. (2004)

Finally, the ability to quantify efficiency and productivity provides management with a control mechanism with which to monitor the performance of production units under its control. The economics, management science, and operations research literatures contain numerous examples of the use of efficiency and productivity measurement techniques for this and related purposes. However, interest in these techniques has spread far beyond their origins, as evidenced by the empirical applications referenced in table 1.1. The recent dates of these studies and the journals in which they appear demonstrate that the techniques are currently in use in fields far removed from their origins. In each of these applications, interesting and challenging issues concerning appropriate behavioral objectives and constraints, and the specification of relevant variables and their measurement, arise. These applications also illustrate the rich variety of analytical techniques that can be used in making efficiency and productivity comparisons. It is worth pondering how each of these examples deals with the long list of problems discussed in this section.

1.3 Definitions and Measures of Economic Efficiency

Economic efficiency has technical and allocative components. The technical component refers to the ability to avoid waste, either by producing as much output as technology and input usage allow or by using as little input as required by technology and output production. Thus, the analysis of technical efficiency can have an output-augmenting orientation or an input-conserving orientation. The allocative component refers to the ability to combine inputs and/or outputs in optimal proportions in light of prevailing prices. Optimal proportions satisfy the first-order conditions for the optimization problem assigned to the production unit.

Koopmans (1951) provided a formal *definition* of technical efficiency: A producer is technically efficient if an increase in any output requires a reduction in at least one other output or an increase in at least one input, and if a reduction in any input requires an increase in at least one other input or a reduction in at least one output. Thus, a technically inefficient producer could produce the same outputs with less of at least one input or could use the same inputs to produce more of at least one output.

Debreu (1951) and Farrell (1957) introduced a *measure* of technical efficiency. With an input-conserving orientation, their measure is defined as (one minus) the maximum equiproportionate (i.e., radial) reduction in all inputs that is feasible with given technology and outputs. With an output-augmenting orientation, their measure is defined as the maximum radial expansion in all outputs that is feasible with given technology and inputs. In both orientations, a value of unity indicates technical efficiency because no radial adjustment is feasible, and a value different from unity indicates the severity of technical inefficiency.

In order to relate the Debreu-Farrell measures to the Koopmans definition, and to relate both to the structure of production technology, it is useful to introduce some notation and terminology. Let producers use inputs $x = (x_1, \ldots, x_N) \in R^N_+$ to produce outputs $y = (y_1, \ldots, y_M) \in R^M_+$. Production technology can be represented by the production set

$$T = \{(y, x) : \text{ x can produce y}\}. \tag{1.1}$$

Koopmans's definition of technical efficiency can now be stated formally as $(y, x) \in T$ is technically efficient if, and only if, $(y', x') \notin T$ for $(y', -x') \geq (y, -x)$.

Technology can also be represented by input sets

$$L(y) = \{x : (y, x) \in T\}, \tag{1.2}$$

which for every $y \in R^M_+$ have input isoquants

$$I(y) = \{x : x \in L(y), \lambda x \notin L(y), \lambda < 1\} \tag{1.3}$$

and input efficient subsets

$$E(y) = \{x : x \in L(y), x' \notin L(y), x' \leq x\}, \tag{1.4}$$

and the three sets satisfy $E(y) \subseteq I(y) \subseteq L(y)$.

Shephard (1953) introduced the input distance function to provide a functional representation of production technology. The input distance function is

$$D_I(y, x) = \max\{\lambda : (x/\lambda) \in L(y)\}. \tag{1.5}$$

For $x \in L(y), D_I(y, x) \geqq 1$, and for $x \in I(y)$, $D_I(y, x) = 1$. Given standard assumptions on T, the input distance function $D_I(y, x)$ is nonincreasing in y and is nondecreasing, homogeneous of degree +1, and concave in x.

The Debreu-Farrell input-oriented measure of technical efficiency TE_I can now be given a somewhat more formal interpretation as the value of the function

$$TE_I(y, x) = \min\{\theta : \theta x \in L(y)\}, \tag{1.6}$$

and it follows from (1.5) that

$$TE_I(y, x) = 1/D_I(y, x). \tag{1.7}$$

For $x \in L(y)$, $TE_I(y, x) \leqq 1$, and for $x \in I(y)$, $TE_I(y, x) = 1$.

Since so much of efficiency measurement is oriented toward output augmentation, it is useful to replicate the above development in that direction. Production technology can be represented by output sets

$$P(x) = \{y : (x, y) \in T\}, \tag{1.8}$$

which for every $x \in R_+^N$ has output isoquants

$$I(x) = \{y : y \in P(x), \lambda y \notin P(x), \lambda > 1\} \tag{1.9}$$

and output efficient subsets

$$E(x) = \{y : y \in P(x), y' \notin P(x), y' \geq y\}, \tag{1.10}$$

and the three sets satisfy $E(x) \subseteq I(x) \subseteq P(x)$.

Shephard's (1970) output distance function provides another functional representation of production technology. The output distance function is

$$D_o(x, y) = \min\{\lambda : (y/\lambda) \in P(x)\}. \tag{1.11}$$

For $y \in P(x)$, $D_o(x, y) \leqq 1$, and for $y \in I(x)$, $D_o(x, y) = 1$. Given standard assumptions on T, the output distance function $D_o(x, y)$ is nonincreasing in x and is nondecreasing, homogeneous of degree +1, and convex in y.

The Debreu-Farrell output-oriented measure of technical efficiency TE_o can now be given a somewhat more formal interpretation as the value of the function

$$TE_o(x, y) = \max\{\phi : \phi y \in P(x)\}, \tag{1.12}$$

and it follows from (1.11) that

$$TE_o(x, y) = [D_o(x, y)]^{-1}. \tag{1.13}$$

For $y \in P(x)$, $TE_o(x, y) \geqq 1$, and for $y \in I(x)$, $TE_o(x, y) = 1$. [Caution: some authors replace (1.12) and (1.13) with $TE_o(x, y) = [\max\{\phi : \phi y \in P(x)\}]^{-1} = D_o(x, y)$, so that $TE_o(x, y) \leqq 1$ just as $TE_I(y, x) \leqq 1$. We follow the convention of defining efficiency of any sort as the ratio of optimal to actual. Consequently, $TE_I(y, x) \leqq 1$ and $TE_o(x, y) \geqq 1$.]

The foregoing analysis presumes that $M > 1$, $N > 1$. In the single input case,

$$D_I(y, x) = x/g(y) \geqq 1 \Longleftrightarrow x \geqq g(y), \tag{1.14}$$

where $g(y) = \min\{x : x \in L(y)\}$ is an input requirement frontier that defines the minimum amount of scalar input x required to produce output vector y. In this case, the input-oriented measure of technical efficiency (1.7) becomes the ratio of minimum to actual input

$$TE_I(y, x) = 1/D_I(y, x) = g(y)/x \leqq 1. \tag{1.15}$$

In the single output case,

$$D_o(x, y) = y/f(x) \leqq 1 \Longleftrightarrow y \leqq f(x), \tag{1.16}$$

where $f(x) = \max\{y : y \in P(x)\}$ is a production frontier that defines the maximum amount of scalar output that can be produced with input vector x. In this

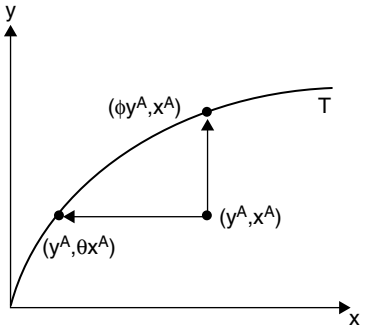


Figure 1.4. Technical Efficiency

case, the output-oriented measure of technical efficiency in (1.13) becomes the ratio of maximum to actual output

$$TE_o(x, y) = [D_o(x, y)]^{-1} = f(x)/y \geqq 1. \tag{1.17}$$

The two technical efficiency measures are illustrated in figures 1.4–1.6. As a preview of things to come, technology is smooth in figure 1.4 and piecewise linear in figures 1.5 and 1.6. This reflects different approaches to using data to estimate technology. The econometric approach introduced in section 1.5 and developed in chapter 2 estimates smooth parametric frontiers, while the mathematical programming approach introduced in section 1.6 and developed in chapter 3 estimates piecewise linear nonparametric frontiers.

In figure 1.4, producer A is located on the interior of T, and its efficiency can be measured horizontally with an input-conserving orientation using (1.6) or vertically with an output-augmenting orientation using (1.12). If an input orientation is selected, $TE_I(y^A, x^A) = \theta x^A / x^A \leqq 1$, while if an output orientation is selected, $TE_o(x^A, y^A) = \phi y^A / y^A \geqq 1$.

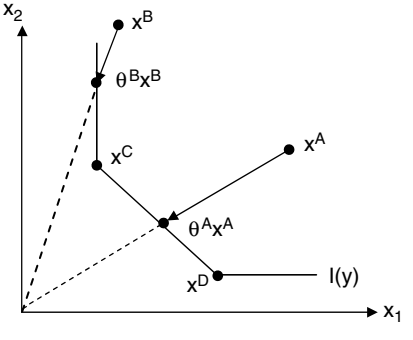


Figure 1.5. Input-Oriented Technical Efficiency

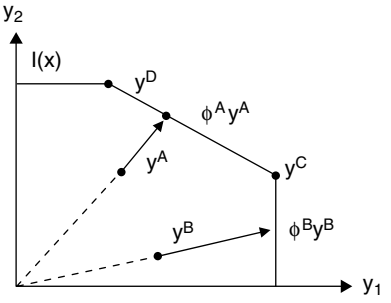


Figure 1.6. Output-Oriented Technical Efficiency

It is also possible to combine the two directions by simultaneously expanding outputs and contracting inputs, either hyperbolically or along a right angle, to arrive at an efficient point on the surface of T between $(y^A, \theta x^A)$ and $(\phi y^A, x^A)$. A hyperbolic measure of technical efficiency TE is defined as

$$TE_H(y, x) = \max\{\alpha : (\alpha y, x/\alpha) \in T\} \geqq 1, \tag{1.18}$$

and $TE_H(y, x)$ is the reciprocal of a hyperbolic distance function $D_H(y, x)$. Under constant returns to scale, $TE_H(y, x) = [TE_o(x, y)]^2 = [TE_I(y, x)]^{-2}$, and $TE_H(y, x)$ is dual to a profitability function. One version of a directional measure of technical efficiency is defined as

$$TE_D(y, x) = \max\{\beta : [(1 + \beta)x] \in T\} \geqq 0, \tag{1.19}$$

and $TE_D(y, x)$ is equal to a directional distance function $D_D(y, x)$. Even without constant returns to scale, $TE_D(y, x)$ can be related to $TE_o(x, y)$ and $TE_I(y, x)$ and is dual to a profit function. The directional measure and its underlying directional distance function are employed to good advantage in chapter 5.

In figure 1.5, input vectors x^A and x^B are on the interior of L(y), and both can be contracted radially and still remain capable of producing output vector y. Input vectors x^C and x^D cannot be contracted radially and still remain capable of producing output vector y because they are located on the input isoquant I(y). Consequently, $TE_I(y, x^C) = TE_I(y, x^D) = 1 > \max\{TE_I(y, x^A), TE_I(y, x^B)\}$. Since the radially scaled input vector $\theta^B x^B$ contains slack in input x_2, there may be some hesitancy in describing input vector $\theta^B x^B$ as being technically efficient in the production of output vector y. No such problem occurs with radially scaled input vector $\theta^A x^A$. Thus, $TE_I(y, \theta^A x^A) = TE_I(y, \theta^B x^B) = 1$ even though $\theta^A x^A \in E(y)$ but $\theta^B x^B \notin E(y)$.

Figure 1.6 tells exactly the same story, but with an output orientation. Output vectors y^C and y^D are technically efficient given input usage x, and output vectors y^A and y^B are not. Radially scaled output vectors $\phi^A y^A$ and $\phi^B y^B$

are technically efficient, even though slack in output y_2 remains at $\phi^B y^B$. Thus, $TE_o(x, \phi^A y^A) = TE_o(x, \phi^B y^B) = 1$ even though $\phi^A y^A \in E(x)$ but $\phi^B y^B \notin E(x)$.

The Debreu-Farrell measures of technical efficiency are widely used. Since they are reciprocals of distance functions, they satisfy several nice properties [as noted first by Shephard (1970) and most thoroughly by Russell (1988, 1990)]. Among these properties are the following:

- $TE_I(y, x)$ is homogeneous of degree -1 in inputs, and $TE_o(x, y)$ is homogeneous of degree -1 in outputs.
- $TE_I(y, x)$ is weakly monotonically decreasing in inputs, and $TE_o(x, y)$ is weakly monotonically decreasing in outputs.
- $TE_I(y, x)$ and $TE_o(x, y)$ are invariant with respect to changes in units of measurement.

On the other hand, they are not perfect. A notable feature of the Debreu-Farrell measures of technical efficiency is that they do not coincide with Koopmans's definition of technical efficiency. Koopmans's definition is demanding, requiring the absence of coordinatewise improvements (simultaneous membership in both efficient subsets), while the Debreu-Farrell measures require only the absence of radial improvements (membership in isoquants). Thus, although the Debreu-Farrell measures correctly identify all Koopmans-efficient producers as being technically efficient, they also identify as being technically efficient any other producers located on an isoquant outside the efficient subset. Consequently, Debreu-Farrell technical efficiency is necessary, but not sufficient, for Koopmans technical efficiency. The possibilities are illustrated in figures 1.5 and 1.6, where $\theta^B x^B$ and $\phi^B y^B$ satisfy the Debreu-Farrell conditions but not the Koopmans requirement because slacks remain at the optimal radial projections.

Much has been made of this property of the Debreu-Farrell measures, but we think the problem is exaggerated. The practical significance of the problem depends on how many observations lie outside the cone spanned by the relevant efficient subset. Hence, the problem disappears in much econometric analysis, in which the parametric form of the function used to estimate production technology (e.g., Cobb-Douglas, but not flexible functional forms such as translog) imposes equality between isoquants and efficient subsets, thereby eliminating slack by assuming it away. The problem assumes greater significance in the mathematical programming approach, in which the nonparametric form of the frontier used to estimate the boundary of the production set imposes slack by a strong (or free) disposability assumption. If the problem is deemed significant in practice, then it is possible to report Debreu-Farrell efficiency scores and slacks separately, side by side. This is rarely done. Instead, much effort has been directed toward finding a "solution" to the problem. Three strategies have been proposed:

- Replace the radial Debreu-Farrell measure with a nonradial measure that projects to efficient subsets (Färe and Lovell, 1978). This guarantees that

an observation (or its projection) is technically efficient if, and only if, it is efficient in Koopmans's sense. However nonradial measures gain this "indication" property at the considerable cost of failing the homogeneity property.
- Develop a measure that incorporates slack and the radial component into an inclusive measure of technical efficiency (Cooper et al., 1999). This measure also gains the indication property, but it has its own problems, including the possibility of negative values.
- Eliminate slack altogether by enforcing strictly positive marginal rates of substitution and transformation. We return to this possibility in section 1.6.4, in a different setting.

Happily, there is no such distinction between definitions and measures of economic efficiency. Defining and measuring economic efficiency require the specification of an economic objective and information on relevant prices. If the objective of a production unit (or the objective assigned to it by the analyst) is cost minimization, then a measure of cost efficiency is provided by the ratio of minimum feasible cost to actual cost. This measure depends on input prices. It attains a maximum value of unity if the producer is cost efficient, and a value less than unity indicates the degree of cost inefficiency. A measure of input-allocative efficiency is obtained residually as the ratio of the measure of cost efficiency to the input-oriented measure of technical efficiency. The modification of this Farrell decomposition of cost efficiency to the output-oriented problem of decomposing revenue efficiency is straightforward. Modifying the procedure to accommodate alternative behavioral objectives is sometimes straightforward and occasionally challenging. So is the incorporation of regulatory and other nontechnological constraints that impede the pursuit of some economic objective.

Suppose that producers face input prices $w = (w_1, \ldots, w_N) \in R^N_{++}$ and seek to minimize cost. Then, a minimum cost function, or a cost frontier, is defined as

$$c(y, w) = \min_x \{w^T x : D_I(y, x) \geqq 1\}. \tag{1.20}$$

If the input sets $L(y)$ are closed and convex, and if inputs are freely disposable, the cost frontier is dual to the input distance function in the sense of (1.20) and

$$D_I(y, x) = \min_w \{w^T x : c(y, w) \geqq 1\}. \tag{1.21}$$

A measure of cost efficiency CE is provided by the ratio of minimum cost to actual cost:

$$CE(x, y, w) = c(y, w)/w^T x \tag{1.22}$$

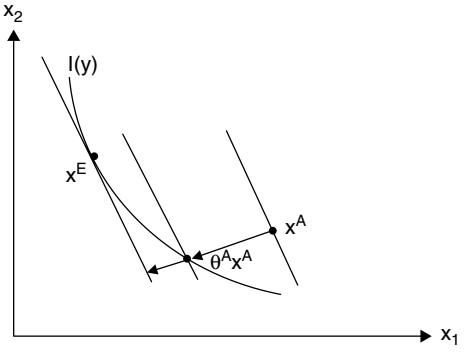


Figure 1.7. Cost Efficiency I

A measure of input-allocative efficiency AE_I is obtained from (1.6) and (1.22) as

$$AE_I(x, y, w) = CE(x, y, w)/TE_I(y, x). \tag{1.23}$$

$CE(x, y, w)$ and its two components are bounded above by unity, and $CE(x, y, w) = TE_I(y, x) \times AE_I(x, y, w)$.

The measurement and decomposition of cost efficiency is illustrated in figures 1.7 and 1.8. In figure 1.7, the input vector x^E minimizes the cost of producing output vector y at input prices w, so $w^T x^E = c(y, w)$. The cost efficiency of x^A is given by the ratio $w^T x^E / w^T x^A = c(y, w)/w^T x^A$. The Debreu-Farrell measure of the technical efficiency of x^A is given by $\theta^A = \theta^A x^A / x^A = w^T(\theta^A x^A)/w^T x^A$. The allocative efficiency of x^A is determined residually as the ratio of cost efficiency to technical efficiency, or by the ratio $w^T x^E / w^T(\theta^A x^A)$. The magnitudes of technical, allocative, and cost inefficiency are all measured by ratios of price-weighted input vectors. The

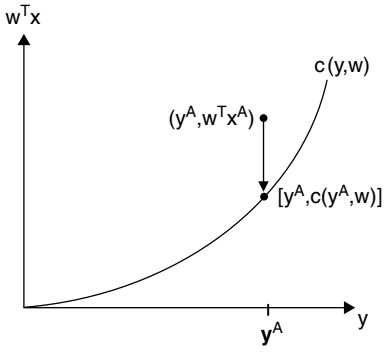


Figure 1.8. Cost Efficiency II

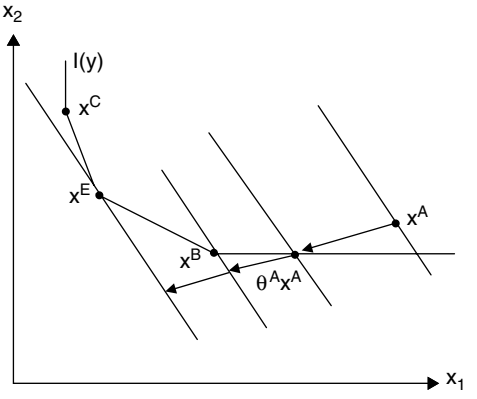


Figure 1.9. Cost Efficiency III

direction of allocative inefficiency is revealed by the input vector difference $(x^E - \theta^A x^A)$. An alternative view of cost efficiency is provided by figure 1.8, in which $CE(x^A, y^A, w) = c(y^A, w)/w^T x^A$.

The measurement and decomposition of cost efficiency are illustrated again in figure 1.9, for the case in which the efficient subset is a proper subset of the isoquant. The analysis proceeds as above, with a twist. The cost efficiency of input vector x^A now has three components, a radial technical component $[w^T(\theta^A x^A)/w^T x^A]$, an input slack component $[w^T x^B/w^T(\theta^A x^A)]$, and an allocative component $(w^T x^E/w^T x^B)$. With input price data, all three components can be identified, although they rarely are. The slack component is routinely assigned to the allocative component.

Suppose next that producers face output prices $p = (p_1, \ldots, p_M) \in R^M_{++}$ and seek to maximize revenue. Then, a maximum revenue function, or a revenue frontier, is defined as

$$r(x, p) = \max_y \{p^T y : D_o(x, y) \leqq 1\}. \tag{1.24}$$

If the output sets $P(x)$ are closed and convex, and if outputs are freely disposable, the revenue frontier is dual to the output distance function in the sense of (1.24) and

$$D_o(x, y) = \max_p \{p^T y : r(x, p) \leqq 1\}. \tag{1.25}$$

A measure of revenue efficiency RE is provided by the ratio of maximum revenue to actual revenue:

$$RE(y, x, p) = r(x, p)/p^T y \tag{1.26}$$

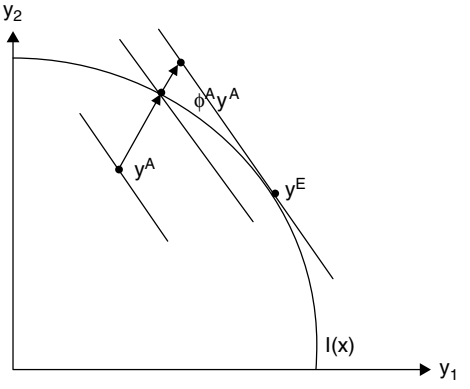


Figure 1.10. Revenue Efficiency I

A measure of output-allocative efficiency AE_o is obtained from (1.12) and (1.26) as

$$AE_o(y, x, p) = RE(y, x, p)/TE_o(x, y). \tag{1.27}$$

RE(y, x, p) and its two components are bounded below by unity, and $RE(y, x, p) = TE_o(x, y) \times AE_o(y, x, p)$.

The measurement and decomposition of revenue efficiency in figures 1.10 and 1.11 follow exactly the same steps. The measurement and decomposition of revenue efficiency in the presence of output slack follow along similar lines as in figure 1.9. Revenue loss attributable to output slack is typically assigned to the output-allocative efficiency component of revenue efficiency.

Cost efficiency and revenue efficiency are important performance indicators, but each reflects just one dimension of a firm's overall performance. A measure of profit efficiency captures both dimensions and relates directly to

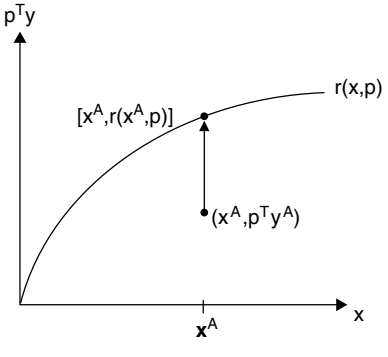


Figure 1.11. Revenue Efficiency II

the bottom line discussed in section 1.1. Suppose that producers face output prices $p \in R^M_{++}$ and input prices $w \in R^N_{++}$ and seek to maximize profit. The maximum profit function, or profit frontier, is defined as

$$\pi(p, w) = \max_{y,x} \left\{ \left(p^T y - w^T x \right) : (y, x) \in T \right\}. \tag{1.28}$$

If the production set T is closed and convex, and if outputs and inputs are freely disposable, the profit frontier is dual to T in the sense of (1.28) and

$$T = \{(y, x) : (p^T y - w^T x) \leqq \pi(p, w) \forall p \in R^M_{++}, w \in R^N_{++}\}. \tag{1.29}$$

A measure of profit efficiency is provided by the ratio of maximum profit to actual profit

$$\pi E(y, x, p, w) = \pi(p, w)/(p^T y - w^T x), \tag{1.30}$$

provided $(p^T y - w^T x) > 0$, in which case $\pi E(y, x, p, w)$ is bounded below by unity. The decomposition of profit efficiency is partially illustrated by figure 1.12, which builds on figure 1.4. Profit at (y^A, x^A) is less than maximum profit at (y^E, x^E), and two possible decompositions of profit efficiency are illustrated. One takes an input-conserving orientation to the measurement of technical efficiency, and the residual allocative component follows the path from $(y^A, \theta x^A)$ to (y^E, x^E). The other takes an output-augmenting orientation to the measurement of technical efficiency, with residual allocative component following the path from $(\phi y^A, x^A)$ to (y^E, x^E). In both approaches the residual allocative component contains an input-allocative efficiency component and an output-allocative efficiency component, although the magnitudes of each component can differ in the two approaches. These two components are hidden from view in the two-dimensional figure 1.12. In both approaches, the residual allocative efficiency component also includes a scale component, which is illustrated in figure 1.12. The direction of the scale component is sensitive to the orientation of the technical efficiency component, which imposes a burden on the analyst to get the orientation right. Because profit efficiency

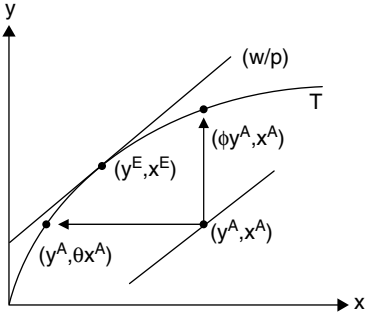


Figure 1.12. Profit Efficiency

involves adjustments to both outputs and inputs, hyperbolic and directional technical efficiency measures are appealing in this context. Whatever the orientation of the technical efficiency measure, profit inefficiency is attributable to technical inefficiency, to an inappropriate scale of operation, to the production of an inappropriate output mix, and to the selection of an inappropriate input mix.

We conclude this section with a brief discussion of dominance. Producer A dominates all other producers for which $(y^A, -x^A) \geq (y, -x)$. This notion is a direct application of Koopmans's definition of efficiency, in which producer A is "more efficient" than all other producers it dominates. Reversing the definition, producer A is dominated by all other producers for which $(y, -x) \geq (y^A, -x^A)$. In figure 1.4, producer A is dominated by all producers to the northwest $\in T$ because they use no more input to produce at least as much output. Similar dominance relationships can be constructed in figures 1.5 and 1.6. In each case, dominance is a physical, or technical, relationship. However, dominance can also be given a value interpretation. In figure 1.8, producer A is dominated (in a cost sense) by all other producers to the southeast on or above $c(y, w)$ because they produce at least as much output at no more cost, and in figure 1.11 producer A is dominated (in a revenue sense) by all other producers to the northwest on or beneath $r(x, p)$ because they use no more input to generate at least as much revenue.

Dominance is an underutilized concept in the field of producer performance evaluation, where the emphasis is on efficiency. This neglect is unfortunate, because dominance information offers a potentially useful complement to an efficiency evaluation, as Tulkens and Vanden Eeckaut (1995, 1999) have demonstrated. Inefficient producers can have many dominators and hence many potential role models from which to learn. To cite one example, Fried et al. (1993) report an average of 22 dominators for each of nearly 9,000 U.S. credit unions.

The identification of dominators can constitute the initial step in a benchmarking exercise. It is possible that dominators utilize superior business practices that are transferable to the benchmarking producer. However, it is also possible that dominance is due to a more favorable operating environment. Although this may be cold comfort to the benchmarking business, it can be very useful to the analyst who does not want to confuse variation in performance with variation in the operating environment. Incorporating variation in the operating environment is an important part of any performance evaluation exercise, and techniques for doing so are discussed below and in subsequent chapters.

1.4 Techniques for Efficiency Measurement

Efficiency measurement involves a comparison of actual performance with optimal performance located on the relevant frontier. Since the true frontier

is unknown, an empirical approximation is needed. The approximation is frequently dubbed a "best-practice" frontier.

The economic theory of production is based on production frontiers and value duals such as cost, revenue, and profit frontiers and on envelope properties yielding cost-minimizing input demands, revenue-maximizing output supplies, and profit-maximizing output supplies and input demands. Emphasis is placed on optimizing behavior subject to constraint. However, for more than 75 years, at least since Cobb and Douglas started running regressions, the empirical analysis of production has been based on a least squares statistical methodology by which estimated functions of interest pass through the data and estimate mean performance. Thus, the frontiers of theory have become the functions of analysis, interest in enveloping data with frontiers has been replaced with the practice of intersecting data with functions, and unlikely efficient outcomes have been neglected in favor of more likely but less efficient outcomes, all as attention has shifted from extreme values to central tendency.

If econometric analysis is to be brought to bear on the investigation of the structure of economic frontiers, and on the measurement of efficiency relative to these frontiers, then conventional econometric techniques require modification. The modifications that have been developed, improved, and implemented in the last three decades run the gamut from trivial to sophisticated. Econometric techniques are introduced in section 1.5 and developed in detail in chapter 2.

In sharp contrast to econometric techniques, mathematical programming techniques are inherently enveloping techniques, and so they require little or no modification to be employed in the analysis of efficiency. This makes them appealing, but they went out of favor long ago in the economics profession. Their theoretical appeal has given way to a perceived practical disadvantage: their ostensible failure to incorporate the statistical noise that drives conventional econometric analysis. This apparent shortcoming notwithstanding, they remain popular in the fields of management science and operations research, and they are making a comeback in economics. Programming techniques are introduced in section 1.6 and developed in detail in chapter 3.

The econometric approach to the construction of frontiers and the estimation of efficiency relative to the constructed frontiers has similarities and differences with the mathematical programming approach. Both are analytically rigorous benchmarking exercises that exploit the distance functions introduced in section 1.3 to measure efficiency relative to a frontier. However, the two approaches use different techniques to envelop data more or less tightly in different ways. In doing so, they make different accommodations for statistical noise and for flexibility in the structure of production technology. It is these two different accommodations that have generated debate about the relative merits of the two approaches. At the risk of oversimplification, the differences between the two approaches boil down to two essential features:

- The econometric approach is stochastic. This enables it to attempt to distinguish the effects of noise from those of inefficiency, thereby providing the basis for statistical inference.
- The programming approach is nonparametric. This enables it to avoid confounding the effects of misspecification of the functional form (of both technology and inefficiency) with those of inefficiency.

A decade or more ago, the implication drawn from these two features was that the programming approach was nonstochastic and the econometric approach was parametric. This had a disturbing consequence. If efficiency analysis is to be taken seriously, producer performance evaluation must be robust to both statistical noise and specification error. Neither approach was thought to be robust to both.

Happily, knowledge has progressed and distinctions have blurred. To praise one approach as being stochastic is not to deny that the other is stochastic, as well, and to praise one approach as being nonparametric is not to damn the other as being rigidly parameterized. Recent explorations into the statistical foundations of the programming approach have provided the basis for statistical inference, and recent applications of flexible functional forms and semiparametric, nonparametric, and Bayesian techniques have freed the econometric approach from its parametric straitjacket. Both techniques are more robust than previously thought. The gap is no longer between one technique and the other, but between best-practice knowledge and average practice implementation. The challenge is to narrow the gap.

It is worth asking whether the two techniques tell consistent stories when applied to the same data. The answer seems to be that the higher the quality of the data, the greater the concordance between the two sets of efficiency estimates. Of the many comparisons available in the literature, we recommend Bauer et al. (1998), who use U.S. banking data, and Cummins and Zi (1998), who use U.S. life insurance company data. Both studies find strong positive rank correlations of point estimates of efficiency between alternative pairs of econometric models and between alternative pairs of programming models, and weaker but nonetheless positive rank correlations of point estimates of efficiency between alternative pairs of econometric and programming models.

Chapters 2 and 3 develop the two approaches, starting with their basic formulations and progressing to more advanced methods. Chapter 4 recasts the parametric econometric approach of chapter 2 into a nonparametric statistical framework and explores the statistical foundations of the programming approach of chapter 3. In addition to these chapters, we recommend comprehensive treatments of the econometric approach by Kumbhakar and Lovell (2000) and of the programming approach by Cooper et al. (2000). Both contain extensive references to analytical developments and empirical applications.

1.5 The Econometric Approach to Efficiency Measurement

Econometric models can be categorized according to the type of data they use (cross section or panel), the type of variables they use (quantities only, or quantities and prices), and the number of equations in the model. In section 1.5.1, we discuss the most widely used model: the single-equation cross-section model. In section 1.5.2, we progress to panel-data models. In both contexts, the efficiency being estimated can be either technical or economic. In section 1.5.3, we discuss multiple equation models, and in section 1.5.4, we discuss shadow price models, which typically involve multiple equations. In these two contexts, the efficiency being estimated is economic, with a focus on allocative inefficiency and its cost.

1.5.1 Single-equation cross-section models

Suppose producers use inputs $x \in R_+^N$ to produce scalar output $y \in R_+$, with technology

$$y_i \leqq f(x_i;\beta)\exp\{v_i\}, \tag{1.31}$$

where β is a parameter vector characterizing the structure of production technology and $i = 1, \ldots, I$ indexes producers. The deterministic production frontier is $f(x_i;\beta)$. Observed output y_i is bounded above by the stochastic production frontier $[f(x_i;\beta)\exp\{v_i\}]$, with the random disturbance term $v_i \gtreqless 0$ included to capture the effects of statistical noise on observed output. The stochastic production frontier reflects what is possible $[f(x_i;\beta)]$ in an environment influenced by external events, favorable and unfavorable, beyond the control of producers $[\exp\{v_i\}]$.

The weak inequality in (1.31) can be converted to an equality through the introduction of a second disturbance term to create

$$y_i = f(x_i;\beta)\exp\{v_i - u_i\}, \tag{1.32}$$

where the disturbance term $u_i \geqq 0$ is included to capture the effect of technical inefficiency on observed output.

Recall from section 1.3 that the Debreu-Farrell output-oriented measure of technical efficiency is the ratio of maximum possible output to actual output (and that some authors use the reciprocal of this measure). Applying definition (1.17) to (1.32) yields

$$TE_o(x_i, y_i) = f(x_i;\beta)\exp\{v_i\}/y_i = \exp\{u_i\} \geqq 1, \tag{1.33}$$

because $u_i \geqq 0$. The problem is to estimate $TE_o(x_i, y_i)$. This requires estimation of (1.32), which is easy and can be accomplished in a number of ways depending on the assumptions one is willing to make. It also requires a decomposition of the residuals into separate estimates of v_i and u_i, which is not so easy.

One approach, first suggested by Winsten (1957) and now known as corrected ordinary least squares (COLS), is to assume that $u_i = 0, i = 1, \ldots, I$, and that $v_i \sim N(0, \sigma_v^2)$. In this case (1.32) collapses to a standard regression model that can be estimated consistently by OLS. The estimated production function, which intersects the data, is then shifted upward by adding the maximum positive residual to the estimated intercept, creating a production frontier that bounds the previous data. The residuals are corrected in the opposite direction and become $\hat{v}_i = v_i - v_i^{max} \leqq 0$, $i = 1, \ldots, I$. The technical efficiency of each producer is estimated from

$$\hat{TE}_o(x_i, y_i) = \exp\{-\hat{v}_i\} \geqq 1, \tag{1.34}$$

and $\hat{TE}_o(x_i, y_i) - 1 \geqq 0$ indicates the percentage by which output can be expanded, on the assumption that $u_i = 0, i = 1, \ldots, I$.

The producer having the largest positive OLS residual supports the COLS production frontier. This makes COLS vulnerable to outliers, although ad hoc sensitivity tests have been proposed. In addition, the structure of the COLS frontier is identical to the structure of the OLS function, apart from the shifted intercept. This structural similarity rules out the possibility that efficient producers are efficient precisely because they exploit available economies and substitution possibilities that average producers do not. The assumption that best practice is just like average practice, but better, defies both common sense and much empirical evidence. Finally, it is troubling that efficiency estimates for all producers are obtained by suppressing the inefficiency error component u_i and are determined exclusively by the single producer having the most favorable noise v_i^{max}. The term $\exp\{u_i\}$ in (1.33) is proxied by the term $\exp\{-\hat{v}\}$ in (1.34). Despite these reservations, and additional concerns raised in chapters 2 and 4, COLS is widely used, presumably because it is easy.

A second approach, suggested by Aigner and Chu (1968), is to make the opposite assumption that $v_i = 0, i = 1, \ldots, I$. In this case, (1.32) collapses to a deterministic production frontier that can be estimated by linear or quadratic programming techniques that minimize either $\Sigma_i u_i$ or $\Sigma_i u_i^2$, subject to the constraint that $u_i = \ln[f(x_i; \beta)/y_i] \geqq 0$ for all producers. The technical efficiency of each producer is estimated from

$$\hat{TE}_o(x_i, y_i) = \exp\{\hat{u}_i\} \geqq 1, \tag{1.35}$$

and $\hat{TE}_o(x_i, y_i) - 1 \geqq 0$ indicates the percentage by which output can be expanded, on the alternative assumption that $v_i = 0, i = 1, \ldots, I$. The $\hat{u}_i$ values are estimated from the slacks in the constraints $[\ln f(x_i; \beta) - \ln y_i \geqq 0, i = 1, \ldots I]$ of the program. Although it appears that the term $\exp\{\hat{u}_i\}$ in (1.35) coincides with the term $\exp\{u_i\}$ in (1.33), the expression in (1.35) is conditioned on the assumption that $v_i = 0$, while the expression in (1.33) is not. In addition, since no distributional assumption is imposed on $u_i \geqq 0$, statistical inference is precluded, and consistency cannot be verified. However, Schmidt (1976) showed that the linear programming "estimate" of β

is maximum likelihood (MLE) if the u_i values follow an exponential distribution, and that the quadratic programming "estimate" of β is maximum likelihood if the u_i values follow a half-normal distribution. Unfortunately, we know virtually nothing about the statistical properties of these estimators, even though they are maximum likelihood. However, Greene (1980) showed that an assumption that the u_i values follow a gamma distribution generates a well-behaved likelihood function that allows statistical inference, although this model does not correspond to any known programming problem. Despite the obvious statistical drawback resulting from its deterministic formulation, the programming approach is also widely used. One reason for its popularity is that it is easy to append monotonicity and curvature constraints to the program, as Hailu and Veeman (2000) have done in their study of water pollution in the Canadian pulp and paper industry.

The third approach, suggested independently by Aigner et al. (1977) and Meeusen and van den Broeck (1977), attempts to remedy the shortcomings of the first two approaches and is known as stochastic frontier analysis (SFA). In this approach, it is assumed that $v_i \sim N(0, \sigma_v^2)$ and that $u_i \geqq 0$ follows either a half-normal or an exponential distribution. The motive behind these two distributional assumptions is to parsimoniously parameterize the notion that relatively high efficiency is more likely than relatively low efficiency. After all, the structure of production is parameterized, so we might as well parameterize the inefficiency distribution, too. In addition, it is assumed that the v_i and the u_i values are distributed independently of each other and of x_i. OLS can be used to obtain consistent estimates of the slope parameters but not the intercept, because $E(v_i - u_i) = E(-u_i) \leqq 0$. However the OLS residuals can be used to test for negative skewness, which is a test for the presence of variation in technical inefficiency. If evidence of negative skewness is found, OLS slope estimates can be used as starting values in a maximum likelihood routine.

Armed with the distributional and independence assumptions, it is possible to derive the likelihood function, which can be maximized with respect to all parameters (β, σ_v^2, and σ_u^2) to obtain consistent estimates of β. However, even with this information, neither team was able to estimate $TE_o(x_i, y_i)$ in (1.33) because they were unable to disentangle the separate contributions of v_i and u_i to the residual. Jondrow et al. (1982) provided an initial solution, by deriving the conditional distribution of $[-u_i|(v_i - u_i)]$, which contains all the information $(v_i - u_i)$ contains about $-u_i$. This enabled them to derive the expected value of this conditional distribution, from which they proposed to estimate the technical efficiency of each producer from

$$\hat{TE}_o(x_i, y_i) = \{\exp\{E[-\hat{u}_i|(v_i - u_i)]\}\}^{-1} \geqq 1, \tag{1.36}$$

which is a function of the MLE parameter estimates. Later, Battese and Coelli (1988) proposed to estimate the technical efficiency of each producer from

$$\hat{TE}_o(x_i, y_i) = \{E[\exp\{-\hat{u}_i\}|(v_i - u_i)]\}^{-1} \geqq 1, \tag{1.37}$$

which is a slightly different function of the same MLE parameter estimates and is preferred because $-\hat{u}_i$ in (1.36) is only the first-order term in the power series approximation to $\exp\{-\hat{u}_i\}$ in (1.37).

Unlike the first two approaches, which suppress either u_i or v_i, SFA sensibly incorporates both noise and inefficiency into the model specification. The price paid is the need to impose distributional and independence assumptions, the prime benefit being the ability to disentangle the two error components. The single parameter half-normal and exponential distributions can be generalized to more flexible two-parameter truncated normal and gamma distributions, as suggested by Stevenson (1980) and Greene (1980), although they rarely are. The independence assumptions seem essential to the MLE procedure. The fact that they can be relaxed in the presence of panel data provides an initial appreciation of the value of panel data, to which we return in section 1.5.2.

The efficiency estimates obtained from (1.36) and (1.37) are unbiased, but their consistency has been questioned, not because they converge to the wrong values, but because in a cross section we get only one look at each producer, and the number of looks cannot increase. However, a new contrary claim of consistency is put forth in chapter 2. The argument is simple and runs as follows: The technical efficiency estimates in (1.36) and (1.37) are conditioned on MLEs of $(v_i - u_i) = \ln y_i - \ln f(x_i; \beta)$, and since β is estimated consistently by MLE, so is technical efficiency, even in a cross section.

For more than a decade, individual efficiencies were estimated using either (1.36) or (1.37). Hypothesis tests frequently were conducted on β and occasionally on σ_u^2/σ_v^2 (or some variant thereof) to test the statistical significance of efficiency variation. However, we did not test hypotheses on either estimator of $TE_o(x_i, y_i)$ because we did not realize that we had enough information to do so. We paid the price of imposing distributions on v_i and u_i, but we did not reap one of the benefits: We did not exploit the fact that distributions imposed on v_i and u_i create distributions for $[-u_i \mid (v_i - u_i)]$ and $[\exp\{-u_i\} \mid (v_i - u_i)]$, which can be used to construct confidence intervals and to test hypotheses on individual efficiencies. This should have been obvious all along, but Horrace and Schmidt (1996) and Bera and Sharma (1999) were the first to develop confidence intervals for efficiency estimators. The published confidence intervals we have seen are depressingly wide, presumably because estimates of σ_u^2/σ_v^2 are relatively small. In such circumstances, the information contained in a ranking of estimated efficiency scores is limited, frequently to the ability to distinguish stars from strugglers.

The preceding discussion has been based on a single output production frontier. However, multiple outputs can be incorporated in a number of ways:

- Estimate a stochastic revenue frontier, with $p^T y$ replacing y and (x, p) replacing x in (1.32). The one-sided error component provides the basis for a measure of revenue efficiency. Applications are rare.

- Estimate a stochastic profit frontier, with $(p^T y - w^T x)$ replacing y and (p, w) replacing x in (1.32). The one-sided error component provides the basis for a measure of profit efficiency. Estimation of profit frontiers is popular, especially in the financial institutions literature. Berger and Mester (1997) provide an extensive application to U.S. banks.
- Estimate a stochastic cost frontier, with $w^T x$ replacing y and (y, w) replacing x in (1.32). Since $w^T x \geqq c(y, w; \beta) \exp\{v_i\}$, this requires changing the sign of the one-sided error component, which provides the basis for a measure of cost efficiency. Applications are numerous.
- Estimate a stochastic input requirement frontier, with the roles of x and y in (1.32) being reversed. This also requires changing the sign of the one-sided error component, which provides the basis for a measure of input use efficiency. Applications are limited to situations in which labor has a very large (variable?) cost share, or in which other inputs are not reported. Kumbhakar and Hjalmarsson (1995) provide an application to employment in Swedish social insurance offices.
- Estimate a stochastic output distance function $D_o(x, y) \exp\{v_i\} \leqq 1 \Rightarrow D_o(x_i, y_i; \beta) \exp\{v_i - u_i\} = 1, u_i \geqq 0$. The one-sided error component provides the basis for an output-oriented measure of technical efficiency. Unlike the models above, a distance function has no natural dependent variable, and at least three alternatives have been proposed. Fuentes et al. (2001) and Atkinson et al. (2003) illustrate alternative specifications and provide applications to Spanish insurance companies and U.S. railroads, respectively.
- Estimate a stochastic input distance function $D_I(y, x) \exp\{v_i\} \geqq 1 \Rightarrow D_I(y_i, x_i; \beta) \exp\{v_i + u_i\} = 1, u_i \geqq 0$. Note the sign change of the one-sided error component, which provides the basis for an input-oriented measure of technical efficiency, and proceed as above.

In the preceding discussion, interest has centered on the estimation of efficiency. A second concern, first raised in section 1.2, involves the incorporation of potential determinants of efficiency. The determinants can include characteristics of the operating environment and characteristics of the manager such as human capital endowments. The logic is that if efficiency is to be improved, we need to know what factors influence it, and this requires distinguishing the influences of the potential determinants from that of the inputs and outputs themselves. Two approaches have been developed:

(1) Let $z \in R^K$ be a vector of exogenous variables thought to be relevant to the production activity. One approach that has been used within and outside the frontier field is to replace $f(x_i; \beta)$ with $f(x_i, z_i; \beta, \gamma)$. The most popular example involves z serving as a proxy for technical change that shifts the production (or cost) frontier. Another popular example involves the inclusion of stage length and load factor in the analysis of airline performance; both are thought to influence operating cost. Although z is relevant in the sense that it is thought to

be an important characteristic of production activity, it does not influence the efficiency of production. The incorporation of potential influences on productive efficiency requires an alternative approach, in which z influences the distance of producers from the relevant frontier.

(2) In the old days, it was common practice to adopt a two-stage approach to the incorporation of potential determinants of productive efficiency. In this approach efficiency was estimated in the first stage using either (1.36) or (1.37), and estimated efficiencies were regressed against a vector of potential influences in the second stage. Deprins and Simar (1989) were perhaps the first to question the statistical validity of this two-stage approach. Later, Battese and Coelli (1995) proposed a single-stage model of general form

$$y_i = f(x_i;\, \beta)\exp\{v_i - u_i(z_i;\, \gamma)\}, \tag{1.38}$$

where $u_i(z_i;\, \gamma) \geqq 0$ and z is a vector of potential influences with parameter vector γ, and they showed how to estimate the model in SFA format. Later, Wang and Schmidt (2002) analyzed alternative specifications for $u_i(z_i;\, \gamma)$ in the single-stage approach; for example, either the mean or the variance of the distribution being truncated below at zero can be made a function of z_i. They also provided detailed theoretical arguments, supported by compelling Monte Carlo evidence, explaining why both stages of the old two-stage procedure are seriously biased. We hope to see no more two-stage SFA models.

1.5.2 Single-equation panel-data models

In a cross section, each producer is observed once. If each producer is observed over a period of time, panel-data techniques can be brought to bear on the problem. At the heart of the approach is the association of a "firm effect" from the panel-data literature with a one-sided inefficiency term from the frontier literature. How this association is formulated and how the model is estimated are what distinguish one model from another. Whatever the model, the principal advantage of having panel data is the ability to observe each producer more than once. It should be possible to parlay this ability into "better" estimates of efficiency than can be obtained from a single cross section.

Schmidt and Sickles (1984) were among the first to consider the use of conventional panel-data techniques in a frontier context. We follow them by writing the panel-data version of the production frontier model (1.32) as

$$y_{it} = f(x_{it};\, \beta)\exp\{v_{it} - u_i\}, \tag{1.39}$$

where a time subscript $t = 1, \ldots, T$ has been added to y, x, and v, but not (yet) to u. We begin by assuming that technical efficiency is time invariant

and not a function of exogenous influences. Four estimation strategies are available.

(1) It is straightforward to adapt the cross-section MLE procedures developed in section 1.5.1 to the panel-data context, as Pitt and Lee (1981) first showed. Allowing u_i to depend on potential influences is also straightforward, as Battese and Coelli (1995) demonstrated. Extending (1.39) by setting $u_{it} = u_{it}(z_{it}; \gamma)$ and specifying one of the elements of z_{it} as a time trend or a time dummy allows technical inefficiency to be time varying, which is especially desirable in long panels. Maximum likelihood estimators of technical efficiency obtained from (1.36) and (1.37) are consistent in T and I. However, MLE requires strong distributional and independence assumptions, and the availability of panel-data techniques enables us to relax some of these assumptions.

(2) The fixed-effects model is similar to cross-section COLS. It imposes no distributional assumption on u_i and allows the u_i values to be correlated with the v_{it} and the x_{it} values. Since the u_i values are treated as fixed, they become producer-specific intercepts $\beta_{oi} = (\beta_o - u_i)$ in (1.39), which can be estimated consistently by OLS. After estimation, the normalization $\beta_o^* = \beta_{oi}^{max}$ generates estimates of $\hat{u}_i = \beta_o^* - \beta_{oi} \geqq 0$, and estimates of producer-specific technical efficiencies are obtained from

$$\hat{TE}_o(x_i, y_i) = [\exp\{-\hat{u}_i\}]^{-1}. \tag{1.40}$$

These estimates are consistent in T and I, and they have the great virtue of allowing the u_i values to be correlated with the regressors. However, the desirable property of consistency in T is offset by the undesirability of assuming time invariance of inefficiency in long panels. In addition, the fixed-effects model has a potentially serious drawback: The firm effects are intended to capture variation in technical efficiency, but they also capture the effects of all phenomena that vary across producers but not through time, such as locational characteristics and regulatory regime.

(3) The random-effects model makes the opposite assumptions on the u_i values, which are allowed to be random, with unspecified distribution having constant mean and variance, but are assumed to be uncorrelated with the v_{it} and the x_{it} values. This allows the inclusion of time-invariant regressors in the model. Defining $\beta_o^{**} = \beta_o - E(u_i)$ and $u_i^{**} = u_i - E(u_i)$, (1.39) can be estimated by generalized least squares (GLS). After estimation, firm-specific estimates of u_i^{**} are obtained from the temporal means of the residuals. Finally, these estimates are normalized to obtain estimates of $\hat{u}_i = u_i^{**max} - u_i^{**}$, from which producer-specific estimates of technical efficiency are obtained from

$$\hat{TE}_o(x_i, y_i) = [\exp\{-\hat{u}_i\}]^{-1}. \tag{1.41}$$

These estimates also are consistent in T and I. The main virtue of GLS is that it allows the inclusion of time-invariant regressors, whose impacts would be confounded with efficiency variation in a fixed-effects model.

(4) Finally, an estimator from Hausman and Taylor (1981) can be adapted to (1.39). It is a mixture of the fixed-effects and random-effects estimators that allows the u_i values to be correlated with some, but not all, regressors and can include time-invariant regressors.

We have explored the tip of the proverbial iceberg. Panel-data econometrics is expanding rapidly, as is its application to frontier models. Details are provided in chapter 2.

1.5.3 Multiple equation models

We begin by reproducing a model popularized long ago by Christensen and Greene (1976). The model is

$$\ln(w^T x)_i = c(\ln y_i,\ \ln w_i;\ \beta) + v_i,$$
$$(w_n x_n / w^T x)_i = s_n(\ln y_i,\ \ln w_i;\ \beta) + v_{ni}, n = 1, \ldots, N-1. \qquad (1.42)$$

This system describes the behavior of a cost-minimizing producer, with the first equation being a cost function and the remaining equations exploiting Shephard's (1953) lemma to generate cost-minimizing input cost shares. The errors $(v_i,\ v_{ni})$ reflect statistical noise and are assumed to be distributed multivariate normal with zero means. The original motivation for appending the cost-share equations was to increase statistical efficiency in estimation, since they contain no parameters not appearing in the cost function. Variants on this multiple equation theme, applied to flexible functional forms such as translog, appear regularly in production (and consumption) economics.

The pursuit of statistical efficiency is laudable, but it causes difficulties when the objective of the exercise is the estimation of economic efficiency. We do not want to impose the assumption of cost minimization that drives Shephard's lemma, so we transform the Christensen-Greene model (1.42) into a stochastic cost frontier model as follows:

$$\ln(w^T x)_i = c(\ln y_i,\ \ln\ w_i;\ \beta) + v_i + T_i + A_i,$$
$$(w_n x_n / w^T x)_i = s_n(\ln y_i,\ \ln\ w_i;\ \beta) + v_{ni} + u_{ni}, n = 1, \ldots, N-1. \qquad (1.43)$$

Here, v_i and the v_{ni} capture the effects of statistical noise. $T_i \geqq 0$ reflects the cost of technical inefficiency, $A_i \geqq 0$ reflects the cost of input-allocative inefficiency, and $(T_i + A_i) \geqq 0$ is the cost of both. Finally, $u_{ni} \gtreqless 0$ captures the departures of actual input cost shares from their cost-efficient magnitudes. Since technical inefficiency is measured radially, it maintains the observed input mix and has no impact on input share equations. However, allocative inefficiency represents an inappropriate input mix, so its cost must be linked to the input cost-share equations by means of a relationship between A_i and u_{ni}, $n = 1, \ldots, N-1$.

The linkage must respect the fact that cost is raised by allocative errors in any input in either direction. The formidable problem is to estimate the technology parameters β and the efficiency error components (T_i, A_i, and u_{ni}) for each producer.

The problem is both conceptual and statistical. The conceptual challenge is to establish a satisfactory linkage between allocative inefficiency (the u_{ni}) and its cost (A_i). The statistical challenge is to estimate a model with so many error components, each of which requires a distribution. The problem remained unresolved until Kumbhakar (1997) obtained analytical results, which Kumbhakar and Tsionas (2005) extended to estimate the model using Bayesian techniques. This is encouraging, because (1.42) remains a workhorse in the nonfrontier literature and, more important, because its extension (1.43) is capable of estimating and decomposing economic efficiency.

There is an appealing alternative. The solution is to remove the influence of allocative inefficiency from the error terms and parameterize it inside the cost frontier and its input cost shares. We turn to this approach below.

1.5.4 Shadow price models

The econometric techniques described in sections 1.5.1–1.5.3 are enveloping techniques. Each treats technical efficiency in terms of distance to a production frontier, economic efficiency in terms of distance to an appropriate economic frontier, and allocative efficiency as a ratio of economic efficiency to technical efficiency. They are in rough concordance on the fundamental notions of frontiers and distance, in keeping with the theoretical developments in section 1.3. They differ mainly in the techniques they employ to construct frontiers and to measure distance. However they all convert a weak inequality to an equality by introducing a one-sided error component.

There is a literature that seeks to measure efficiency without explicit recourse to frontiers, and indeed, it contains many papers in which the word "frontier" does not appear. In this literature, little attempt is made to envelop data or to associate efficiency with distance to an enveloping surface. Unlike most econometric efficiency analysis, the focus is on allocative efficiency. Instead of attempting to model allocative inefficiency by means of error components, as in (1.43), allocative inefficiency is modeled parametrically by means of additional parameters to be estimated.

The literature seems to have originated with Hopper (1965), who found subsistence agriculture in India to attain a high degree of allocative efficiency, supporting the "poor but efficient" hypothesis. He reached this conclusion by using OLS to estimate Cobb-Douglas production functions (not frontiers), then to calculate the value of the marginal product of each input, and then to make two comparisons: the value of an input's marginal product for different outputs, and the values of an input's marginal product with its price. In each comparison equality implies allocative efficiency, and the sign and magnitude of an inequality indicate the direction and severity (and the cost,

which can be calculated since the production function parameters have been estimated) of the allocative inefficiency. Hopper's work was heavily criticized, and enormously influential.

In a nutshell, the shadow price models that have followed have simply parameterized Hopper's comparisons, with inequalities being replaced with parameters to be estimated. Thus, assuming $M = 1$ for simplicity and following Lau and Yotopoulos (1971) and Yotopoulos and Lau (1973), the inequality

$$y \leqq f(x;\beta) \tag{1.44}$$

is parameterized as

$$y = \phi f(x;\beta). \tag{1.45}$$

There is no notion of a production frontier here, since in moving from (1.44) to (1.45) the obvious requirement that $\max\{\phi\} \leqq 1$ is ignored. Indeed, so far this is just a Hoch (1955)–Mundlak (1961) management bias production function model, in which different intercepts are intended to capture the effects of variation in the (unobserved) management input. But it gets better.

If producers seek to maximize profit, then the inequalities

$$\partial\phi f(x;\beta)/\partial x_n \gtreqless (w_n/p), n = 1, \ldots, N \tag{1.46}$$

are parameterized as

$$\partial\phi f(x;\beta)/\partial x_n = \theta_n(w_n/p), \tag{1.47}$$

where $\theta_n \gtreqless 1$ indicate under- or overutilization of x_n relative to the profit-maximizing values. All that remains is to endow $f(x;\beta)$ with a functional form, and estimation of (β, ϕ, θ_n) provides a more sophisticated framework within which to implement Hopper's procedures. A host of hypotheses can be tested concerning the existence and nature of technical and allocative efficiency, without recourse to the notion of a frontier and error components.

The shadow price approach gained momentum following the popularity of the Averch-Johnson (1962) hypothesis. This hypothesis asserted that regulated utilities allowed to earn a "fair" rate of return on their invested capital would rationally overcapitalize, leading to higher than minimum cost and thus to customer rates that were higher than necessary.

The analysis proceeds roughly as above. A producer's cost

$$w^T x \geqq c(y, w;\beta) \tag{1.48}$$

is parameterized as

$$w^T x = (1/\phi)c(y, \theta w;\beta), \tag{1.49}$$

where θw is a vector of shadow prices. Now, $\phi \leqq 1$ reflects technical inefficiency and $\theta_n \gtreqless 1$ reflects allocative inefficiency, and there is an explicit notion of a

cost frontier. A producer's input demands

$$x_n \gtreqless x_n(y, w; \beta) \tag{1.50}$$

are parameterized as

$$x_n = (1/\phi)x_n(y, \theta w; \beta). \tag{1.51}$$

Although x_n may be allocatively inefficient for the input prices w that a producer actually pays, it is allocatively efficient for the shadow price vector θw.

The Averch-Johnson hypothesis asserts that rate-of-return regulation lowers the shadow price of capital beneath the cost of capital, leading to rational overcapitalization. The situation is depicted in figure 1.13. Given exogenous output y and input prices w_K and w_L, the cost-minimizing input combination occurs at x^E. The actual input combination occurs at x^A, which is technically efficient but allocatively inefficient, involving overcapitalization. Since the actual input combination must be allocatively efficient for some price ratio, the problem boils down to one of estimating the distortion factor θ along with the technology parameters β. In the two-input case illustrated in figure 1.13, there is one distortion parameter, while in the N input case there are $N - 1$ distortion parameters. The hypothesis of interest is that $\theta < 1$, the cost of which is given by the ratio $[c(y, \theta w; \beta)/c(y, w; \beta)] \geqq 1$, which is the reciprocal of the cost-efficiency measure (1.22) translated to this analytical framework.

Comparing (1.49) and (1.51) with (1.43) makes it clear that in the shadow price approach both sources of cost inefficiency have been moved from error components to the functions to be estimated. Although the error components approach to estimation and decomposition of economic efficiency has proved intractable so far, the shadow price approach has proved successful and has become very popular. It is also possible to combine the two approaches, by modeling technical efficiency as an error component and modeling allocative

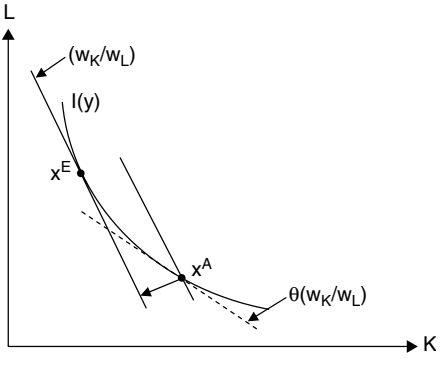


Figure 1.13. The Averch-Johnson Hypothesis

efficiency parametrically. Kumbhakar and Lovell (2000) discuss estimation strategies for the pure shadow price model and the combined model.

When modeling the behavior of producers who are constrained in their pursuit of a conventional objective, or who pursue an unconventional objective, analysts have two choices. The preferred choice is to model objective and constraint(s) correctly, derive the first-order conditions, and construct an estimating model based on the assumption that producers are efficient. This can be hard work, as Färe and Logan (1983) have demonstrated for the case of the profit-seeking rate-of-return–regulated producer. An easier alternative approach, illustrated above, is to model such producers as being unconstrained in their pursuit of a conventional objective, allow for failure to satisfy first-order conditions, and check to see if the direction of the estimated allocative inefficiency is consistent with what one would expect if in fact the producers were constrained or pursued some other objective. That is, use a model that is inappropriate but familiar, and look for allocative inefficiency by comparing shadow price ratios with actual price ratios.

In a related situation the analyst does not know the constraints or the objective of producers, perhaps because there are competing paradigms at hand. In this case, it is feasible to use the familiar model and use estimated shadow prices to provide an indirect test of the competing paradigms.

These are the two purposes that the shadow price approach most frequently serves. Thus, allocative inefficiency in the unconstrained pursuit of cost minimization or profit maximization suggests allocative efficiency in a more complicated environment, and departures of shadow price ratios from actual price ratios provide the basis for hypothesis tests. The model has been used frequently to test the Averch-Johnson hypothesis, and more generally as a framework for testing allocative efficiency hypotheses in a wide variety of contexts. Two other examples come to mind, primarily because they are current and have not yet been subjected to analysis using the shadow price approach. The impact of domestic content legislation could be explored within the shadow price framework. Another popular hypothesis that could be tested within this framework is that of discrimination, against minorities or immigrants or whatever group is of interest.

1.6 The Mathematical Programming Approach to Efficiency Measurement

The mathematical programming approach to the construction of frontiers and the measurement of efficiency relative to the constructed frontiers goes by the descriptive title of data envelopment analysis, with the interesting acronym DEA. It truly does envelop a data set; it makes no accommodation for noise and so does not "nearly" envelop a data set the way the deterministic kernel of a stochastic frontier does. Moreover, subject to certain assumptions about the structure of production technology, it envelops the data as tightly as possible.

Like the econometric approach, the programming approach can be categorized according to the type of data available (cross section or panel) and according to the types of variables available (quantities only, or quantities and prices). With quantities only, technical efficiency can be estimated, while with quantities and prices economic efficiency can be estimated and decomposed into its technical and allocative components. However, DEA was developed in a public-sector, not-for-profit environment, in which prices are suspect at best and missing at worst. Consequently, the vast majority of DEA studies use quantity data only and estimate technical efficiency only, despite the fact that the procedures are easily adapted to the estimation of economic efficiency in a setting in which prices are available and reliable.

In section 1.6.1, we analyze plain vanilla DEA to estimate technical efficiency. In section 1.6.2, we discuss one of many possible DEA models of economic efficiency. In section 1.6.3, we discuss the application of DEA to panel data, although the most popular such application occurs in the analysis of productivity change (which we discuss in section 1.8.3). In section 1.6.4, we discuss a technical issue, the imposition of weight restrictions, which has important economic implications. Finally, in section 1.6.5, we offer a brief introduction to the statistical foundations of DEA, a subject covered more fully in chapter 4.

1.6.1 Basic DEA

Producers use inputs $x \in R_+^N$ to produce outputs $y \in R_+^M$. The research objective is to estimate the performance of each producer relative to best observed practice in a sample of $i = 1, \ldots, I$ producers. To this end, weights are attached to each producer's inputs and outputs so as to solve the problem

$$\begin{aligned} &\text{Min}_{\upsilon,\mu}\, \upsilon^T x_o / \mu^T y_o \\ &\text{Subject to } \upsilon^T x_i / \mu^T y_i \geqq 1, i = 1, \ldots, o, \ldots, I \\ &\upsilon, \mu \geqq 0 \end{aligned} \tag{1.52}$$

Here (x_o, y_o) are the vectors of inputs and outputs of the producer under evaluation, and (x_i, y_i) are the vectors of inputs and outputs of the ith producer in the sample. The problem seeks a set of nonnegative weights, or multipliers, that minimize the weighted input-to-output ratio of the producer under evaluation, subject to the constraints that when these weights are assigned to every producer in the sample, their weighted input-to-output ratios are bounded below by one. Associate the multipliers (υ, μ) with shadow prices, and think of the objective in the problem as one of minimizing the ratio of shadow cost to shadow revenue.

The nonlinear program (1.52) can be converted to a dual pair of linear programs. The first DEA model is known as the CCR model, after Charnes, Cooper, and Rhodes (1978). The "multiplier" program appears in the right column of (1.53) below, where X is an $N \times I$ sample input matrix with columns

of producer input vectors x_i, and Y is an $M \times I$ sample output matrix with columns of producer output vectors y_i. Think of the multiplier program as one of minimizing shadow cost, subject to the constraint that shadow revenue is normalized to one, and subject to the constraints that when these multipliers are assigned to all producers in the sample, no producer earns positive shadow profit:

CCR Envelopment Program	*CCR Multiplier Program*
$\text{Max}_{\phi,\lambda}\,\phi$	$\text{Min}_{\upsilon,\mu}\,\upsilon^T x_o$
Subject to $X\lambda \leqq x_o$	Subject to $\mu^T y_o = 1$
$\phi y_o \leqq Y\lambda$	$\upsilon^T X - \mu^T Y \geqq 0$
$\lambda \geqq 0$	$\upsilon, \mu \geqq 0$

(1.53)

Because the multiplier program is a linear program, it has a dual, which is also a linear program. The dual "envelopment" program appears in the left column of (1.53), where ϕ is a scalar and λ is an $I \times 1$ intensity vector. In the envelopment program, the performance of a producer is evaluated in terms of its ability to expand its output vector subject to the constraints imposed by best practice observed in the sample. If radial expansion is possible for a producer, its optimal $\phi > 1$, while if radial expansion is not possible, its optimal $\phi = 1$. Noting the output orientation of the envelopment program, it follows that ϕ is the DEA estimator of $TE_o(x, y)$ defined in (1.12). Noting that ϕ is a radial efficiency measure, and recalling the divergence between Koopmans's definition of technical efficiency and the Debreu-Farrell measure of technical efficiency, it follows that optimal $\phi = 1$ is necessary, but not sufficient, for technical efficiency since $(\phi y_o, x_o)$ may contain slack in any of its M + N dimensions. At optimum, $\phi = 1$ characterizes technical efficiency in the sense of Debreu and Farrell, while $\{\phi = 1,\ X\lambda = x_o,\ \phi y_o = Y\lambda\}$ characterizes technical efficiency in the sense of Koopmans.

The output-oriented CCR model is partly illustrated in figure 1.14, for the $M = 2$ case. Producer A is technically inefficient, with optimal projection $\phi^A y^A$ occurring at a convex combination of efficient producers D and C on the output isoquant $I^{CCR}(x)$, so $\lambda^D > 0$, $\lambda^C > 0$, with all other elements of the intensity vector being zero. The efficient role models D and C are similar to, and a linear combination of them is better than, inefficient producer A being evaluated. The envelopment program provides this information. The multiplier program provides information on the trade-off between the two outputs at the optimal projection. The trade-off is given by the optimal shadow price ratio $-(\mu_1/\mu_2)$. The fact that this shadow price ratio might differ from the market price ratio, if one exists, plays a role in the DEA model of economic efficiency in section 1.6.2. The multiplier program also provides information on input trade-offs $-(\upsilon_n/\upsilon_k)$ and output–input trade-offs (μ_m/υ), although this information is not portrayed in figure 1.14.

Problem (1.53) is solved I times, once for each producer in the sample, to generate I optimal values of (ϕ, λ) and I optimal values of (υ, μ). It thus

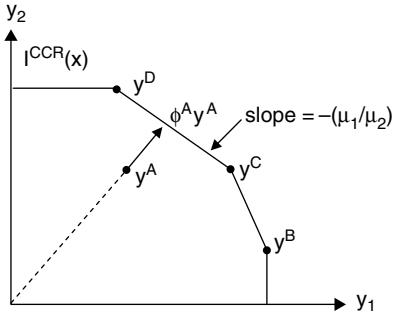


Figure 1.14. The Output-Oriented CCR Model

provides a wealth of information about the performance of each producer in the sample and about the structure of production technology.

The CCR production set corresponding to T in (1.1) is obtained from the envelopment problem in (1.53) as $T^{CCR} = \{(y, x) : y \leqq Y\lambda,\ X\lambda \leqq x,\ \lambda \geqq 0\}$ and imposes three restrictions on the technology. These restrictions are constant returns to scale, strong disposability of outputs and inputs, and convexity. Each of these restrictions can be relaxed.

Constant returns to scale is the restriction that is most commonly relaxed. Variable returns to scale is modeled by adding a free variable υ_o to the multiplier program, which is equivalent to adding a convexity constraint $\Sigma_i\lambda_i = 1$ to the envelopment program. The variable returns to scale model was introduced by Afriat (1972), but is better known as the BCC model after Banker, Charnes, and Cooper (1984). The BCC envelopment and multiplier programs become

BCC Envelopment Program	*BCC Multiplier Program*
$\text{Max}_{\phi,\lambda}\,\phi$	$\text{Min}_{\upsilon,\upsilon o,\mu}\,\upsilon^T x_o + \upsilon_o$
Subject to $X\lambda \leqq x_o$	Subject to $\mu^T y_o = 1$
$\phi y_o \leqq Y\lambda$	$\upsilon^T X + \upsilon_o - \mu^T Y \geqq 0$
$\lambda \geqq 0, \Sigma_i\lambda_i = 1$	$\upsilon, \mu \geqq 0, \upsilon_o$ free

(1.54)

The interpretation of the BCC envelopment and multiplier programs is essentially the same as for the CCR model, but the BCC production set shrinks, becoming $T^{BCC} = \{(y, x) : y \leqq Y\lambda,\ X\lambda \leqq x,\ \lambda \geqq 0,\ \Sigma_i\lambda_i = 1\}$. T^{BCC} exhibits variable returns to scale, because only convex combinations of efficient producers form the best-practice frontier. For this reason, it envelops the data more tightly than T^{CCR} does.

The difference between the two production sets is illustrated in figure 1.15. Because T^{BCC} envelops the data more tightly than T^{CCR} does, efficiency estimates are generally higher with a BCC specification, and rankings can differ in the two specifications. As in the CCR model, the BCC envelopment program provides efficiency estimates and identifies efficient role models. Also as in the CCR model, the BCC multiplier program estimates optimal shadow

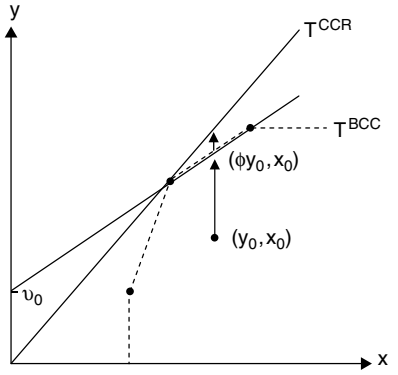


Figure 1.15. Returns to Scale in DEA

price ratios, but it also provides information on the nature of scale economies. The optimal projection to T^{BCC} occurs at $(\phi y_o, x_o)$. At this projection, the output–input trade-off is μ/υ. The vertical intercept of the supporting hyperplane $y = \upsilon_o + \upsilon x_o$ at $(\phi y_o, x_o)$ is positive. This indicates decreasing returns to scale at $(\phi y_o, x_o)$, which should be apparent from figure 1.15. More generally, $\upsilon_o \lessgtr 0$ signals that a producer is operating in a region of increasing, constant or decreasing returns to scale.

Notice the shape of T^{BCC} in figure 1.15. Requiring strictly positive input to produce nonzero output is a consequence of not allowing for the possibility of inactivity and of imposing convexity on T^{BCC}. This creates a somewhat strained notion of variable returns to scale, one that is well removed from the classical S-shaped production frontier that reflects Frisch's (1965) "ultra-passum" law. Petersen (1990) has attempted to introduce more flexibility into the DEA approach to measuring scale economies by dispensing with the assumption of convexity of T while maintaining the assumption of convexity of L(y) and P(x).

The CCR and BCC models differ in their treatment of scale economies, as reflected by the additional equality constraint $\Sigma_i \lambda_i = 1$ and free variable υ_o in the BCC model. Just as (μ, υ) are shadow prices of outputs and inputs, υ_o is the shadow value of the convexity constraint $\Sigma_i \lambda_i = 1$. It is possible to conduct a test of the null hypothesis that $\upsilon_o = 0$, or that the convexity constraint $\Sigma_i \lambda_i = 1$ is redundant. This is a test for constant returns to scale and is discussed along with other hypothesis tests in chapter 4. However, a qualification is in order concerning the interpretation of the multipliers. Most efficient producers are located at vertices, and it is possible that some inefficient producers are projected to vertices. At vertices, shadow prices of variables (υ, μ) in the CCR and BCC models, and of the convexity constraint (υ_o) in the BCC model, are not unique.

The CCR and BCC envelopment programs are output oriented, just as the econometric problem (1.32) is. It is a simple matter to obtain analogous

input-oriented envelopment programs, by converting the envelopment programs to minimization programs and converting the multiplier problems to maximization programs (details appear in chapter 3). The choice between the two orientations depends on the objective assigned to producers. If producers are required to meet market demands, and if they can freely adjust input usage, then an input orientation is appropriate.

The assumption of strong disposability is rarely relaxed, despite the obvious interest in relaxing the free disposability of surplus inputs or unwanted outputs. One popular exception occurs in environmental economics, in which producers use purchased inputs to produce marketed outputs and undesirable byproducts such as air or water pollution. In this case, the byproducts may or may not be *privately* freely disposable, depending on whether the regulator is watching, but they are surely *socially* weakly or expensively disposable. The value of relaxing the strong output disposability assumption lies in its potential to provide evidence on the marginal private cost of abatement. This evidence can be compared with estimates of the marginal social benefit of abatement to inform public policy.

Without going into details [which are provided by Färe et al. (1989, 1993) and a host of subsequent writers], the essence of weak disposability is captured in figure 1.16. Here, y_2 is a marketed output and y_1 is an undesirable byproduct. A conventional output set exhibiting strong disposability is bounded by the output isoquant $I^S(x)$ with solid line segments. The corresponding output set exhibiting weak disposability of the byproduct is bounded by the output isoquant $I^W(x)$ with dashed line segments. $L^W(x) \subseteq L^S(x)$, and that part of $L^S(x)$ not included in $L^W(x)$ provides an indication of the amount of marketed output foregone if the byproduct is not freely disposable. Disposal is free with technology $L^S(x)$, and abatement is costly with technology $L^W(x)$. For $y_1 < y_1^*$, the conventional strong disposal output set allows abatement of y_1 to be privately free, as indicated by the horizontal solid line segment along which $(\mu_1/\mu_2) = 0$. In contrast, the weak disposal output set makes abatement privately costly, as indicated by the positively sloped dashed line segments to the left of y_1^*. Moreover, increased abatement becomes increasingly costly, since the shadow price ratio $(\mu_1/\mu_2) > 0$ increases with additional abatement.

In figure 1.16, the marginal cost of abatement is reflected in the amount of y_2 (and hence revenue) that must be sacrificed to reduce the byproduct. With given inputs and technology, reducing air pollution requires a reduction in electricity generation. Allowing x or technology to vary would allow the cost of abatement to reflect the additional input or the new technology (and hence cost) required to abate with no loss in marketed output. With given electricity generation, reducing air pollution could be accomplished by installing scrubbers or by upgrading technology.

The assumption of convexity of output sets $P(x)$ and input sets $L(y)$ also is rarely relaxed, despite the belief of many, expressed by McFadden (1978, pp. 8–9), that its importance lies more in its analytical convenience than in its technological realism. In the previous context of scale economies, feasibility

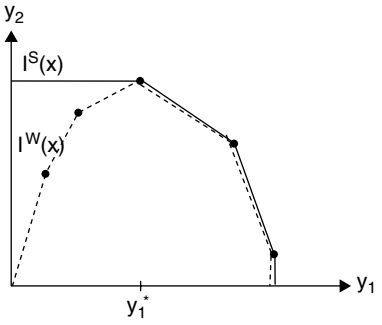


Figure 1.16. Weak Disposability of y_1

of an activity (y, x) does not necessarily imply feasibility of all scaled activities $(\lambda y, \lambda x)$, $\lambda > 0$, which motivates relaxing the assumption of constant returns to scale. In the present context, feasibility of two distinct activities (y^A, x^A) and (y^B, x^B) does not necessarily imply feasibility of all convex combinations of them, which motivates relaxing the assumption of convexity.

Deprins et al. (1984) were the first to relax convexity. They constructed a "free disposal hull" (FDH) of the data that relaxes convexity while maintaining strong disposability and allowing for variable returns to scale. An FDH output set is contrasted with a BCC output set in figure 1.17. The BCC output set is bounded by the output isoquant $I^{BCC}(x)$ as indicated by the solid line segments. The FDH output set dispenses with convexity but retains strong disposability, and is bounded by the output isoquant $I^{FDH}(x)$ as indicated by the dashed line segments. The contrast between FDH and DEA input sets and production sets is structurally identical. In each case, dispensing with convexity creates frontiers that have a staircase shape. This makes slacks a much more serious problem in FDH than in DEA, and it complicates the FDH multiplier program.

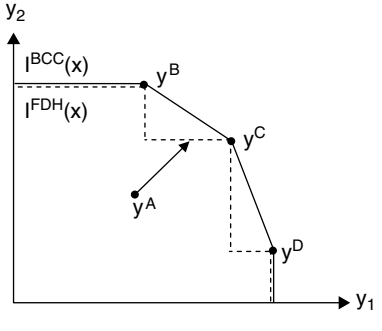


Figure 1.17. An FDH Output Set