1

Introduction

1.0 The Objectives of this Study

We have four goals we wish to accomplish in this study. The first two are interlinked, and involve the development of a model of efficient producer behavior, and the simultaneous development of a taxonomy of possible types of departure from efficiency, in a variety of environments. These two goals are accomplished by constructing a variety of production frontiers, and measuring distance to the frontiers. The third goal is the development of an analytical and computational technique for examining the first two. The technique we use, linear programming, is one of several available techniques, but is the one we prefer. The fourth goal is a demonstration of the wide applicability of the approach we take to modeling producer behavior. We meet this fourth goal in a number of ways, but primarily by adopting an attitude throughout the study that this is a study in applied production analysis. We focus on the empirical relevance of producer frontiers and distance to them. We apply the linear programming techniques to artificial data to illustrate the type of information they can generate. And we frequently suggest problems to which these ideas can be applied.

Conventional microeconomic theory is based on the assumption of optimizing behavior. Thus it is assumed that producers optimize from a technical or engineering perspective by not wasting resources. Loosely speaking, this means that producers operate somewhere on the boundary, rather than on the interior, of their production possibility sets. Producers are also assumed to optimize from an economic perspective by solving some allocation problem that involves prices. Thus cost min-

imizing producers are assumed to allocate resources efficiently so as to operate on rather than above their minimum cost frontiers, and similarly for producers seeking other economic objectives. However for a variety of reasons not all producers succeed in solving both types of optimization problem in all circumstances. For this reason it is important to have a way of analyzing the degree to which producers fail to optimize and the extent of the departures from technical and economic efficiency. Ideally this analysis is integrated with the analysis of the structure of efficient behavior, and developed simultaneously with it, so that we have a model of efficient producer behavior and of departure from efficient behavior. The first two goals of this study are achieved through the development of such an integrated analysis.

There exist more than one way to meet the first two goals of our study. The approach we adopt is to formulate a model of efficient production and economic behavior using linear programming techniques. The fact that they are "tightly bounding" or "closely enveloping" techniques makes them ideally suited to the type of economic problem under consideration. Their bounding property enables them to serve well the goal of representing frontiers, and their tightness property provides a yardstick against which to measure efficiency.

The linear programming methodology is used to achieve the third goal of this study. The final goal is reached in part by actually formulating and then solving linear programs for various production problems, and in part by suggesting a variety of other problems for which these techniques have been used or would be suitable.

Interest in the general topic of production frontiers and the measurement of efficiency relative to these frontiers has grown greatly in the last decade. We shall develop the history of thought on the subject as we go, but it is not too much of an oversimplification to say that the few technical studies that were published prior to the mid-1970s were characterized primarily by the dust they gathered. Since then, however, interest has blossomed. It is now possible to find published "efficiency" studies in virtually every country and about virtually every conceivable market or nonmarket production activity. Interest has spread to policy issues of great import.

1.1 The Motivation for this Study

The basic motivation for writing a book on production frontiers and efficiency is an empirical one. Theoretical production analysis has always focused on production activity as an optimization process. The basic tools have been the (maximum) production frontier, the (minimum) cost frontier, the (maximum) profit frontier, and so on. For practically as long, empirical production analysis has focused on central tendency, or "average" or "most likely" relationships constructed by intersecting data with a function rather than surrounding data with a frontier. This dichotomy between theoretical frontiers and empirical functions has long bothered us, and we have not been alone. There have emerged from time to time attempts to reconcile the two phenomena by developing quantitative techniques for surrounding rather than intersecting data. We view this study as an effort to build an empirically oriented approach to the envelopment of production data that integrates the construction of production frontiers with the measurement and interpretation of efficiency relative to the constructed frontiers.

In addition to the pursuit of a logically consistent approach to the empirical development of producer theory, we are motivated to model production frontiers and producer efficiency by other considerations. Among the more persuasive to us are the following.

(a) The structure of economic frontiers is of inherent interest. It is important to know whether, for example, technologically efficient production is characterized by economies of scale or scope, and whether cost efficient production is characterized by subadditivity.

(b) The structure of economic frontiers can be different from the structure of economic functions constructed from the same data. Best practice is not just better than average practice, it may also be structurally different. It is important to know whether, and if so in what ways, the structure of efficient production differs from the structure of average production. Best practice may be better than average practice precisely because it exploits available substitution possibilities or scale opportunities that average practice does not. Public policy based on the structure of best practice frontiers may be very different from policy based on the structure of average practice functions.

(c) The distance between production frontiers and observed producers is of obvious policy interest. It is important to know how

inefficient observed production is on average, and at what cost. It is also important to know what types of producers are most and least efficient, and what type of inefficiency is most prevalent. The distance between best practice and average practice was the focus of Salter's influential work; citing previous work of others, Salter (1966; 95-99) found the ratio of best practice to mean practice labor productivity to range from just over one to just under two in a variety of industries. Similar findings have been reported by Klotz, Madoo, and Hansen (1980), and Albriktsen and Førsund (1990). Although one may quarrel with the use of labor productivity as a performance measure, the point is that a wide gulf between best practice and average practice is not unusual.

Think of motivations (a)–(c) in the context of producers operating in conventional market environments pursuing conventional economic objectives. They apply with equal force to other producers, those operating in regulated or otherwise restricted market environments and those pursuing objectives other than the usual goal of profit maximization. Although the theory of such types of producers is well developed, it has not been integrated with any approach to measuring their efficiency. Thus a further motivation is to develop a framework for analyzing both phenomena.

(d) It is useful to have a model of producers pursuing conventional objectives under constraint that generates constrained frontiers and constrained efficiency measures. The reason is simple: it is important not to falsely attribute the effects of the constraint to a failure to solve the optimization problem. An example is provided by budget-constrained (as well as resource-constrained, the typical assumption) profit maximization. Another example is provided by profit maximization under regulatory constraint that restricts input usage (e.g., rate of return regulation or affirmative action programs) or output supply (e.g., local content protection requirements or command economy assortment and delivery requirements).

(e) It is useful to have a model of producers pursuing unconventional objectives that not only characterizes unconventional frontiers but at the same time measures performance relative to such frontiers. The reason is that it is important not to confuse successful pursuit of an unconventional objective with failure to achieve a

misspecified conventional objective. An example is the labor-managed Illyrian firm; another is the profit-constrained revenue-maximizing firm. The list of possibilities is long.

The five motives just cited refer to an interest in characterizing efficient production technology, and measuring efficiency relative to that technology, under a variety of producer objectives and environmental constraints. Another motive relates to the techniques we apply to the problem. We employ the tools of mathematical programming throughout the study, and we use these same tools in our empirical work. By using linear programming techniques to develop our analysis of frontiers and efficiency we are extending an approach initiated many years ago by von Neumann (1938 [1945]), and many others. The linear programming approach to production analysis has a rich, if relatively brief, intellectual history, and we are following in distinguished footsteps. Although production analysis is dominated by the parametric approach of Hicks (1946), Samuelson (1947), and others, linear programming remains a useful way of modeling production activity. By using linear programming techniques in empirical work, we find ourselves in the minority once again, at least among economists we know, although the minority is distinguished. This reveals further motivation for our study.

(f) It is useful to expose the profession, the bulk of a generation of which has grown up on a diet rich in calculus and richer still in least squares, to linear programming techniques. At a theoretical or modeling level, they are ideally suited to the task of constructing frontiers and measuring distances to calculated frontiers. At an empirical level they free investigators from having to impose unwanted structure on economic relationships of interest. At both levels they shift the emphasis of the investigation from a most likely relationship reflecting central tendency to a less likely relationship that focuses on extremal tendency.

Finally we have an interdisciplinary motive for conducting this study. Economics is not the only discipline interested in frontiers and efficiency. We are not alone. If interest on the part of economists in frontiers can be said to have been rekindled in the 1970s when the work of Debreu (1951), Koopmans (1951, 1957), and Farrell (1957) was rediscovered, then our rekindled interest is approximately the same age as the interest of the operations research–management science discipline in the same subject. A large and valuable body of work has emerged in the OR/MS field, and the two bodies of work have until recently developed

mostly independently. The OR/MS approach has its own orientation, and relies heavily on linear programming techniques. These techniques and the way they are applied deserve a wide audience, which we hope to stimulate.

(g) The OR/MS discipline has developed an approach (informatively dubbed "data envelopment analysis," or DEA) to the construction of production frontiers and efficiency measurement that deserves wider exposure among economists. It employs linear programming techniques, and it is similar in structure to a part of the framework we propose.

In light of the similarities between the economics and the OR/MS interests in efficiency measurement and its empirical applicability, it seems useful to quote at length from Lewin and Minton (1986), who seek to define a research agenda for determining organizational effectiveness. In their opinion (p. 529),

> "... it would be desirable to have a theory-based mathematics for calculating the relative effectiveness of an organization (over time or in comparison to other referent organizations) which is:

1 capable of analytically identifying relatively most effective organizations in comparison to relatively least effective organizations;

2 capable of deriving a single summary measure of relative effectiveness of organizations in terms of their utilization of resources and their environmental factors to produce desired outcomes;

3 able to handle noncommensurate, conflicting multiple outcome measures, multiple resource factors and multiple environmental factors outside the control of the organization being evaluated; and not be dependent on a set of *a priori* weights or prices for the resources utilized, environmental factors or outcome measures;

4 able to handle qualitative factors such as participant satisfaction, extent of information processing available, degree of competition, etc.;

5 able to provide insights as to factors which contribute to relative effectiveness ratings; and

6 able to maintain equity in the evaluation."

These desiderata closely parallel our own objectives.

1.2 The Strands of Thought Drawn Together in this Study

Within the general area of production economics, we draw on three bodies of literature in developing our analysis of frontiers and efficiency. The first is the body of work concerned with the measurement of efficiency in production. We stress that our interest is in measurement rather than causation, not because we think that a search for the causes of the pattern of efficiency is unimportant but (i) because uncovering the pattern comes first, and (ii) because our comparative advantage lies with measurement rather than hypothesizing about causal factors. This means that we shall have little to say about property rights, and principals and agents, and incentive mechanisms, and competition versus monopoly, and private versus public provision, and so on. The second body of work we draw on in our analysis is the approach to production analysis developed by Shephard (1953, 1970, 1974). We find Shephard's approach valuable for a number of reasons to be detailed below, but one outstanding virtue of his approach is that it is based on distance functions. This is particularly useful because distance functions are intimately related to (radial) efficiency measures, and also because the use of distance functions makes it clear that multiple outputs are no harder to deal with than multiple inputs. The third body of work we draw on in our analysis is the linear programming activity analysis approach to the theory of production. Our model of physical (production) and value (cost or profit or other) frontiers is general, however, and can be analyzed using other techniques. There is nothing in the underlying theory that requires a linear programming formulation. We use it simply because it provides an elegant formulation of production economics and, at the same time, it provides a computationally feasible method of calculation, both of the appropriate frontier and of the distance to it. We now briefly consider these three strands of thought in turn.

1.2.1 The Measurement of Efficiency in Production Economics

The matter of productive efficiency has been of interest since Adam Smith's pin factory and before. However a rigorous analytical approach to the measurement of efficiency in production can be said to have originated with the work of Koopmans and Debreu. Koopmans (1951; 60) provided a definition of what we refer to as technical efficiency: an input–output vector is technically efficient if, and only if, increasing any output or decreasing any input is possible only by decreasing some other out-

put or increasing some other input. Lest readers worry about whether Koopmans' definition implies the existence of (or our ability to determine) an absolute frontier, Farrell (1957; 255) and much later Charnes and Cooper (1985; 72) remind us of the empirical necessity of treating Koopmans' definition of technical efficiency as a *relative* notion, relative to best observed practice in the reference set or comparison group.

This provides a way of differentiating efficient from inefficient production states, but it offers no guidance concerning either the degree of inefficiency of an inefficient vector or the identification of an efficient vector or combination of efficient vectors with which to compare an inefficient vector. This issue was addressed by Debreu (1951), who offered the first measure of productive efficiency with his "coefficient of resource utilization." Debreu's measure is a radial measure of technical efficiency. Radial measures are nice because they focus on the maximum feasible *equiproportionate* reduction in all variable inputs, or the maximum feasible *equiproportionate* expansion of all outputs, in contrast to the more popular but less desirable maximization of output per unit of labor. Radial measures are also nice because they are independent of unit of measurement. However radial measures have a drawback: achievement of the maximum feasible input contraction or output expansion suggests technical efficiency, even though there may remain slack in outputs or inputs. That is, an input–output vector labeled efficient on the basis of Debreu's radial measure may be technically inefficient on the basis of Koopmans' definition because it may lie on the boundary of the production possibilities set but not on the efficient subset of the boundary. The trick is to derive an operationally useful efficiency measure that calls a vector efficient if and only if it satisfies the Koopmans definition. Various ways of dealing with this problem are introduced in Chapter 3 and used throughout the book.

Farrell (1957) extended the work initiated by Koopmans and Debreu by noting that production efficiency has a second component reflecting the ability of producers to select the "right" technically efficient input–output vector in light of prevailing input and output prices. This led Farrell to define overall productive efficiency as the product of technical and allocative, or price, efficiency. Implicit in the notion of allocative efficiency is a specific behavioral assumption about the goal of the producer; Farrell considered cost minimization in competitive input markets, although other behavioral assumptions can be considered. We do so throughout the book.

Although the natural focus of most economists is on markets and their prices and thus on allocative rather than technical efficiency, and although Farrell introduced the notion of allocative efficiency and its measurement, he expressed a concern about our ability to measure prices accurately enough to make good use of allocative efficiency measurement, and hence of overall economic efficiency measurement. This concern expressed by Farrell (1957; 261) has greatly influenced the OR/MS work on efficiency measurement; see Charnes and Cooper (1985; 94), who cite Farrell's concern as one of several motivations for the typical OR/MS emphasis on the measurement of technical efficiency. Notwithstanding the concerns of Farrell and many OR/MS practitioners, however, we show throughout the book that the linear programming techniques are quite capable of solving all sorts of price-dependent efficiency problems. If there is a problem, it lies with the availability and reliability of price data, not with the analytical and empirical technique we use to measure economic efficiency.

It should also be noted that the decomposition process initiated by Farrell has been taken further. Technical efficiency has been decomposed into the product of measures of scale efficiency, input congestion and "pure" technical efficiency by Färe, Grosskopf, and Lovell (1983). Efficiency measurement can be oriented toward the output side of the producer's operations, in which case revenue efficiency can be measured and decomposed into the product of (output price) allocative efficiency and technical efficiency, and the latter can be further decomposed into the product of scale efficiency, output congestion, and "pure" technical efficiency. Efficiency measurement can be oriented toward inputs and outputs together, in which case profit efficiency can be measured and decomposed, although the decomposition is not so straightforward as in the cases of cost efficiency and revenue efficiency. Several other types of efficiency, corresponding to different objectives or to the presence of different constraints, have been proposed, and some are considered in the chapters to follow. The practical advantage of being able to decompose efficiency, however defined and measured, lies in the resulting ability to quantify the magnitudes, and hence the relative importance, of the components. Knowing the magnitude and cost of inefficiency is useful, but so too is knowing where it is most problematic, e.g., internal to the production unit or in certain input markets or in certain output markets.

1.2.2 Shephard's Direct and Indirect Dual Approaches to Analyzing Production Technology and Producer Behavior

The second strand of thought we draw upon in this study relates to Shephard's (1970, 1974) models of technology and his (1953, 1970, 1974) distance functions. The models of technology include direct, indirect, and price space formulations, which we use in Chapters 3-4, 5-6 and 7, respectively. Although Shephard introduced each of these models of technology, he provided piecewise linear formulations for only the direct input and output correspondences (1970; 283-92, 1974; 5-13).

In contrast to the traditional scalar-valued production function, direct input and output correspondences readily admit multiple outputs and multiple inputs. They are thus well suited to characterize all kinds of technologies without having to resort to possibly unwarranted output aggregation prior to analysis.

Our indirect models of production, while based on Shephard's work, differ from his formulation in one respect. Shephard defined the indirect output correspondence in terms of the cost function (1974; 16), and he defined the indirect input correspondence in terms of the revenue function (1974; 24). Our formulations, which are equivalent to those of Shephard, are based on Shephard and Färe (1980), and they are defined on the direct correspondences with the addition of the appropriate value constraint. This formulation serves to clarify the notion of an indirect production correspondence, and it is particularly useful in the specification of linear programming models of indirect efficiency measurement.

Our two price space formulations of technology are based on Shephard's "cost structure correspondence" (1970; 232) and his "revenue correspondence" (1970; 234), and they are mappings from output space into subsets of cost-deflated input prices and from input space into subsets of revenue-deflated output prices, respectively. These two price space formulations lead naturally to the construction of two families of dual price efficiency measures.

The original distance function introduced by Shephard (1953; 5) is in our terminology a direct input distance function defined on the direct input correspondence. This distance function treats (multiple) outputs as given and contracts input vectors as much as possible consistent with technological feasibility of the contracted input vector. Among its several useful properties, the most important for our purposes is the fact that the reciprocal of the direct input distance function has been proposed by Debreu (1951) as a coefficient of resource utilization, and by