# 2

# The Econometric Approach to Efficiency Analysis

William H. Greene

## 2.1 Introduction

Chapter 1 describes two broad paradigms for measuring economic efficiency, one based on an essentially nonparametric, programming approach to analysis of observed outcomes, and one based on an econometric approach to estimation of theory-based models of production, cost, or profit. This chapter presents an overview of techniques for econometric analysis of technical (production) and economic (cost) efficiency. The stochastic frontier model of Aigner, Lovell, and Schmidt (1977) is now the standard econometric platform for this type of analysis. I survey the underlying models and econometric techniques that have been used in studying technical inefficiency in the stochastic frontier framework and present some of the recent developments in econometric methodology. Applications that illustrate some of the computations are presented in the final section.

### 2.1.1 Modeling production

The empirical estimation of production and cost functions is a standard exercise in econometrics. The *frontier production function* or *production frontier* is an extension of the familiar regression model based on the theoretical premise that a *production function*, or its dual, the *cost function*, or the convex conjugate of the two, the *profit function*, represents an ideal, the *maximum output* attainable given a set of inputs, the *minimum cost* of producing that output given the prices of the inputs, or the *maximum profit* attainable given the inputs,



FRIED: "CHAP02" — 2007/8/24 — 19:02 — PAGE 92 — #1

outputs, and prices of the inputs. The estimation of frontier functions is the econometric exercise of making the empirical implementation consistent with the underlying theoretical proposition that no observed agent can exceed the ideal. In practice, the frontier function *model* is (essentially) a regression model that is fit with the recognition of the theoretical constraint that all observations lie within the theoretical extreme. Measurement of (in)efficiency is, then, the empirical estimation of the extent to which observed agents (fail to) achieve the theoretical ideal. My interest in this chapter is in this latter function. The estimated model of production, cost, or profit is the means to the objective of measuring inefficiency. As intuition might suggest at this point, the exercise here is a formal analysis of the "residuals" from the production or cost model. The theory of optimization, production, and/or cost provides a description of the ultimate source of deviations from this theoretical ideal.

### 2.1.2 History of thought

The literature on frontier production and cost functions and the calculation of efficiency measures begins with Debreu (1951) and Farrell (1957) [though there are intellectual antecedents, e.g., Hicks's (1935) suggestion that monopolists would enjoy their position through the attainment of a quiet life rather than through the pursuit of economic profits, a conjecture formalized somewhat more by Leibenstein (1966, 1975)]. Farrell suggested that one could usefully analyze technical efficiency in terms of realized deviations from an idealized frontier isoquant. This approach falls naturally into an econometric approach in which the *inefficiency* is identified with disturbances in a regression model.

The empirical estimation of production functions had begun long before Farrell's work, arguably with Cobb and Douglas (1928). However, until the 1950s, production functions were largely used as devices for studying the functional distribution of income between capital and labor at the macroeconomic level. The celebrated contribution of Arrow et al. (1961) marks a milestone in this literature. The origins of empirical analysis of microeconomic production structures can be more reasonably identified with the work of Dean (1951, a leather belt shop), Johnston (1959, electricity generation), and, in his seminal work on electric power generation, Nerlove (1963). It is noteworthy that all three of these focus on costs rather than production, though Nerlove, following Samuelson (1938) and Shephard (1953), highlighted the dual relationship between cost and production.[1] Empirical attention to production functions at a disaggregated level is a literature that began to emerge in earnest in the 1960s (see, e.g., Hildebrand and Liu, 1965; Zellner and Revankar, 1969).

### 2.1.3 Empirical antecedents

The empirical literature on production and cost developed largely independently of the discourse on frontier modeling. Least squares or some variant

was generally used to pass a function through the middle of a cloud of points, and residuals of both signs were, as in other areas of study, not singled out for special treatment. The focal points of the studies in this literature were the estimated parameters of the production structure, not the individual deviations from the estimated function. An argument was made that these "averaging" estimators were estimating the average, rather than the "best-practice" technology. Farrell's arguments provided an intellectual basis for redirecting attention from the production function specifically to the deviations from that function, and respecifying the model and the techniques accordingly. A series of papers including Aigner and Chu (1968) and Timmer (1971) proposed specific econometric models that were consistent with the frontier notions of Debreu (1951) and Farrell (1957). The contemporary line of research on econometric models begins with the nearly simultaneous appearance of the canonical papers of Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977), who proposed the stochastic frontier models that applied researchers now use to combine the underlying theoretical propositions with a practical econometric framework. The current literature on production frontiers and efficiency estimation combines these two lines of research.

### 2.1.4 Organization of the survey

This survey presents an overview of this literature and proceeds as follows:

Section 2.2 presents the microeconomic theoretical underpinnings of the empirical models. As in the other parts of our presentation, this section gives only a cursory survey because of the very large literature on which it is based. The interested reader can find considerable additional detail in chapter 1 of this book and in a gateway to the larger literature, chapter 2 of Kumbhakar and Lovell (2000).

Section 2.3 constructs the basic econometric framework for the econometric analysis of efficiency. This section also presents some intermediate results on "deterministic" (orthodox) frontier models that adhere strictly to the microeconomic theory. This part is brief. It is of some historical interest and contains some useful perspective for the more recent work. However, with little exception, current research on the deterministic approach to efficiency analysis takes place in the environment of "data envelopment analysis" (DEA), which is the subject of chapter 3 of this book.[2] This section provides a bridge between the formulation of orthodox frontier models and the modern stochastic frontier models.

Section 2.4 introduces the stochastic production frontier model and presents results on formulation and estimation of this model. Section 2.5 extends the stochastic frontier model to the analysis of cost and profits and describes the important extension of the frontier concept to multiple-output technologies.

Section 2.6 turns to a major econometric issue, that of accommodating heterogeneity in the production model. The assumptions made in sections 2.4

and 2.5 regarding the stochastic nature of technical inefficiency are narrow and arguably unrealistic. Inefficiency is viewed as simply a random shock distributed homogeneously across firms. These assumptions are relaxed at the end of section 2.5 and in section 2.6. Here, I examine proposed models that allow the mean and variance of inefficiency to vary across firms, thus producing a richer, albeit considerably more complex, formulation. This part of the econometric model extends the theory to the practical consideration of observed and unobserved influences that are absent from the pure theory but are a crucial aspect of the real-world application.

The econometric analysis continues in section 2.7 with the development of models for panel data. Once again, this is a modeling issue that provides a means to stretch the theory to producer behavior as it evolves through time. The analysis pursued here goes beyond the econometric issue of how to exploit the useful features of longitudinal data. The literature on panel data estimation of frontier models also addresses the fundamental question of how and whether inefficiency varies over time, and how econometric models can be made to accommodate the theoretical propositions.

The formal measurement of inefficiency is considered in sections 2.8 and 2.9. The use of the frontier function model for estimating firm-level inefficiency that was suggested in sections 2.3 and 2.4 is formalized in the stochastic frontier model in section 2.8. Section 2.9 considers the separate issue of allocative inefficiency. In this area of study, the distinction between errors in optimization and the consequences of those errors for the goals or objectives of optimization is made explicit. Thus, for example, the effect of optimization errors in demand systems is viewed apart from the ultimate impact on the costs of production.

Section 2.10 describes contemporary software for frontier estimation and illustrates some of the computations with "live" data sets. Some conclusions are drawn in section 2.11.

### 2.1.5 Preface

The literature on stochastic frontier estimation was already large at the time of the 1993 edition of this survey and it has grown vastly in the decade plus since then. It is simply not possible to touch upon all aspects of all the research that has been done and is ongoing. [Even the book-length treatise Kumbhakar and Lovell (2000) leaves the reader to their own devices to explore the received empirical studies.] In this survey, I introduce a number of topics and present some of the most familiar econometric estimators and issues. Since the earlier rendition of this survey, two topics have received great attention in the literature are given correspondingly greater coverage here: the statistical analysis of the inefficiency estimators (the Jondrow et al., 1982, estimator and counterparts) and panel data estimation. A few topics are treated relatively superficially, not for lack of interest but because, for better or for worse, they have not yet had great influence on how empirical work is done in this

area. These include Bayesian estimation and semi- and nonparametric estimation. Yet another topic falls somewhere between the mainstream and these. In the analysis of inefficiency, we recognize that, in terms of costs, inefficiency can arise from two sources: *technical inefficiency*, which arises when, given the chosen inputs, output falls short of the ideal; and *allocative inefficiency*, which arises from suboptimal input choices given prices and output. Technical inefficiency (the difference between output and maximal output) is, in some sense, "pure" in that we can single out the source. Cost inefficiency, in contrast, is a blend of the two sources, technical and allocative inefficiency. Decomposition of cost inefficiency into its two components in a theoretically appropriate manner (the so-called "Greene problem") has posed a vexing challenge in this literature (see Greene, 1993, 1997, 2003c). The estimation of "allocative" inefficiency has received some attention in the research of the past two decades, with some interesting and creative results. However, the estimation of allocative inefficiency and the decomposition have received much less attention than the more straightforward analysis of technical inefficiency on the production side and *economic* inefficiency (i.e., the blend) on the cost side. This is due partly to the very heavy data and analytical/technical requirements of the received approaches but mainly, when all is said and done, to the persistent absence of a practical theoretically consistent solution to the original problem. Formal analysis of allocative inefficiency requires estimation of both a cost or production function and a complete demand system. I introduce this topic below but spend less space on it than on the estimation of technical and "economic" (cost) efficiency.

Note, finally, that the range of applications of the techniques described here is also huge. Frontier analysis has been used to study inefficiency in hospital costs, electric power, commercial fishing, farming, manufacturing of many sorts, public provision of transportation and sewer services, education, labor markets, and a huge array of other settings.[3] Both space and time precludes any attempt to survey this side of the literature here. I hope the community of researchers whose work is not explicitly cited here can forgive the omission of their work, again, not for lack of interest, but for lack of space. My intent in this chapter is to survey methods; reluctantly, I leave it to the reader to explore the vast range of applications. The extensive table in chapter 1 (which unfortunately is limited to twenty-first century contributions) should be very helpful.

There have been numerous general survey-style studies of the frontiers literature, including, of course, the earlier editions of this work: Førsund et al. (1980) and Greene (1993, 1997). Notable among these surveys are Bauer (1990), Battese (1992), Schmidt (1985), Cornwell and Schmidt (1996), Kalirajan and Shand (1999), and Murillo-Zamorano (2004). There are book-length treatments, as well, including Kumbhakar and Lovell (2000) and Coelli, Rao, and Battese (1998).[4] Given all of these, I felt it necessary to give some thought to the end purpose of the present exercise. First, obviously, it is an opportunity to give some exposure to the last five or six years of innovative

research in the area. Primarily, however, I see my purpose here as providing the interested entrant to the area a bridge from the basic principles of econometrics and microeconomics to the specialized literature in econometric estimation of inefficiency. As such, it is not necessary to provide a complete compendium of either the theory or the received empirical literature. (That is fortunate, given its volume.) Thus, my intent is to introduce the econometric practice of efficiency estimation to the newcomer to the field.

## 2.2 Production and Production Functions

Let's begin by defining a producer as an economic agent that takes a set of *inputs* and transforms them either in form or in location into a set of *outputs*. We keep the definition nonspecific because we desire to encompass service organizations such as travel agents or law or medical offices. Service businesses often rearrange or redistribute information or claims to resources, which is to say, move resources rather than transform them. The production of *public services* provides one of the more interesting and important applications of the techniques discussed in this study (see, e.g., Pestieau and Tulkens, 1993).

### 2.2.1 Production

It is useful to think in terms of a producer as a simple machine. An electric motor provides a good example. The inputs are easily definable, consisting of a lump of capital, the motor, itself, and electricity that flows into the motor as a precisely defined and measurable quantity of the energy input. The motor produces two likewise precisely measurable (at least in principle) outputs, "work," consisting of the rotation of a shaft, and heat due to friction, which might be viewed in economic terms as waste, or a negative or undesirable output (see, e.g., Atkinson and Dorfman, 2005). Thus, in this setting, we consider production to be the process of transforming the two inputs into the economically useful output, work. The question of "usefulness" is crucial to the analysis. Whether the byproducts of production are "useless" is less than obvious. Consider the production of electricity by steam generation. The excess steam from the process might or might not be economically useful (it is in some cities, e.g., New York and Moscow), depending, in the final analysis, on relative prices. Conceding the importance of the distinction, we depart at this point from the issue and focus our attention on the production of economic "goods" that have been agreed upon a priori to be "useful" in some sense.

The economic concept of production generalizes from a simple, well-defined engineering relationship to higher levels of aggregation such as farms, plants, firms, industries, or, for some purposes, whole economies that engage in the process of transforming labor and capital into gross domestic product by some ill-defined production process. Although this interpretation stretches

the concept perhaps to its logical limit, it is worth noting that the first empirical analyses of production functions, by Cobb and Douglas (1928), were precisely studies of the functional distribution of income between capital and labor in the context of an aggregate (macroeconomic) production function.

### 2.2.2 Modeling production

The *production function* aspect of this area of study is a well-documented part of the model. The "function" itself is, as of the time of the observation, a relationship between inputs and outputs. It is most useful to think of it simply as a body of knowledge. The various technical aspects of production, such as factor substitution, economies of scale, or input demand elasticities, while interesting in their own right, are only of tangential interest in the present context. To the extent that a particular specification, Cobb-Douglas versus translog, for example, imposes restrictions on these features, which then distort our efficiency measures, we are interested in functional form. But, this is not the primary focus.

The Cobb-Douglas and translog models overwhelmingly dominate the applications literature in stochastic frontier and econometric inefficiency estimation. (In contrast, the received literature in DEA—by construction—is dominated by linear specifications.) The issue of functional form for the production or cost function (or distance, profit, etc.) is generally tangential to the analysis and not given much attention. There have been a number of studies specifically focused on the functional form of the model. In an early entry to this literature, Caves, Christensen (one of the creators of the translog model), and Trethaway (1980) employed a Box-Cox functional form in the translog model to accommodate zero values for some of the outputs.[5] The same consideration motivated Martinez-Budria, Jara-Diaz, and Ramos-Real (2003) in their choice of a quadratic cost function to study the Spanish electricity industry. Another proposal to generalize the functional form of the frontier model is the Fourier flexible function used by Huang and Wang (2004) and Tsionas (2004).

In a production (or cost) model, the choice of functional form brings a series of implications with respect to the shape of the implied isoquants and the values of elasticities of factor demand and factor substitution. In particular, the Cobb-Douglas production function has universally smooth and convex isoquants. The implied cost function is likewise well behaved. The price to be paid for this good behavior is the strong assumption that demand elasticities and factor shares are constant for given input prices (for all outputs), and that Allen elasticities of factor substitution are all –1. Cost functions are often used in efficiency analysis because they allow the analyst to specify a model with multiple inputs. This is not straightforward on the production side, though distance functions (see section 2.5.4) also provide an avenue. The Cobb-Douglas multiple-output cost function has the unfortunate implication that in output space, the output possibility frontiers are all convex instead

of concave—thus implying output specialization. These considerations have generally motivated the choice of flexible (second-order) functional forms, and in this setting, the translog production model for one output and $K$ inputs,

$$\ln y = \alpha + \sum_{k=1}^{K} \beta_k \ln x_k + \frac{1}{2} \sum_{k=1}^{K} \sum_{m=1}^{K} \gamma_{km} \ln x_k \ln x_m,$$

or the translog multiple-output cost function for $K$ inputs and $L$ outputs,

$$\begin{aligned} \ln C = \alpha &+ \sum_{k=1}^{K} \beta_k \ln w_k + \frac{1}{2} \sum_{k=1}^{K} \sum_{m=1}^{K} \gamma_{km} \ln w_k \ln w_m \\ &+ \sum_{s=1}^{L} \delta_s \ln y_s + \frac{1}{2} \sum_{s=1}^{L} \sum_{t=1}^{L} \phi_{st} \ln y_s \ln y_t \\ &+ \sum_{k=1}^{K} \sum_{s=1}^{L} \theta_{ks} \ln w_k \ln y_s, \end{aligned}$$

is most commonly used (see, e.g., Kumbhakar, 1989). These models do relax the restrictions on demand elasticities and elasticities of substitution. However, the generality of the functional form produces a side effect: They are not monotonic or globally convex, as is the Cobb-Douglas model. Imposing the appropriate curvature on a translog model is a generally challenging problem. [See Salvanes and Tjotta (1998) for methodological commentary.] Researchers typically (one would hope) "check" the regularity conditions after estimation. Kleit and Terrell (2001) in an analysis of the U.S. electricity industry used a Bayesian estimator that directly imposes the necessary curvature requirements on a two-output translog cost function. The necessary conditions, which are data dependent—they will be a function of the specific observations—are (1) monotonicity: $s_k = \partial \ln C / \partial \ln w_k = \beta_k + \sum_m \gamma_{km} \ln w_m \geq 0$, $k = 1, \ldots, K$ (nonnegative factor shares); and (2) concavity: $\mathbf{\Gamma} - \mathbf{S} + \mathbf{s}\mathbf{s}^{\mathrm{T}}$ negative semidefinite, where $\mathbf{\Gamma} = [\gamma_{km}]$, $\mathbf{S} = \mathrm{diag}[s_k]$, and $\mathbf{s} = [s_1, s_2, \ldots, s_k]^{\mathrm{T}}$. Monotonicity in the outputs requires $\partial \ln C / \partial \ln y_s = \delta_s + \sum_r \phi_{sr} \ln y_r > 0$. As one might imagine, imposing data- and parameter-dependent constraints such as these during estimation poses a considerable challenge. In this study, Kleit and Terrell selectively cull the observations during estimation, retaining those that satisfy the restrictions. Another recent study, O'Donnell and Coelli (2005) also suggest a Bayesian estimator, theirs for a translog distance function in which they impose the necessary curvature restrictions a priori, parametrically. Finally, Griffiths, O'Donnell, Tan, and Cruz (2000) impose the regularity conditions on a system of cost and cost-share equations.

The preceding is an issue that receives relatively little attention in the stochastic frontier applications, though it is somewhat more frequently examined in the more conventional applications in production and cost modeling

(e.g., Christensen and Greene, 1976).[6] I acknowledge this aspect of modeling production and cost at this point, but consistent with others, and in the interest of brevity, I do not return to it in what follows.

### 2.2.3 Defining efficiency

The analysis of economic inefficiency stands at the heart of this entire exercise. If one takes classical microeconomics at face value, this is a fruitless exercise, at least regarding "competitive" markets. Functioning markets and the survivor principle simply do not tolerate inefficiency. But, this clearly conflicts with even the most casual empiricism. Also note that analysis of regulated industries and government enterprises (including buses, trains, railroads, nursing homes, waste hauling services, sewage carriage, etc.) has been among the most frequent recent applications of frontier modeling. Because the orthodoxy of classical microeconomics plays a much lesser role here, the conflict between theory and practice is less compelling. I eschew a philosophical discussion of the *concept* of inefficiency, technical, allocative, or otherwise. (For a very readable, if somewhat glib discussion, the reader may begin with Førsund, Lovell, and Schmidt, 1980).[7] Alvarez, Arias, and Greene (2005) also pursue this issue from an econometric perspective. In what follows, producers are characterized as efficient if they have produced as much as possible with the inputs they have actually employed or if they have produced that output at minimum cost. I formalize the notions as we proceed.

By technical efficiency, I mean here to characterize the relationship between observed production and some ideal, or potential production. In the case of a single output, we can think in terms of *total factor productivity*, the ratio of actual output to the optimal value as specified by a "production function." Two crucial issues, which receive only occasional mention in this chapter, are the functional form of the production function and the appropriate list of inputs. In both cases, specification errors can bring systematic errors in the *measurement* of efficiency.

AQ: In Paragraph "We define ..." line $\mathbf{x} \in \mathrm{R}_k^+$, into the first two equations in this paragraph, please confirm that the boxes correct characters. You noted at copyediting review that nothing was missing.

We define production as a process of transformation of a set of inputs, denoted $\mathbf{x} \in \mathrm{R}_K^+$, into a set of outputs, $\mathbf{y} \in \mathrm{R}_M^+$. Henceforth, the notation $\mathbf{z}$, in boldface, denotes a column vector of the variables in question, whereas the same symbol $z$, in italics and not boldface, denotes a scalar, or a single input or output. The process of transformation (rearrangement, etc.) takes place in the context of a body of knowledge called the *production function*. We denote this process of transformation by the equation $T(\mathbf{y}, \mathbf{x}) = 0$. (The use of 0 as the normalization seems natural if we take a broad view of production against a backdrop of the laws of conservation—if $\mathbf{y}$ is defined broadly enough: Neither energy nor matter can be created nor destroyed by the transformation.)

I should emphasize that the production function is not static; technical change is ongoing. It is interesting to reflect that, given the broadest definition, the force of technical change would be only to change the mix of outputs

obtained from a given set of inputs, not the quantities in total. The electric motor provides a good example. A "more efficient" motor produces more work and less heat (at least by the yardstick that most people would use), but, in fact, the total amount of energy that flows from it will be the same before and after our hypothetical technical advance. The notion of greater efficiency in this setting is couched not in terms of total output, which must be constant, but in terms of a subset of the outputs that are judged as useful against the remaining outputs that arguably are less so.

The state of knowledge of a production process is characterized by an *input requirements set*

$$L(\mathbf{y}) = \{\mathbf{x} : (\mathbf{y}, \mathbf{x}) \text{ is producible}\}.$$

That is to say, for the vector of outputs $\mathbf{y}$, any member of the input requirements set is *sufficient* to produce the output vector. Note that this makes no mention of efficiency, nor does it define the production function per se, except indirectly insofar as it also defines the set of inputs that is *insufficient* to produce $\mathbf{y}$ [the complement of $L(\mathbf{y})$ in $\mathrm{R}_K^+$] and, therefore, defines the limits of the producer's abilities. The production function is defined by the *isoquant*

$$I(\mathbf{y}) = \{\mathbf{x} : \mathbf{x} \in L(\mathbf{y}) \text{ and } \lambda\mathbf{x} \notin L(\mathbf{y}) \text{ if } 0 \leq \lambda < 1\}.$$

The isoquant thus defines the boundary of the input requirement set. The isoquant is defined in terms of contraction of an input point. A more general definition is the *efficient subset*

$$\begin{aligned}\mathrm{ES}(\mathbf{y}) = \{\mathbf{x} : \mathbf{x} \in L(\mathbf{y}) \text{ and } \mathbf{x}' \notin L(\mathbf{y}) \text{ for } \mathbf{x}' \text{ when } \mathbf{x}'_k \leq \mathbf{x}_k \forall k \\ \text{and } \mathbf{x}'_k < \mathbf{x}_j \text{ for some } j\}.\end{aligned}$$

The distinction between these two similar definitions is shown in figure 2.1. Note that $\mathbf{x}^A = (x_1^A, x_2^A)'$ is on the isoquant but is not in the efficient subset, since there is slack in $x_2^A$. But $\mathbf{x}^B$ is in both $I(\mathbf{y})$ and $\mathrm{ES}(\mathbf{y})$. When the input

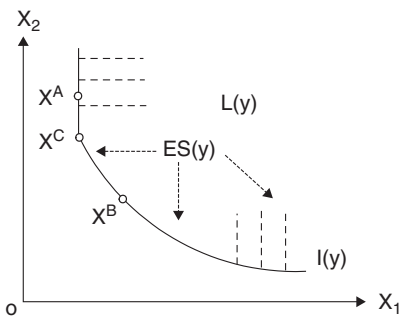


**Figure 2.1.** Input Requirements

AQ: Confirm that box is a correct character in line "[the complement ...] and".

requirements set is strictly convex, as is commonly assumed in econometric applications, the distinction disappears, but the distinction between these two sets is crucially important in DEA (discussed in chapter 3).

Shephard's (1953) *input distance function* is

$$D_I(\mathbf{y}, \mathbf{x}) = \max \left\{ \lambda : \left[ \frac{1}{\lambda} \right] \mathbf{x} \in L(\mathbf{y}) \right\}.$$

It is clear that $D_I(\mathbf{y}, \mathbf{x}) \geq 1$ and that the isoquant is the set of $\mathbf{x}$ values for which $D_I(\mathbf{y}, \mathbf{x}) = 1$. The Debreu (1951)–Farrell (1957) input-based measure of technical efficiency is

$$TE(\mathbf{y}, \mathbf{x}) = \min\{\theta : \theta \mathbf{x} \in \mathrm{L}(\mathbf{y})\}.$$

From the definitions, it follows that $\mathrm{TE}(\mathbf{y}, \mathbf{x}) \leq 1$ and that $\mathrm{TE}(\mathbf{y}, \mathbf{x}) = 1/D_I(\mathbf{y}, \mathbf{x})$. The Debreu-Farrell measure provides a natural starting point for the analysis of efficiency.

The Debreu-Farrell measure is strictly defined in terms of production and is a measure of *technical efficiency*. It does have a significant flaw in that it is wedded to radial contraction or expansion of the input vector. Consider, in figure 2.2, the implied inefficiency of input vector $\mathbf{X}^A$. Figure 2.2 is a conventional isoquant/isocost graph for a single output being produced with two inputs, with price ratio represented by the slope of the isocost line, $\mathbf{ww}'$. With the input vector $\mathbf{X}^A$ normalized to length one, the Debreu-Farrell measure of technical efficiency would be $\theta$, but in economic terms, this measure clearly understates the degree of inefficiency. By scaling back both inputs by the proportion $\theta$, the producer could reach the isoquant and thus achieve technical efficiency, but by reallocating production in favor of input $x_1$ and away from $x_2$, the same output could be produced at even lower cost. Thus, producer A is both technically inefficient and *allocatively inefficient*. The overall efficiency or economic efficiency of producer A is only $\alpha$. Allocative inefficiency and

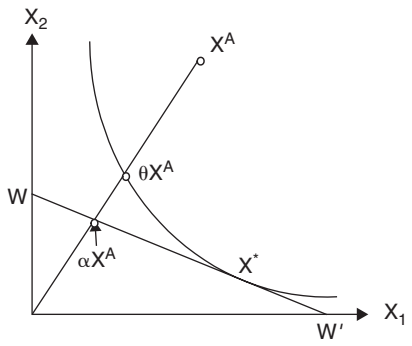


**Figure 2.2.** Technical and Allocative Inefficiency

its implications for econometric measurement of inefficiency are discussed in section 2.9. Empirically decomposing (observable) overall inefficiency, $1 - \alpha$, into its (theoretical, latent) components, technical inefficiency, $(1 - \theta)$, and allocative inefficiency, $(\theta - \alpha)$, is an ongoing and complex effort in the empirical literature on efficiency estimation.

## 2.3 Frontier Production and Cost Functions

The theoretical underpinnings of a model of production are suggested above.[8] Here we take as given the existence of a well-defined production structure characterized by smooth, continuous, continuously differentiable, quasi-concave production or transformation function. Producers are assumed to be price takers in their input markets, so input prices are treated as exogenous. The empirical measurement of TE($\mathbf{y}$, $\mathbf{x}$) requires definition of a transformation function. For most of this analysis, we are concerned with a single-output production frontier. Let

$$y \leq f(\mathbf{x})$$

denote the production function for the single output, $y$, using input vector $\mathbf{x}$. Then, an output-based Debreu-Farrell style measure of technical efficiency is.

$$\mathrm{TE}(y, \mathbf{x}) = \frac{y}{f(\mathbf{x})} \leq 1.$$

Note that the measure is the conventional measure of total factor productivity and that it need not equal the input-based measure defined earlier.

Our econometric framework embodies the Debreu-Farrell interpretation as well as the textbook definition of a production function. Thus, we begin with a model such as

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta})\mathrm{TE}_i,$$

where $0 < \mathrm{TE}(y_i, \mathbf{x}_i) \leq 1, \boldsymbol{\beta}$ is the vector of parameters of the production function to be estimated, and $i$ indexes the $i$th of $N$ firms in a sample to be analyzed. For present purposes, $\boldsymbol{\beta}$ is of secondary interest in the analysis. For example, in the setting of the translog model, parametric functions such as elasticities of substitution or economies of scale are of only marginal interest. The production model is usually linear in the logs of the variables, so the empirical counterpart takes the form

$$\ln y_i = \ln f(\mathbf{x}_i, \boldsymbol{\beta}) + \ln \mathrm{TE}_i = \ln f(\mathbf{x}_i, \boldsymbol{\beta}) - u_i,$$

where $u_i \geq 0$ is a measure of *technical inefficiency* since $u_i = -\ln \mathrm{TE}_i \approx 1 - \mathrm{TE}_i$. Note that

$$\mathrm{TE}_i = \exp(-u_i).$$

[See Jondrow et al. (1982) and Battese and Coelli (1992) for discussion and analysis of the distinction between these two measures.] The preceding provides the central pillar of the econometric models of production that are described below.

Formal econometric analysis of models of this sort as frontier production functions begins with Aigner and Chu's (1968) reformulation of a Cobb-Douglas model. A parallel literature is devoted to the subject of DEA. The centerpiece of DEA is the use of linear programming to wrap a quasi-convex hull around the data in essentially the fashion of Farrell's efficient unit isoquant. The hull delineates the efficient subset defined above, so, by implication, points observed inside the hull are deemed observations on inefficient producers. DEA differs fundamentally from the econometric approach in its interpretation of the data-generating mechanism but is close enough in its philosophical underpinnings to merit at least some consideration here. I turn to the *technique* of DEA in the discussion of deterministic frontiers below.[9]

### 2.3.1 Least squares regression–based estimation of frontier functions

In most applications, the production model, $f(\mathbf{x}_i, \boldsymbol{\beta})$, is linear in the logs of the inputs or functions of them, and the log of the output variable appears on the left-hand side of the estimating equation. It is convenient to maintain that formulation and write

$$\ln y_i = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i,$$

where $\varepsilon_i = -u_i$, and $\mathbf{x}_i$ is the set of whatever functions of the inputs enter the empirical model. We assume that $\varepsilon_i$ is randomly distributed across firms. An important assumption, to be dropped later, is that the distribution of $\varepsilon_i$ is independent of all variables in the model. For present purposes, we assume that $\varepsilon_i$ is a *nonzero* (negative) mean, constant variance, and otherwise ordinary regression disturbance. The assumptions thus far include $E[\varepsilon_i|\mathbf{x}_i] \leq 0$, but absent any other special considerations, this is a classical linear regression model.[10] The model can thus be written

$$\ln y_i = (\alpha + E[\varepsilon_i]) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + (\varepsilon_i - E[\varepsilon_i]) = \alpha * + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i^*.$$

This defines a classical linear regression model. Normality of the disturbance is precluded, since $\varepsilon_i^*$ is the difference between a random variable that is always negative and its mean. Nonetheless, the model's parameters can be consistently estimated by ordinary least squares (OLS) since OLS is robust to nonnormality. Thus, the technical parameters of the production function, with the exception of the constant term, can be estimated consistently, if not efficiently by OLS. If the distribution of $\varepsilon$ were known, the parameters could be estimated more efficiently by maximum likelihood (ML). Since the constant term usually reveals nothing more than the units of measurement of the left-hand side variable in

this model, one might wonder whether all of this is much ado about nothing, or at least very little. But, one might argue that, in the present setting, the constant is the *only* parameter of interest. Remember, it is the residuals and, by construction, now $E[u_i|\mathbf{x}_i]$ that are the objects of estimation. Three approaches may be taken to examine these components.

(1) Since only the constant term in the model is inconsistent, any information useful for comparing firms to each other that would be conveyed by estimation of $u_i$ from the residuals can be obtained directly from the OLS residuals,

$$e_i = \ln y_i - a^* - \mathbf{b}^{\mathrm{T}}\mathbf{x}_i = -u_i + E[u_i],$$

where $\mathbf{b}$ is the least squares coefficient vector in the regression of $\ln y_i$ on a constant and $\mathbf{x}_i$.

Thus, for example, $e_i - e_m$ is an unbiased and pointwise consistent estimator of $u_j - u_m$. Likewise, the ratio estimator $\exp(e_i)/\exp(e_m)$ estimates

$$\frac{\mathrm{TE}_i \exp(E[u_i])}{\mathrm{TE}_m \exp(E[u_m])} = \frac{\mathrm{TE}_i}{\mathrm{TE}_m}$$

consistently (albeit with a finite sample bias because of the nonlinearity of the function). For purposes of comparison of firms only, one could simply ignore the frontier aspect of the model in estimation and proceed with the results of OLS. This does preclude any sort of estimator of $\mathrm{TE}_i$ or of $E[u_i]$, but for now this is not consequential.

(2) Since the only deficiency in the OLS estimates is a displacement of the constant term, one might proceed simply by "fixing" the regression model. Two approaches have been suggested. Both are based on the result that the OLS estimator of the slope parameters is consistent and unbiased, so the OLS residuals are pointwise consistent estimators of linear translations of the original $u_i$ values. One simple remedy is to shift the estimated production function upward until all residuals except one, on which we hang the function, are negative. The intercept is shifted to obtain the corrected OLS (COLS) constant,

$$a_{\mathrm{COLS}} = a^* + \max_i e_i.$$

All of the COLS residuals,

$$e_{i,\mathrm{COLS}} = e_i - \max_i e_i,$$

satisfy the theoretical restriction. Proofs of the consistency of this COLS estimator, which require only that, in a random sample drawn from the population $u_i$, $\operatorname{plim} \min_i u_i = 0$, appear in Gabrielsen (1975) and Greene (1980a). The logic of the estimator was first suggested much earlier by Winsten (1957). A lengthy application with an extension to panel data appears in Simar (1992).

In spite of the methodological problems to be noted below, this has been a popular approach in the analysis of panel data (see, e.g., Cornwell, Schmidt, and Sickles, 1990; Evans et al., 2000a, 2000b).

(3) An alternative approach that requires a parametric model of the distribution of $u_i$ is modified OLS (MOLS). [The terminology was suggested by Lovell (1993, p. 21).] The OLS residuals, save for the constant displacement, are pointwise consistent estimates of their population counterparts, $-u_i$. The mean of the OLS residuals is useless—it is zero by construction. But, since the displacement is constant, the variance and any higher order *central* moment of (the negatives of) the OLS residuals will be a consistent estimator of the counterpart of $u_i$. Thus, if the parameters of $E[u_i]$ are identified through the variance or, perhaps, higher moments or other statistics, then consistent estimation of the deeper model parameters may be completed by using the method of moments. For a simple example, suppose that $u_i$ has an exponential distribution with mean $\lambda$. Then, the variance of $u_i$ is $\lambda^2$, so the standard deviation of the OLS residuals is a consistent estimator of $E[u_i] = \lambda$. Since this is a one-parameter distribution, the entire model for $u_i$ can be characterized by this parameter and functions of it.[11] The estimated frontier function can now be displaced upward by this estimate of $E[u_i]$. This MOLS method is a bit less orthodox than the COLS approach described above since it is unlikely to result in a full set of negative residuals. The typical result is shown in figure 2.3.

A counterpart to the preceding is possible for analysis of the costs of production. In this case, the working assumption is that the estimated cost function lies under all the data, rather than above them.

The received literature contains discussion of the notion of an "average" frontier (an oxymoron, perhaps), as opposed to the "best-practice" frontier, based on the distinction between OLS and some technique, such as ML, which takes account of the frontier nature of the underlying model. One could argue that the former is being defined with respect to an estimator, OLS, rather than with respect to a definable, theoretically specified model. Whether the

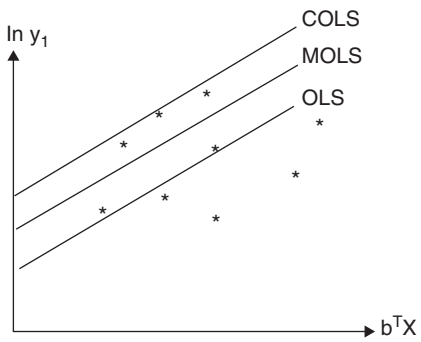


**Figure 2.3.** OLS Production Frontier Estimators

distinction is meaningful in an economic sense is questionable. There is some precedent for raising the question of whether the technology in use "at the frontier" differs from that in the middle of the pack, so to speak (see Klotz, Madoo, and Hansen, 1980), but the simple scaling of a loglinear production function is unlikely to reveal much about this possibility. Indeed, the implied radial expansion of the production function thus formulated might reveal nothing more than different rates of adoption of Hicks neutral technical innovations. But Førsund and Jansen (1977) argue that this difference or, more generally, differential rates of adjustment of capital stocks across firms in an industry *do* create a meaningful distinction between average and best-practice production frontiers. Some firms in an industry might be achieving the maximum output attainable, that is, be locating themselves on the frontier that applies to them, but might not have completely adjusted their capital stock to the most up-to-date, technically advanced available. Thus, the best-practice frontier for an industry that reflects this full adjustment would lie outside the frontiers applicable to some of the constituent firms (see Førsund and Hjalmarsson, 1974, for additional discussion). The description, as I show later, is akin to the motivation for the stochastic frontier model. However, the posited differences between firms are more systematic in this description.

### 2.3.2 Deterministic frontier models

Frontier functions as specified above, in which the deviation of an observation from the theoretical maximum is attributed solely to the inefficiency of the firm, are labeled *deterministic frontier functions*. This is in contrast to the specification of the frontier in which the maximum output that a producer can obtain is assumed to be determined both by the production function and by random external factors such as luck or unexpected disturbances in a related market. Under this second interpretation, the model is recast as a *stochastic frontier production function*, which is the subject of section 2.4.

Aigner and Chu (1968) suggested a loglinear (Cobb-Douglas) production function,

$$Y_i = AX_{1i}^{\beta_1} X_{2i}^{\beta_2} U_i,$$

in which $U_i$ (which corresponds to $\text{TE}_i$) is a random disturbance between 0 and 1. Taking logs produces

$$\begin{aligned} \ln Y_i &= \alpha + \sum_{k=1}^{K} \beta_k x_{ki} + \varepsilon_i \\ &= \alpha + \sum_{k=1}^{K} \beta_k x_{ki} - u_i, \end{aligned}$$

where $\alpha = \ln A, x_{ki} = \ln X_{ki}$, and $\varepsilon_i = \ln U_i$. The nonstochastic part of the right-hand side is viewed as the frontier. It is labeled "deterministic" because the stochastic component of the model is entirely contained in the (in)efficiency term, $-u_i$. Aigner and Chu (1968) suggested two methods of computing the parameters that would constrain the residuals $u_i$ to be nonnegative, linear programming,

$$\min_{\alpha,\beta} \sum_{i=1}^{N} \varepsilon_i \text{ subject to } \ln y_i - \alpha - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i \leq 0 \forall i,$$

and quadratic programming,

$$\min_{\alpha,\beta} \sum_{i=1}^{N} \varepsilon_i^2 \text{ subject to } \ln y_i - \alpha - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i \leq 0 \forall i.$$

In both applications, the slack variables associated with the constraints produce the estimates of $-u_i$. A number of early applications, such as Førsund and Jansen (1977), built upon this approach both to study the technical aspects of production and to analyze technical efficiency.

The Aigner-Chu (1968) approach satisfies the original objective. One can compare the individual residuals based on the programming results,

$$\hat{u}_i = \hat{\alpha} + \hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{x}_i - \ln y_i,$$

to each other or to an absolute standard to assess the degree of technical (in)efficiency represented in the sample. A summary measure that could characterize the entire sample would be the

$$\text{average technical inefficiency} = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i.$$

Another useful statistic would be the

$$\text{average technical inefficiency} = \frac{1}{N} \sum_{i=1}^{N} e^{-\hat{u}_i} = \hat{E}[\mathrm{TE}_i].$$

This semiparametric approach was applied in a series of studies including Førsund and Hjalmarsson (1979), Albriktsen and Førsund (1990), and Førsund (1992). In these studies, the authors specified the generalized production function proposed by Zellner and Revankar (1969),

$$\gamma_0 \ln y_i + \gamma_1 y_i = \alpha + \sum_{k=1}^{K} \beta_k x_{ki},$$

and minimized the sum of the residuals subject to the additional constraints $\sum_k \beta_k = 1$ and $(\gamma_0, \gamma_1, \beta_k, k = 1, \ldots, K) > 0$. The foci of the applications are economies of scale and technical efficiency.

#### 2.3.2.1 Statistical issues

The programming procedures are not based explicitly on an assumed statistical model. The properties of the "estimators" are therefore ambiguous—they would depend on what process actually did generate the data. (This is the logic of much of the contemporary discussion of how to bridge the econometric and DEA approaches to efficiency estimation. See, e.g., Simar and Wilson, 1998, 1999; see also chapter 4 of this volume.) The programming estimators have the notable disadvantage that they do not naturally produce standard errors for the coefficients, so inference is precluded. For present purposes, the main disadvantage is that absent a more detailed specification, consistency of the estimates cannot be verified, nor, as such, can consistency of the inefficiency estimates, $-u_i$. The programming procedures might, however, have the virtue of robustness to specification errors in the distribution of $u_i$, though this, too, remains to be verified and would depend on an underlying statistical specification (see Simar, 1996; Cazals, Florens, and Simar, 2002). Under the presumption that there is some common underlying stochastic process generating the observed data, one could proceed from here by using bootstrapping to attempt to deduce the properties of the estimators. (Again, this is an approach that has been employed to study the behavior of DEA techniques; see Simar and Wilson, 1998, 1999; see also chapter 4 this volume.) However, from a larger standpoint, it is a moot point, because the estimators themselves are no longer employed in estimating inefficiency. DEA has supplanted the linear programming approach, and the quadratic programming approach is now only of historical interest.

Schmidt (1976) observed that the Aigner-Chu optimization criteria could be construed as the log-likelihood functions for models in which one-sided residuals were distributed as exponential for the linear programming estimator, and half-normal for the quadratic programming approach. This would appear to endow the programming procedures with a statistical pedigree. However, neither log-likelihood function has a zero root, and the Hessians of both log-likelihoods are singular. The former contains a diagonal block of zeros, while the latter has a zero eigenvalue.[12] Therefore, one cannot base statistical inference on the standard results for ML estimators (MLEs) in these settings. The inference problem remains.

The statistical problem with Schmidt's estimators is a violation of the regularity conditions for MLE. This leaves the possibility that, for other distributions, the regularity conditions might be met, and as such, a well-behaved likelihood function for a one-sided disturbance might still be definable. Greene (1980a) proposed a model based on the gamma distribution,

$$h(u_i) = \frac{\theta^P}{\Gamma(P)} u_i^{P-1} e^{-\theta u_i}, \quad u_i \geq 0, \theta > 0, P > 2.$$

The density is defined for all $P > 0$, but $P > 2$ is required for a well-behaved log-likelihood function for the frontier model.[13] The gamma frontier model

does produce a bona fide MLE, with all of the familiar properties. In principle, the log-likelihood,

$$\ln L(\alpha, \boldsymbol{\beta}, P, \theta) = P \ln \theta - N \ln \Gamma(P) + (P-1) \sum_{i=1}^{N} \ln u_i - \theta \sum_{i=1}^{N} u_i,$$
$$\theta > 0, P > 2, u_i = \alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i - y_i > 0$$

can be maximized by conventional methods. The restriction that all sample residuals must be kept strictly positive for the estimator to be computable turns out to be a persistent and major complication for iterative search methods. However, inference can proceed as in more conventional problems. In spite of the practical complications, there have been several applications, including Greene (1980a, 1980b), Stevenson (1980), Aguilar (1988), Hunt, Kim, and Warren (1986), Chen and Tang (1989), and Hunt-McCool and Warren (1993). An extension that adds firm-specific effects to the efficiency term is described in Deprins and Simar (1989a). Like other deterministic frontiers, the gamma frontier model above is largely of historical interest. The contemporary work has focused on the stochastic frontier model as a preferable approach, for reasons discussed below. However, the gamma frontier model has retained some currency as the subject of several more recent studies and, in particular, as the platform for several Bayesian estimators (see Tsionas, 2000b, 2002; Huang, 2004; Koop et al., 1999; discussed below).

#### 2.3.2.2 Deterministic cost frontiers

Førsund and Jansen (1977) formulated a hybrid of the linear programming approaches and the parametric model above to extend the analysis to costs of production. The Førsund and Jansen specification departs from a homothetic production function,[14]

$$y_i = F[f(\mathbf{x}_i)], F'[f(\mathbf{x}_i)] > 0, f(t\mathbf{x}_i) = tf(\mathbf{x}_i) \forall \mathbf{x}_i.$$

The empirical model is obtained by specifying

$$y_i = F[f(\mathbf{x}_i) v_i],$$

where $v_i$ has a beta density (with parameters $\theta + 1$ and 1)

$$h(v_i) = (1+\theta) v_i^{\theta}, 0 < v_i < 1, \theta > 0.$$

The cost function that corresponds to this production function is

$$\ln C_i = \ln F^{-1}(y_i) + \ln c(\mathbf{w}_i) - \ln v_i,$$

where $\mathbf{w}_i$ is the vector of input prices, and $c(\mathbf{w}_i)$ is the unit cost function. The authors then derive the corresponding log-likelihood function. The parameters of the production function are obtained by using linear programming

to minimize $\sum_{i=1}^{N} \ln C_i$ subject to the constraints that observed costs lie *on or above* the cost frontier.[15] There are three aspects of this approach that are of interest here. First, this construction is an alternative attempt to derive the linear programming criterion as the solution to an ML problem.[16] Second, this is one of the first applications to estimate a *cost frontier* instead of a production frontier. There is some controversy to this exercise owing to the possible contribution of allocative inefficiency to the observed estimates of firm inefficiency. Third, there is a subtle sleight-of-hand used in formulating the cost function. If the technical inefficiency component, $v_i$, were to enter the production function more naturally, *outside* the transformation of the core function, the form in which it entered the cost frontier would be far more complicated. On the other hand, if the inefficiency entered the production function in the place of $v_i\mathbf{x}_i$, inside the homogeneous kernel function (in the form of input-oriented inefficiency), then its appearance in the cost function would be yet more complicated (see, e.g., Kumbhakar and Tsionas, 2005a; Kurkalova and Carriquiry, 2003).

#### 2.3.2.3 COLS and MOLS estimators

The slope parameters in the deterministic frontier models can be estimated consistently by OLS. The constant term can be consistently estimated simply by shifting the least squares line upward sufficiently that the largest residual is zero. The resulting efficiency measures are $-\hat{u}_i = e_i - \max_i e_i$. Thus, absolute estimators of the efficiency measures in this model are directly computable using nothing more elaborate than OLS. In the gamma frontier model, $a$, the OLS estimate of $\alpha$ converges to plim $a = \alpha - E[u_i] = \alpha - (P/\theta)$. So, another approach would be to correct the constant term using estimates of $P$ and $\theta$. The gamma model also produces individual estimates of technical efficiency. A summary statistic that might also prove useful is $E[u_i] = P/\theta = \mu$, which can be estimated with the corrected residuals. Likewise, an estimate of $\text{var}[u_i] = P/\theta^2 = \sigma_u^2$ is produced by the least squares residual variance. Combining the two produces a standardized mean $\mu/\sigma_u = \sqrt{P}$. Here, as elsewhere, functions of the OLS parameter estimates and residuals can be used to obtain estimates of the underlying structural parameters. Consistent estimators of $\theta = P/\mu$ and $P = \theta\mu$ are easily computed. Using this correction to the least squares constant term produces the MOLS estimator. Another useful parameter to estimate is $E[\exp(-u_i)] = [\theta/(1+\theta)]^P$. A similar approach is taken by Afriat (1972), who suggests that $u_i$ be assumed to have a one-parameter gamma distribution, with $\theta = 1$ in the preceding. Richmond (1974) builds on Afriat's model to obtain the distribution of $e_i^{-u}$ and then derives $E[\exp(-u_i)]$ and other population moments.[17] Both authors suggest that the OLS residuals be used to estimate these parameters. As Richmond demonstrates, $P$ can be consistently estimated simply by using the standard deviation of the OLS residuals.

#### 2.3.2.4 Methodological questions

A fundamental practical problem with the gamma and all other deterministic frontiers is that any measurement error and any other outcome of stochastic variation in the dependent variable must be embedded in the one-sided disturbance. In any sample, a single errant observation can have profound effects on the estimates. Unlike measurement error in $y_i$, this outlier problem is not alleviated by resorting to large sample results.

There have been a number of other contributions to the econometrics literature on specification and estimation of deterministic frontier functions. Two important papers that anticipated the stochastic frontier model discussed in the next section are Timmer (1971), which proposed a probabilistic approach to frontier modeling that allowed *some* residuals to be positive, and Aigner, Amemiya, and Poirier (1976), who, in a precursor to Aigner et al. (1977), focused on asymmetry in the distribution of the disturbance as a reflection of technical inefficiency. Applications of the parametric form of the deterministic frontier model are relatively uncommon. The technical problems are quite surmountable, but the inherent problem with the stochastic specification and the implications of measurement error render it problematic. The nonparametric approach based on linear programming has an intuitive appeal and now dominates this part of the literature on frontier estimation.

### 2.3.3 Data envelopment analysis

DEA is a body of techniques for analyzing production, cost, revenue, and profit data, essentially, without parameterizing the technology. This constitutes a growth industry in the management science literature, and appears with some frequency in economics, as well.[18] We begin from the premise that *there exists a production frontier* that acts to constrain the producers in an industry. With heterogeneity across producers, they will be observed to array themselves at varying distances from the efficient frontier. By wrapping a hull around the observed data, we can reveal which among the set of observed producers are closest to that frontier (or farthest from it). Presumably, the larger the sample, the more precisely this information will be revealed. In principle, the DEA procedure constructs a piecewise linear, quasi-convex hull around the data points in the input space. As in our earlier discussions, technical efficiency requires production on the frontier, which in this case is the observed best practice. Thus, DEA is based fundamentally on a comparison among observed producers. Once again, to argue that this defines or estimates an ideal in any sense requires the analyst to assume, first, that there exists an ideal production point and, second, that producers strive to achieve that goal. Without belaboring the obvious, it is not difficult to construct situations in which the second of these would be difficult to maintain. The service sectors of the recently dismantled centrally planned economies of Eastern Europe come to mind as cases in point.

There are many survey-style treatments of DEA, including chapter 3 of this book. Because this chapter is devoted to econometric approaches to efficiency analysis, I eschew presentation of any of the mathematical details. Another brief (tight) and very readable sketch of the body of techniques is given in Murillo-Zamorano (2004, pp. 37–46).

The DEA method of modeling technical and allocative efficiency is largely atheoretical. Its main strength may be its lack of parameterization; it requires no assumptions about the form of the technology. The piecewise linearity of the efficient isoquant might be problematic from a theoretical viewpoint, but that is the price for the lack of parameterization. The main drawback is that shared with the other deterministic frontier estimators. Any deviation of an observation from the frontier must be attributed to inefficiency.[19] There is no provision for statistical noise or measurement error in the model. The problem is compounded in this setting by the absence of a definable set of statistical properties. Recent explorations in the use of bootstrapping methods has begun to suggest solutions to this particular shortcoming (see, e.g., Xue and Harker, 1999; Simar and Wilson, 1998, 1999; Tsionas, 2001b, which used efficiency measures produced by a DEA as priors for inefficiency in a hierarchical Bayes estimation of a stochastic frontier).

I do not return to the subject of DEA in this chapter, so at this point I note a few of the numerous comparisons that have been made between (nonparametric) DEA and statistics-based frontier methods, both deterministic and stochastic. Several studies have analyzed data with both DEA and parametric, deterministic frontier estimators. For example, Bjurek, Hjalmarsson, and Forsund (1990) used the techniques described above to study the Swedish social insurance system. Førsund (1992) did a similar analysis of Norwegian ferries. In both studies, the authors do not observe radical differences in the results with the various procedures. That is perhaps not surprising since the main differences in their specifications concerned functional form: Cobb-Douglas for the parametric models, and piecewise linear for the nonparametric ones. The differences in the inferences one draws often differ more sharply when the statistical underpinnings are made more detailed in the stochastic frontier model, but even here, the evidence is mixed. Ray and Mukherjee (1995), using the Christensen and Greene (1976) data on U.S. electricity generation, found good agreement between DEA and stochastic frontier-based estimates. Murillo-Zamorano and Vega-Cervera (2001) find similar results for a later (1990) sample of U.S. electricity generators. Cummins and Zi (1998) also found concordance in their analysis of the U.S. insurance industry. Finally, Chakraborty, Biswas, and Lewis (2001) found in analyzing public education in Utah that the empirical results using the various techniques are largely similar. These studies do stand in contrast to Ferrier and Lovell (1990), who found major differences between DEA and stochastic frontier-based inefficiency estimates in a multiple-out distance function fit in a large sample of American banks. Bauer et al. (1998) likewise found substantial differences between parametric and nonparametric efficiency estimates for a sample of U.S. banks. In

sum, the evidence is mixed, but it does appear that, quite frequently, the overall pictures drawn by DEA and statistical frontier-based techniques are similar. That the two broad classes of techniques fail to produce the same pictures of inefficiencies poses a dilemma for regulators hoping to use the methods to evaluate their constituents (and, since they have the same theoretical underpinning, casts suspicion on both methods). As noted above, this has arisen in more than one study of the banking industry. Bauer et al. (1998) discuss specific conditions that should appear in efficiency methods to be used for evaluating financial institutions, with exactly this consideration in mind.

## 2.4 The Stochastic Frontier Model

The stochastic production frontier proposed by Aigner et al. (1977) and Meeusen and van den Broeck (1977)[20] is motivated by the idea that deviations from the production "frontier" might not be entirely under the control of the firm being studied. Under the interpretation of the deterministic frontier of the preceding section, some external events, for example, an unusually high number of random equipment failures, or even bad weather, might ultimately appear to the analyst as inefficiency. Worse yet, any error or imperfection in the specification of the model or measurement of its component variables, including the (log) output, could likewise translate into increased inefficiency measures. This is an unattractive feature of any deterministic frontier specification. A more appealing formulation holds that any particular firm faces its own production frontier, and that frontier is randomly placed by the whole collection of stochastic elements that might enter the model outside the control of the firm. [This is a similar argument to Førsund and Jansen's (1977) rationale for an average vs. best-practice frontier function.] An appropriate formulation is

$$y_i = f(\mathbf{x}_i)\mathrm{TE}_i e^{v_i},$$

where all terms are as defined above and $v_i$ is unrestricted. The latter term embodies measurement errors, any other statistical noise, and random variation of the frontier across firms. The reformulated model is

$$\ln y_i = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + v_i - u_i = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i.$$

(The issue of functional form was considered in section 2.2.2. I use the linear specification above generically here.) As before, $u_i > 0$, but $v_i$ may take any value. A symmetric distribution, such as the normal distribution, is usually assumed for $v_i$. Thus, the *stochastic frontier* is $\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + v_i$, and as before, $u_i$ represents the inefficiency.

Note, before beginning this lengthy section, that the ultimate objective in the econometric estimation of frontier models is to construct an estimate of $u_i$ or at least $u_i - \min_i u_i$. The first step, of course, is to compute the technology

parameters, $\alpha$, $\boldsymbol{\beta}$, $\sigma_u$, and $\sigma_v$ (and any other parameters). It does follow that, if the frontier model estimates are inappropriate or inconsistent, then estimation of the inefficiency component of $\varepsilon_i$, that is, $u_i$, is likely to be problematic, as well. So, we first consider estimation of the technology parameters. Estimation of $u_i$ is considered in detail in section 2.8.

### 2.4.1 Implications of least squares

In the basic specification above, both components of the compound disturbance are generally assumed to be independent and identically distributed (iid) across observations.[21] As long as $E[v_i - u_i]$ is constant, the OLS estimates of the slope parameters of the frontier function are unbiased and consistent. The average inefficiency present in the distribution is reflected in the asymmetry of the distribution, a quantity that is easily estimable, even with the results of OLS, with the third moment of the residuals,

$$m_3 = \frac{1}{N}\sum_{i=1}^{N}(\hat{\varepsilon}_i - \hat{E}[\varepsilon_i])^3,$$

however estimated, as long as the slope estimators are consistent. By expanding

$$\mu_3 = E[v_i - (u_i - E[u_i])]^3,$$

we see that, in fact, the skewness of the distribution of the estimable disturbance, $\varepsilon_i$, is simply the negative of that of the latent inefficiency component, $u_i$. So, for example, regardless of the assumed underlying distribution, the negative of the third moment of the OLS residuals provides a consistent estimator of the skewness of the distribution of $u_i$. Since this statistic has units of measurement equal to the cube of those of the log of output, one might, as a useful first step in any analysis, examine the conventional normalized measure, $\sqrt{b_3} = -m_3/s^3$, where $s$ is the sample standard deviation of the residuals. Values between 0 and 4 are typical. A Wald test of the hypothesis of no systematic inefficiency in the distribution could be based on the familiar chi-squared test,[22]

$$\chi_1^2 = \frac{1}{6}\left[\frac{-m_3}{s^3}\right]^2.$$

The skewness coefficient of the least squares residuals in any finite sample could have the "wrong" sign (positive in this case). This might cast doubt on the specification of the stochastic frontier model and suggest that the Wald test is meaningless.[23] Other tests of the stochastic frontier specification are presented in Schmidt and Lin (1984). The skewness of the residuals turns out to be an important indicator of the specification of the stochastic frontier model. I emphasize, however, that this is merely a sample statistic subject to sampling variability. The skewness is only suggestive—$m_3$ could be positive even if the

stochastic frontier model is correct. Indeed, for a nonnormal specification of the random components, $\mu_3$ could be positive in the population.

### 2.4.2 Forming the likelihood function

We begin with a general formulation of the model and then narrow the specification to the particular models that have been proposed in the contemporary literature. The generic form of the stochastic frontier is

$$\begin{aligned}\ln y_i &= \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + v_i - u_i \\ &= \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i.\end{aligned}$$

It is convenient to start with the simplest assumptions, that

(a) $f_v(v_i)$ is a symmetric distribution;
(b) $v_i$ and $u_i$ are statistically independent of each other; and
(c) $v_i$ and $u_i$ are independent and identically distributed across observations.

Thus, our starting point has both error components with constant means 0 and $\mu$ and variances $\sigma_v^2$ and $\sigma_u^2$, respectively, over all observations. To form the density of $\ln y_i$ that underlies the likelihood function, we use these assumptions to obtain the joint density of the components,

$$f_{v,u}(v_i, u_i) = f_v(v_i) f_u(u_i).$$

Then, $\varepsilon_i = v_i - u_i$, so

$$f_{\varepsilon,u}(\varepsilon_i, u_i) = f_u(u_i) f_v(\varepsilon_i + u_i).$$

[The Jacobian of the transformation from $(v, u)$ to $(\varepsilon, u)$ is $\det \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1} = 1$.]
Finally, to obtain the marginal density of $\varepsilon_i$, we integrate $u_i$ out of the joint density:

$$f_\varepsilon(\varepsilon_i) = \int_0^\infty f_u(u_i) f_v(\varepsilon_i + u_i)\, du_i$$

The final step gives the contribution of observation $i$ to the log-likelihood

$$\ln L_i(\alpha, \boldsymbol{\beta}, \sigma_u^2, \sigma_v^2 | \ln y_i, \mathbf{x}_i) = \ln f_\varepsilon(y_i - \alpha - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i | \sigma_u^2, \sigma_v^2).$$

In several important cases examined below, the integral has a convenient closed form so that estimation of the model by ML or through Bayesian methods based on the likelihood is straightforward. Note, however, that with current techniques of simulation-based estimation, closed forms for integrals such as this are not always necessary for estimation.[24]

The derivation above requires a trivial modification for a cost frontier. In this case,

$$\ln C_i = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + v_i + u_i.$$

(For convenience here, we retain the symbol $\mathbf{x}$ for the variables in the frontier function, though in a cost function, they would be output and the input prices, not the inputs.) Then, $\varepsilon_i = v_i + u_i$ and $f_{\varepsilon,u}(\varepsilon_i, u_i) = f_u(u_i)f_v(\varepsilon_i - u_i)$. Since $v_i$ is assumed to have a symmetric distribution, the second term may be written $f_v(\varepsilon_i - u_i) = f_v(u_i - \varepsilon_i)$. Making this simple change, we see that in order to form the density for log cost for a particular model in which observations lie above the frontier, it is necessary only to reverse the sign of $\varepsilon_i$ where it appears in the functional form. An example below will illustrate.

### 2.4.3 The normal–half-normal model

The compound disturbance in the stochastic frontier model, while asymmetrically distributed, is, for most choices of the disturbance distributions, otherwise well behaved. MLE is generally straightforward. The literature on stochastic frontier models begins with Aigner et al.'s (1977) normal–half-normal model, in which

$$f_v(v_i) = N[0, \sigma_v^2] = (1/\sigma_v)\phi(v_i/\sigma_v), -\infty < v_i < \infty$$

and

$$u_i = |U_i| \text{ where } f_U(U_i) = N[0, \sigma_u^2] = (1/\sigma_u)\phi(U_i/\sigma_u), -\infty < U_i < \infty,$$

where $\phi(.)$ denotes the standard normal density. The resulting density for $u_i$ is

$$f_u(u_i) = [1/\Phi(0)](1/\sigma_u)\phi(u_i/\sigma_u), 0 \le u_i < \infty,$$

where $\Phi(.)$ is the standard normal cumulative distribution function (CDF). The symmetrically distributed $v_i$ is usually to be assumed to be normal, which we denote $f(v_i) = N[0, \sigma_v^2]$. The distribution of the compound random variable $\varepsilon_i = (v_i - u_i)$ has been derived by Weinstein (1964) and is discussed in Aigner et al. (1977).[25] The end result, maintaining the form above, is

$$f_\varepsilon(\varepsilon_i) = \frac{2}{\sqrt{2\pi(\sigma_u^2 + \sigma_v^2)}}\left[\Phi\left(\frac{-\varepsilon_i(\sigma_u/\sigma_v)}{\sqrt{\sigma_u^2 + \sigma_v^2}}\right)\right]\exp\left(\frac{-\varepsilon_i^2}{2(\sigma_u^2 + \sigma_v^2)}\right).$$

A convenient parameterization that also produces a useful interpretation is $\sigma^2 = (\sigma_u^2 + \sigma_v^2)$ and $\lambda = \sigma_u/\sigma_v$.[26] Then,

$$f_\varepsilon(\varepsilon_i) = \frac{2}{\sigma\sqrt{2\pi}}\phi\left(\frac{\varepsilon_i}{\sigma}\right)\left[\Phi\left(\frac{-\varepsilon_i\lambda}{\sigma}\right)\right].$$

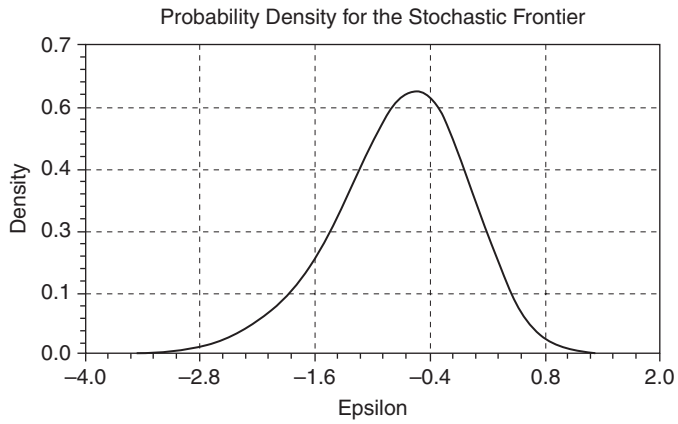


**Figure 2.4.** Density of a Normal Minus a Half-Normal

This density is skewed in the negative direction (see the above discussion). Figure 2.4 illustrates the shape of the distribution for $\lambda = 2$ and $\sigma = 1$. The constructed parameter $\lambda = \sigma_u/\sigma_v$ characterizes the distribution. If $\lambda \to +\infty$, the deterministic frontier results. If $\lambda \to 0$, the implication is that there is no inefficiency in the disturbance, and the model can be efficiently estimated by OLS.

With the assumption of a half-normal distribution, we obtain $E[u] = \sigma_u\sqrt{2/\pi}$ and $\text{var}[u_i] = \sigma_u^2[(\pi - 2)/\pi]$. A common slip in applications is to treat $\sigma_u^2$ as the variance of $u_i$. In fact, this overstates the variance by a factor of nearly 3! Since $\sigma_u$ is not the standard deviation of $u_i$, it gives a somewhat misleading picture of the amount of inefficiency that the estimates suggest is present in the data. Likewise, although $\lambda$ is indicative of this aspect of the model, it is primarily a convenient normalization, not necessarily a directly interpretable parameter of the distribution. It might seem that the variance ratio $\sigma_u^2/\sigma^2$ would be a useful indicator of the influence of the inefficiency component in the overall variance. But again, the variance of the truncated-normal random variable $u_i$ is $\text{var}[U_i|U_i > 0] = [(\pi - 2)/\pi]\sigma_u^2$, not $\sigma_u^2$. In the decomposition of the total variance into two components, the contribution of $u_i$ is

$$\frac{\text{var}[u]}{\text{var}[\varepsilon]} = \frac{[(\pi - 2)/\pi]\sigma_u^2}{[(\pi - 2)/\pi]\sigma_u^2 + \sigma_v^2}.$$

Further details on estimation of the half-normal model may be found in Aigner et al. (1977) and in Greene (2003a). The parameter $\lambda$ is the inefficiency component of the model. The simple regression model results if $\lambda$ equals zero. The implication would be that every firm operates on its frontier. This does not imply, however, that one can "test" for inefficiency by the usual means, because the polar value, $\lambda = 0$, is on the boundary of the parameter space, not

in its interior. Standard tests, such as the Lagrange multiplier test, are likely to be problematic.[27]

The log-likelihood function for the normal–half-normal stochastic frontier model is

$$\text{Ln } L(\alpha, \beta, \sigma, \lambda) = -N \ln \sigma - \text{ constant} + \sum_{i=1}^{N} \left\{ \ln \Phi \left[ \frac{-\varepsilon_i \lambda}{\sigma} \right] - \frac{1}{2} \left[ \frac{\varepsilon_i}{\sigma} \right]^2 \right\},$$

where

$\varepsilon_i = \ln y_i - \alpha - \beta^{\mathrm{T}} x_i$, $\lambda = \sigma_u / \sigma_v$, $\sigma^2 = \sigma_u^2 + \sigma_v^2$, and $\Phi$ = the standard normal CDF.

The log-likelihood function is quite straightforward to maximize and has been integrated into several contemporary commercial computer packages, including *Frontier 4.1* (Coelli, 1996), *LIMDEP* (Greene, 2000), *Stata* (Stata, Inc., 2005), and *TSP* (TSP International, 2005; see also Greene, 2003a, for discussion of maximizing this log-likelihood). The normal–half-normal model has an intriguing and useful feature. Regarding an above point about the "incorrect" skewness of the least squares, Waldman (1982) has shown that in estimation of a stochastic production (cost) frontier with the normal–half-normal model, if the OLS residuals, ignoring the frontier function altogether, are positively (negatively) skewed (i.e., in the wrong direction), then the maximizers of the log-likelihood are OLS for ($\alpha$, $\boldsymbol{\beta}$, $\sigma^2$) and zero for $\lambda$.[28] This is a very useful self-diagnostic on specification and estimation of this frontier model.[29]

### 2.4.4 Normal–exponential and normal–gamma models

The assumption of half-normality has seemed unduly narrow, and numerous alternatives have been suggested. Meeusen and van den Broeck (1977) and Aigner et al. (1977) presented the log-likelihood and associated results for an exponentially distributed disturbance,[30]

$$f_u(u_i) = \theta \exp(-\theta u_i), \theta > 0, u_i \geq 0.$$

In the exponential model, $\sigma_u = 1/\theta$. To maintain continuity, it is helpful to use this parameterization. With this assumption,

$$\text{Ln } L(\alpha, \beta, \sigma_v, \sigma_u) = \sum_{i=1}^{N} \left[ -\ln \sigma_u + \frac{1}{2} \left( \frac{\sigma_v}{\sigma_u} \right)^2 + \ln \Phi \left( \frac{-(\varepsilon_i + \sigma_v^2 / \sigma_u)}{\sigma_v} \right) + \frac{\varepsilon_i}{\sigma_u} \right].$$

MLE with this distribution is straightforward, as well, although, as discussed below, there can be some complications involved with obtaining starting values.[31] The asymptotic covariance matrix of the estimated parameters is

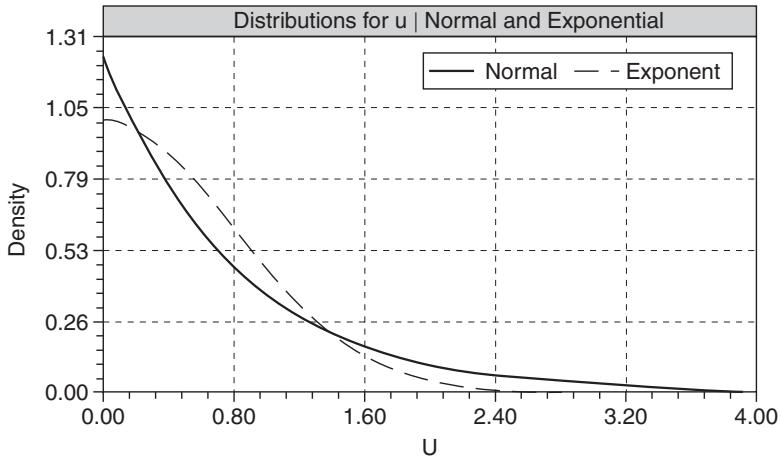


**Figure 2.5.** Half-Normal and Exponential Distributions

typically estimated by the Berndt, Hall, Hall and Hausman "outer product of gradients" method (Greene, 2003a), though the analytic Hessians are not overly burdensome (see Aigner et al., 1977).

The exponential model actually is qualitatively different from the half-normal. Figure 2.5 shows the half-normal distribution with $\sigma_u = 0.8$—this is the one that underlies figure 2.4—and the exponential distribution with $\theta = 1.659$, which implies the same standard deviation $[0.8(\pi - 2)/\pi = 0.603]$. As shown in figure 2.5, for a given value of the parameter, the exponential implies a tighter clustering of the values near zero. In practice, as explored below, this seems to make only a marginal difference in the estimated inefficiencies.

### 2.4.5 Bayesian estimation

Since van den Broeck et al. (1994) and Koop et al. (1994, 1995), there has been an active and rapid development of Bayesian estimators for stochastic frontier models.[32] Rather than treat this as a separate literature, which it is decidedly not, here I detail the basic form of the method and describe some of the numerous extensions in the different sections below, for example, on the gamma model and on multiple-output cost and distance functions. For reasons noted shortly, the basic platform for the Bayesian frontier estimator is the normal–exponential model. [I draw on Koop and Steel (2001) as a useful pedagogy for the interested reader.[33] Also, in the interest of readability, I deviate slightly from the conventional notation in Bayesian applications in which densities are usually formulated in terms of precision parameters (reciprocals of variances) rather than the natural parameters of the original model.]

The log of the likelihood function for the normal–exponential model is

$$\text{Ln } L(\text{data}; \alpha, \boldsymbol{\beta}, \sigma_v, \sigma_u) = \sum_{i=1}^{N} \left[ -\ln \sigma_u + \frac{1}{2}\left(\frac{\sigma_v}{\sigma_u}\right)^2 + \ln \Phi\left(\frac{-((v_i - u_i) + \sigma_v^2/\sigma_u)}{\sigma_v}\right) + \frac{v_i - u_i}{\sigma_u} \right]$$

where $v_i - u_i = y_i - \alpha - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i$. Estimation proceeds (in principle) by specifying priors over $\boldsymbol{\Theta} = (\alpha, \boldsymbol{\beta}, \sigma_v, \sigma_u)$ and then deriving inferences from the joint posterior $p(\boldsymbol{\Theta}|\text{data})$. In general, the joint posterior for this model cannot be derived in closed form, so direct analysis is not feasible. Using Gibbs sampling and known conditional posteriors, it is possible use Markov chain Monte Carlo (MCMC) methods to sample from the marginal posteriors and use that device to learn about the parameters and inefficiencies. In particular, for the model parameters, we are interested in estimating $E[\boldsymbol{\Theta}|\text{data}]$ and var$[\boldsymbol{\Theta}|\text{data}]$ and perhaps even more fully characterizing the density $f(\boldsymbol{\Theta}|\text{data})$. In addition, we are interested in estimating the posterior mean inefficiencies $E[u_i|\text{data}]$ and in constructing confidence intervals (or their Bayesian counterparts, highest posterior density [HPD] intervals), again, conditioned on the data.[34] The preceding does not include features of $u_i$ in the estimation. One might, ex post, estimate $E[u_i|\text{data}]$ (see van den Broeck et al., 1994); however, it is more natural in this setting to include $(u_1, \ldots, u_N)$ with $\boldsymbol{\Theta}$ and estimate the conditional means with those of the other parameters. (The method is known as *data augmentation*; see Albert and Chib, 1993.) We develop the priors for the model components, then the conditional posteriors, and, finally, the Gibbs sampler for inference based on the joint posterior.

Priors for the parameters are specified as follows: A diffuse (improper, uninformative) prior for $(\alpha, \boldsymbol{\beta})$ would have $p(\alpha, \boldsymbol{\beta}) \propto 1$ over all of $\mathrm{R}^{K+1}$. Another typical approach is to specify a proper, but relatively diffuse prior, $p(\alpha, \boldsymbol{\beta}) \sim N[(\alpha^0, \boldsymbol{\beta}^0), \mathbf{W}]$ where $(\alpha^0, \boldsymbol{\beta}^0)$ is generally $(0, \mathbf{0})$ and $\mathbf{W}$ is large enough to avoid having an informative prior unduly influence the posterior.[35] For the stochastic elements of the frontier model, we specify $p(v_i|\sigma_v) \sim$ normal $(0, \sigma_v^2)$ and $p(u_i|\sigma_u) \sim$ exponential$(\sigma_u)$ independent of $v_i$. [Note that this is the departure point for extensions such as the gamma model (see discussion in Koop and Steel, 2001, and the analysis below) or a Dirichlet process (see the discussion of semiparametric estimators below).] For specifying the priors over the variance parameters, Koop and Steel (2001) note that

> the researcher can, of course, use any prior in an attempt to reflect his/her prior beliefs. However, a proper prior for $1/\sigma_v$ and $\sigma_u$ [maintaining our notation, not theirs] is advisable: Fernandez et al. (1997) show that Bayesian inference is not feasible (in the sense that

AQ: Confirm that box is not a missing character in line "Prior for ... $R^{k+1}$.

> the posterior is not well defined) under the usual improper priors for $1/\sigma_v$ and $\sigma_u$. (p. 523)

Priors for assumed independent variance parameters in stochastic frontier models are usually assumed to follow gamma distributions:

$$p(1/\sigma_v) \sim G(1/\sigma_v|\phi_v, P_v) = \frac{\phi_v^{P_v}}{\Gamma(P_v)} \exp\left[-\phi_v(1/\sigma_v)\right] (1/\sigma_v)^{P_v-1}, 1/\sigma_v \geq 0$$

The usual noninformative prior for $1/\sigma_v$ has $\phi_v = P_v = 0$ producing $p(1/\sigma_v) = (1/\sigma_v)^{-1}$, but use of this is precluded here. Different applications use different values—there is little uniformity as to how to choose these values, though in keeping with the aforementioned, values that produce more diffuse priors are preferred. For the parameter of the exponential inefficiency distribution, we likewise specify a gamma density:

$$p(\sigma_u) \sim G(\sigma_u|\phi_u, P_u) = \frac{\phi_v^{P_u}}{\Gamma(P_u)} \exp\left[-\phi_u\sigma_u\right] \sigma_u^{P_v-1}, \sigma_u \geq 0$$

Choice of the priors for the hyperparameters for the inefficiency distribution presents something of a difficulty, since, as above, a diffuse prior derails posterior inference. Many (possibly most) of the received applications have adopted a suggestion by Koop et al. (1997, and elsewhere). Setting $P_u = 1$ produces an exponential distribution for the prior over $\sigma_u$. We then set $\phi_u$ so that the prior median of efficiency, $\exp(-u_i)$, has a reasonable value. This involves setting $\phi_u = -\ln \tau^*$, where $\tau^*$ is the desired median. The value 0.875 is now conventional in the literature; hence, $\phi_u = 0.1335$. (Note that this is a fairly tight, quite informative prior. Koop et al. (1997, 2001) and subsequent researchers note that results seem not to be too dependent on this assumption.) The remaining detail is how to model the inefficiencies for the data augmentation. But that is already done in hierarchical fashion, since

$$p(u_i|\sigma_u) = \sigma_u \exp(-\sigma_u u_i).$$

We now have the joint prior for the parameters and $\mathbf{u} = (u_1, \ldots, u_N)$,

$$\begin{aligned} p(\boldsymbol{\Theta}, \mathbf{u}) &= p(\alpha, \boldsymbol{\beta}) p(1/\sigma_v) p(\sigma_u) p(u_1, \ldots, u_N|\sigma_u) \\ &= p(\alpha, \boldsymbol{\beta}) p(1/\sigma_v) p(\sigma_u) \prod_{i=1}^{N} p(u_i|\sigma_u) \end{aligned}$$

In order to draw inferences about the model components, we require information about the joint posterior

$$p(\boldsymbol{\Theta}, \mathbf{u}|\text{data}) \propto p(\boldsymbol{\Theta}, \mathbf{u}) L(\text{data}; \boldsymbol{\Theta}, \mathbf{u}).$$

The full posterior does not exist in closed form, so no analytic forms are available for the interesting characteristics, such as $E[\boldsymbol{\Theta}, \mathbf{u}|\text{data}]$. The strategy to be adopted is to infer these values by random sampling from the posterior

and, relying on the laws of large numbers, use statistics such as means and variances to infer the characteristics. However, no method exists for random sampling from this joint posterior. The Gibbs sampler provides the needed device. In broad terms, we desire to sample from

$$p(\boldsymbol{\Theta}, \mathbf{u}|\text{data}) = p[(\alpha, \boldsymbol{\beta}), 1/\sigma_v, \sigma_u, u_1, \ldots, u_N|\text{data}].$$

As noted, this is not feasible. However, it has been shown (see Casella and George, 1992) that the following strategy, Gibbs sampling, produces a set of samples from the marginal posteriors, which is all we need: We construct the conditional posteriors

$$\begin{aligned}
&p[(\alpha, \boldsymbol{\beta})|1/\sigma_v, \sigma_u, u_1, \ldots, u_N, \text{data}],\\
&p[1/\sigma_v|(\alpha, \boldsymbol{\beta}), \sigma_u, u_1, \ldots, u_N, \text{data}],\\
&p[\sigma_u|(\alpha, \boldsymbol{\beta}), 1/\sigma_v, u_1, \ldots, u_N, \text{data}],\\
&p[u_i|(\alpha, \boldsymbol{\beta}), 1/\sigma_v, \sigma_u|\text{data}], i = 1, \ldots, N.
\end{aligned}$$

Random samples from these, cycling in seriatim (an MCMC iteration), produces a set of random samples from the respective marginal posteriors. (The order of the draws does not matter.) The cycle begins at some value within the range of each variable. Many thousands of cycles are used, with the first several thousand discarded to eliminate the effects of the initial conditions—for example, received reported values range from 10,000 with the first 5,000 discarded to 75,000 with every fifth draw after 10,000 retained.

It remains to derive the conditional posteriors. For the stochastic frontier model, these are known: With all other values including $u_i, i = 1, \ldots, N$ known,

$$p[(\alpha, \boldsymbol{\beta})|1/\sigma_v, \sigma_u, u_1, \ldots, u_N, \text{data}] = p(\alpha, \boldsymbol{\beta}) \times N[(a, \mathbf{b}), \sigma_v^2 \mathbf{A}],$$

where $(a, \mathbf{b})$ are the least squares coefficients in the linear regression of $y_i + u_i$ on a constant and $\mathbf{x}_i$, and $\mathbf{A}$ is the inverse of the usual second moment matrix for $[1, \mathbf{x}_i]$. Recall $p(\alpha, \boldsymbol{\beta}) = 1$ in our application. Random draws from the multivariate normal distribution are simple to draw; they can be built up from primitive draws from the standard normal with straightforward calculations (see Greene, 2003a, p. 922). The conditional densities for the variance parameters are a bit more complicated. For the symmetric distribution,

$$p[1/\sigma_v|(\alpha, \boldsymbol{\beta}), \sigma_u, u_1, \ldots, u_N, \text{data}] = \gamma(f, P^*),$$

where $f = \phi_v + \frac{1}{2N}\sum_{i=1}^{N}(y_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^2$ and $P^* = P_v + N/2$. For the inefficiency parameter,

$$p[\sigma_u|(\alpha, \boldsymbol{\beta}), 1/\sigma_v, u_1, \ldots, u_N, \text{data}] = \gamma\left(\frac{1}{N}\sum_{i=1}^{N} u_i - \ln \tau^*, N + 1\right).$$

Sampling from the gamma distribution is less straightforward than from the normal but can be done with published software, such as IMSL Libraries (Absoft, 2005). Finally, for the data augmentation step, we have

$$p[u_i|(\alpha,\beta), 1/\sigma_v, \sigma_u, \text{data}) = N^+[-(y_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) + \sigma_v^2/\sigma_u, \sigma_v^2],$$

where $N^+[.]$ denotes the truncated normal distribution.[36] Sampling from a truncated normal distribution with a given underlying mean and standard deviation is also straightforward. Some authors (e.g., Tsionas, 2002) suggest acceptance/rejection—draw an observation, and either accept it if it is positive, or reject it and draw another. A simpler and smoother method requiring but a single draw is based on the inverse probability transform: $u_{i,r} = \mu + \sigma\Phi^{-1}[F_{i,r} + (1 - F_{i,r})\Phi(-\mu/\sigma)]$, where the subscript $i$, $r$ denotes the $r$th draw for observation $i$, $\mu$ and $\sigma$ are the mean and standard deviation noted above, $\Phi^{-1}(.)$ is the inverse function of the standard normal CDF, and $F_{i,r}$ is a single standard uniform, $U[0, 1]$ draw.

These equations define the Gibbs sampler, which can be used to produce samples from the desired marginal posterior distributions. Thus, after the iterations, the simple means and variances of the draws produce estimates of the means and variances of the conditional distributions, $f[(\alpha, \boldsymbol{\beta})|\text{data}]$, $f(1/\sigma_v|\text{data})$, $f(\sigma_u|\text{data})$, and $f(u_i|\text{data})$. (Note, again, that the last of these is not an estimator of $u_i$; it is an estimator of $\mathrm{E}[u_i|\text{data}]$. No amount of data, manipulated in Bayesian or classical fashion, produces a convergent estimator of $u_i$; we only estimate the mean of the conditional distribution.)

### 2.4.6 The normal–gamma model

Stevenson (1980) and Greene (1980a, 1980b) also proposed results for the gamma/normal distribution. In this case,

$$f_u(u_i) = \frac{\sigma_u^{-P}}{\Gamma(P)} \exp(-u_i/\sigma_u) u_i^{P-1}, u_i \geq 0, P > 0.$$

Note that the exponential results if $P = 1$. Unlike the case of the deterministic gamma frontier model, this model only requires $P$ to be positive. The convolution with $v_i$ makes the resulting distribution of $\varepsilon_i$ regular for all positive values of $P$. Stevenson limited his attention to the Erlang form (integer values of $P$, 1.0, and 2.0), which greatly restricts the model. Beckers and Hammond (1987) first formally derived the log-likelihood for the convolution of a normal and a gamma variate. The resulting functional form was intractable, however. Greene (1990) produced an alternative formulation that highlights the relationship of the gamma model to the exponential model considered above. The

log-likelihood function with the normal–gamma mixture is

$$\text{Ln } L(\alpha, \beta, \sigma_v, \sigma_u) = \sum_{i=1}^{N} \left[ \begin{array}{l} -P \ln \sigma_u - \ln \Gamma(P) + \ln q(P-1, \varepsilon_i) \\ +\frac{1}{2}\left(\frac{\sigma_v}{\sigma_u}\right)^2 + \ln \Phi\left(\frac{-\varepsilon_i + \sigma_v^2/\sigma_u}{\sigma_v}\right) + \frac{\varepsilon_i}{\sigma_u} \end{array} \right],$$

where
$q(r, \varepsilon_i) = E\left[z^r | z > 0, \varepsilon_i\right], z \sim N[-\varepsilon_i + \sigma_v^2/\sigma_u, \sigma_v^2]$.

The log-likelihood function can be written

$$\text{Ln } L(\alpha, \beta, \sigma_v, \sigma_u) = \ln L_{\text{Exponential}} + \sum_{i=1}^{N} [-(P-1) \ln \sigma_u - \ln \Gamma(P) + \ln q(P-1, \varepsilon_i)].$$

The $q(r, \varepsilon)$ function is a (possibly) fractional moment of the truncated normal distribution.[37] The additional terms drop out if $P$ equals 1, producing the exponential model.

The gamma formulation has some appeal in that it allows for different distributions of inefficiency. Figure 2.6 suggests the possibilities. The heaviest plot in figure 2.6 shows the distribution with $P = 0.8$. When $P < 1$, the distribution only asymptotes to the vertical axis, which implies that there is large mass near zero. The middle plot has $P = 1$, which is the exponential distribution shown in figure 2.5. The lower plot shows that, with $P > 1$, the distribution can be pulled away from zero, which is likely

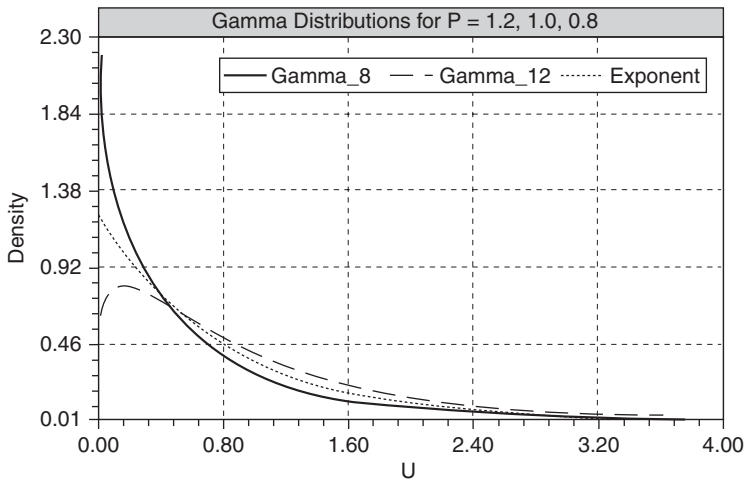


**Figure 2.6.** Gamma Distributions

to be a more reasonable characterization of inefficiency at least for some applications.

Greene's formulation of the gamma model brought some substantial differences from the half-normal specification in an empirical application.[38] However, the greatly increased complexity of the procedure has somewhat inhibited its application.[39] van den Broeck et al. (1994) and Koop et al. (1995) have taken a Bayesian approach to the specification of the inefficiency term in the stochastic frontier model. The former study considers several prior distributions for $u_i$ including half- and truncated-normal, exponential, and the Erlang form of Greene's normal–gamma model.[40] Several other Bayesian applications (e.g., Tsionas, 2002; Huang, 2004) have employed the gamma model in stochastic frontier formulations—it provides a comfortable fit for the inverted gamma conjugate priors for variance parameters. Ritter and Simar (1997), however, have analyzed this model extensively and expressed considerable skepticism about its usability for classical formulation. They found that the parameters are very weakly identified, and estimation is exceedingly difficult. Accurate computation of the fractional moments is extremely difficult. Note, however, that the model that Ritter and Simar (1997) focus on has only a constant term, so their results may lack generality—they do stand as a caveat for researchers nonetheless. Greene (2000, 2003a, 2003b) proposes a more general approach than Ritter and Simar's, based on MSL that seems largely to overcome the previous problems of computation of the log-likelihood and its derivatives.

#### 2.4.6.1 Classical estimation of the normal–gamma model

Several recent studies have resurrected the normal–gamma model. Greene (2003a) has produced an alternative approach to computing the complex part of the log-likelihood function, the expectations from the truncated normal distribution, using Monte Carlo simulation, rather than attempting to approximate the integrals directly. The method appears to produce more satisfactory results. The obstacle to estimation is accurate computation of $q(r, \varepsilon_i) = E[z^r | z > 0]$ where $z \sim N[\mu_i, \sigma_v^2], \mu_i = -(\varepsilon_i + \sigma_v^2/\sigma_u)$. Since it is an expectation, and an otherwise straightforward function, the function can be consistently (pointwise) estimated with $\hat{q}(r, \varepsilon_i) = (1/Q) \sum_{q=1}^{Q} z_{iq}^r$, where $z_{iq}$ is a random draw from the indicated truncated normal distribution. The MSL estimator then replaces $q(r, \varepsilon_i)$ with $\hat{q}(r, \varepsilon_i)$. The remaining complication is how to obtain the random draws. Acceptance/rejection of a sample of draws from the untruncated normal population is unacceptable, since the resulting function is no longer smooth and continuous (different observations will be rejected depending on the parameters), it will take huge numbers of draws, and it will be very inaccurate in the tails. The alternative is to use the inverse probability transform, which translates random draws one for one. The strategy is implemented by using the generic formula for sampling from the truncated normal

distribution,

$$z_{iq} = \mu_i + \sigma\Phi^{-1}[(1 - F_q)P_L + F_q],$$

where $\varepsilon_i = y_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \mu_i = -\varepsilon_i - \sigma_v^2/\sigma_u, \sigma = \sigma_v$, and $P_L = \Phi(-\mu_i/\sigma)$, and $F_q$ is a draw from the continuous uniform (0, 1) distribution. Combining all terms, then,

$$\begin{aligned}&\operatorname{Ln} L_S(\alpha, \boldsymbol{\beta}, \sigma_v, \sigma_u)\\ &= \sum_{i=1}^{N}\left[\begin{array}{l} -P\ln\sigma_u - \ln\Gamma(P) + \frac{1}{2}\left(\frac{\sigma_v}{\sigma_u}\right)^2 + \ln\Phi\left(\frac{-\varepsilon_i + \sigma_v^2/\sigma_u}{\sigma_v}\right) + \frac{\varepsilon_i}{\sigma_u}\\ +\ln\left\{\frac{1}{Q}\sum_{q=1}^{Q}\left(\mu_i + \sigma_v\Phi^{-1}(F_{iq} + (1 - F_{iq})\Phi(-\mu_i/\sigma_v)\right)^{P-1}\right\}\end{array}\right].\end{aligned}$$

As complicated as it is, this form is vastly simpler than the Pochammer function invoked by Beckers and Hammond (1987) or the direct integration in Greene (1990). The function and its derivatives are smooth and continuous in all the parameters of the model. Further details appear in Greene (2003b, 2004a). Vitaliano (2003) is a recent application.

#### 2.4.6.2 Bayesian estimation of the normal–gamma model

Owing to its flexibility and its natural similarity to familiar forms of priors, the gamma model has also attracted researchers employing Bayesian methods to estimate stochastic frontier models. Tsionas (2002) begins with a normal–exponential model and an assumed panel-data setting. Each unit has its own parameter vector, $\boldsymbol{\beta}_i$, which is assumed to be generated by a prior normal density, $N[\boldsymbol{\beta}^0, \boldsymbol{\Omega}]$. Posterior means are derived for all the production parameters using the MCMC simulation method that is now standard in Bayesian applications. Finally, the posterior distribution for the inefficiencies, $u_{it}$ is obtained as a truncated normal variable with a specified mean and variance that is a function of the other parameters in his model. Thus, estimates of $u_{it}$ are obtained after obtaining posterior estimates of the other parameters.[41] (The estimation of $u_{it}$ and the use of panel data are both subjects of later sections of this chapter, but the use of the normal–gamma model remains somewhat out of the mainstream of empirical applications, so it seems appropriate to continue the thread of the discussion here rather than later, because this model does not figure prominently in most of the rest of the discussion.) Interestingly enough, after developing the model for panel-data applications, Tsionas (2002) applied it to a cross section—the Christensen and Greene (1976) electricity data. It seems likely that some of the fairly extreme empirical results in his paper were a consequence of stretching the panel-data estimator to samples of one in a cross section—his results appear to imply an average efficiency in the sample of more than 99%, which is considerably at odds with earlier findings with the same data set.) Tsionas proceeded to extend his model to

half-normal and Erlang (gamma with $P = 1, 2, 3$) distributions, employing similar methodologies in each case.

Van den Broeck et al. (1994) and Koop et al. (1995) have also examined the normal–Erlang model using Bayesian MCMC techniques. Surprisingly, in an *earlier* paper, Tsionas (2000b), again employing MCMC techniques, examined the implications of a noninteger value of $P$ in the normal–gamma model. Suggestions elsewhere notwithstanding, he found that variation to noninteger values of $P$, even within a fairly narrow range, does produce substantive differences in the appearance of the inefficiency distribution. He continues to examine the model with various values of $P$. In an indicator of the complexity of the estimation problem, in his analysis, it becomes necessary to fix one of the other model parameters at an assumed value to proceed with estimation. In their Capital Asset Pricing Model (CAPM) study of mutual fund performance, Annaert et al. (2001) also fit the Erlang model with $P = 1, 2$, and 3 and then probabilistically pooled the three sets of estimates. With $P$ fixed in each case, the estimator itself is easily fit using the straightforward MCMC methods mentioned above. In sum, the normal–gamma model with a free shape parameter has posed an ongoing challenge in the Bayesian literature, but one that has attracted a fair amount of attention. Ultimately, the flexibility of the two-parameter distribution and the variety of shapes that it can accommodate do have an appeal. (One might surmise that the convenience of the conjugate prior with the flexibility of the two-parameter gamma model make it an irresistible target in this literature.) In the most recent attack on this vexing estimation problem, Huang (2004) develops a full likelihood-based Bayesian estimator for the normal–gamma model without the Erlang restriction. His results on inefficiency estimates are essentially the same as Tsionas's; in his full model with parameter heterogeneity, the modal efficiency is roughly 0.99 (Huang's figure 4). The estimates presented in Huang's table 1 suggest that the overall distribution of inefficiency is roughly exponential with a mean and standard deviation of $1/77.4337 = 0.0129$. Both of these sets of implausible results are considerably at odds with other evidence of inefficiency in the Christensen and Greene data.[42] Finally, Griffin and Steel (2004) propose a Dirichlet (semiparametric) specification for the inefficiencies in a semiparametric formulation of a Bayesian model. In passing, they also fit the normal–gamma (fully parametric) model. The application is based on Koop et al.'s (1997) hospital data, so we cannot compare the results to the foregoing. They do (apparently) find that for most of their sample the normal–gamma model tracks the semiparametric model fairly well, and far better than the normal–exponential model, which might be expected. Migon and Medici (2001) also propose methodology for the normal–gamma model but do not use it in their applications. (Unlike most other studies, they ultimately gravitated to a normal–lognormal model.)

In summary, then, it would appear that Greene (2003b) and Tsionas (2002)/Huang (2004) have reported considerable progress in the 20-plus year development of this strand of literature. Both estimation strategies

based on simulation—the former in the classical tradition, the latter in the Bayesian paradigm—appear to be reasonably (not extremely) straightforward to implement.[43] What remains unsettled, at least as a caveat, is the Ritter and Simar (1997) argument that the model is difficult to identify. The applications seem to suggest otherwise, but extensive analysis remains to be done.

There have been numerous Bayesian applications in the stochastic frontier literature. A significant proportion of them are listed above, and nearly all of the remainder (that I have located) appear at one point or another below.[44] As in most applications, since the specifications are stringent in their specification of noninformative (diffuse) priors, the results usually differ marginally, if at all, from MLEs derived from the classical approach.[45] There are, however, some aspects of Bayesian estimation in the stochastic frontier literature that are worthy of note. First, there are now Bayesian applications to problems that have not received much attention in the classical literature, for example, O'Donnell and Coelli's (2005) application in which they imposed the theoretical curvature conditions on a translog distance function. The estimation of technical or cost inefficiency poses an unusual challenge for Bayesian estimators, however. Since estimates of inefficiency (technical or cost) are individual observation specific, it is not possible to obtain them without assuming an informative prior. Thus, Koop et al. (1994), Tsionas (2002), and Huang (2004) all assume a gamma prior for $\ln u_i$ with a known mean (and variance). Obviously, the results are sensitive to the assumption. The technique of data augmentation (Albert and Chib, 1993) is often used as a means to the end of posterior parameter mean estimation in models with missing data (e.g., the probit model). The estimates of the missing data values are generally of no intrinsic interest and are not analyzed at any length in the posterior analysis. The same technique is used in estimating $u_i$ in stochastic frontier models, but in this setting, the augmented data are not a means to an end—they are the end. However, it is here that it is necessary to assume a fairly strongly informative prior in order to have a tractable posterior with finite variance. I return to this issue in some detail below.

In sum, some of the Bayesian applications merely demonstrate the existence of counterparts to classical estimators. Given diffuse priors, this produces little more than an alternative method (MCMC) of maximizing the likelihood function and then calling the new "estimate" something with a different name. (See Kim and Schmidt, 2000, for some evidence on this point.) But, at the same time, innovative applications that extend the model, such as Tsionas's (2003) dynamic model and Atkinson and Dorfman's (2005) distance function model, have begun to appear, as well. As of this writing, this strand of the literature remains a significant minority. I revisit it at various points below, but my treatment, like the literature it surveys, focuses primarily on classical, ML-based applications.

### 2.4.7 The truncated-normal model

Stevenson (1980) argued that the zero mean assumed in the Aigner et al. (1977) model was an unnecessary restriction. He produced results for a *truncated* as opposed to *half-normal* distribution. That is, the one-sided error term, $u_i$ is obtained by truncating at zero the distribution of a variable with possibly nonzero mean. The complete parameterization is

$$v_i \sim N[0, \sigma_v^2],$$
$$U_i \sim N[\mu, \sigma_u^2], u_i = |U_i|.$$

For convenience, let us use the parameterizations given above for $\lambda$ and $\sigma$. Then, the log-likelihood is

$$\text{Ln } L(\alpha, \beta, \sigma, \lambda, \mu) = -N\left[\ln\sigma + \frac{1}{2}\ln 2\pi + \ln\Phi(\mu/\sigma_u)\right] + \sum_{i=1}^{N}\left[-\frac{1}{2}\left(\frac{\varepsilon_i + \mu}{\sigma}\right)^2 + \ln\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\varepsilon_i\lambda}{\sigma}\right)\right],$$

where $\sigma_u = \lambda\sigma/\sqrt{1+\lambda^2}$ (a derivation appears in Kumbhakar and Lovell, 2000). Starting values for the iterations in the stochastic frontier models are typically obtained by manipulating the results of OLS to obtain method-of-moments estimators for the parameters of the underlying distribution. There does not appear to be a convenient method-of-moments estimator for the mean of the truncated normal distribution. But MLE presents no unusual difficulty. The obvious starting value for the iterations would be the estimates for a half-normal model and zero for $\mu$. The benefit of this additional level of generality is the relaxation of a possibly erroneous restriction. A cost appears to be that the log-likelihood is sometimes ill-behaved when $\mu$ is unrestricted. As such, estimation of a nonzero $\mu$ often inflates the standard errors of the other parameter estimators considerably, sometimes attends extreme values of the other parameters, and quite frequently impedes or prevents convergence of the iterations. It is also unclear how the restriction of $\mu$ to zero, as is usually done, would affect efficiency estimates. The Bayesian applications of this model (e.g., Tsionas, 2001a; Holloway et al., 2005) have apparently encountered less difficulty in estimation of this model.

As explored in section 2.6, the parameters of the underlying distribution of $u_i$ provide a mechanism for introducing heterogeneity into the distribution of inefficiency. The mean of the distribution (or the variance or both) could depend on factors such as industry, location, and capital vintage. One way such factors might be introduced into the model could be to use

$$\mu_i = \mu_0 + \boldsymbol{\mu}_1^{\mathrm{T}}\mathbf{z}_i,$$

where $\mathbf{z}_i$ is any variables that should appear in this part of the model. As noted, we revisit this possibility further below.

### 2.4.8 Estimation by COLS method-of-moments estimators

The parameters of the stochastic frontier model can be estimated using the second and third central moments of the OLS residuals, $m_2$ and $m_3$. For the half-normal model, the moment equations are

$$m_2 = \left[\frac{\pi - 2}{\pi}\right]\sigma_u^2 + \sigma_v^2,$$
$$m_3 = \sqrt{\frac{2}{\pi}}\left[1 - \left(\frac{4}{\pi}\right)\right]\sigma_u^3.$$

(Note that $m_3$ is negative, since the offset in $\varepsilon_i$ by $u_i$ is negative.) Thus, $\sigma_u$ and $\sigma_v$ are easily estimable. Since $E[u_i] = (2/\pi)^{1/2}\sigma_u$, the adjustment of the OLS constant term is $\hat{\alpha} = a + \hat{\sigma}_u\sqrt{2/\pi}$. These MOLS estimators are consistent, but inefficient in comparison to the MLEs. The degree to which they are inefficient remains to be determined, but it is a moot point, since with current software, full MLE is no more difficult than least squares.

Waldman (1982) has pointed out an intriguing quirk in the half-normal model. Normally, there are two roots of the log-likelihood function for the stochastic frontier model: one at the OLS estimates and another at the MLE. In theory, the distribution of the compound disturbance is skewed to the left. But, if the model is badly specified, the OLS residuals can be skewed in the opposite direction. In this instance, the OLS results are the MLEs, and consequently, one must estimate the one-sided terms as 0.0.[46] (Note that if this occurs, the MOLS estimate of $\sigma$ is undefined.) One might view this as a built-in diagnostic, since the phenomenon is likely to arise in a badly specified model or in an inappropriate application. This "failure"—I use the term advisedly here, since analysts might differ on whether the estimation tools or the analyst has failed—occurs relatively frequently. Coelli's (1995) formulation may be more convenient in this regard (see note 26). He suggests the moment estimators

$$\hat{\sigma}^2 = m_2 + \left(\frac{2}{\pi}\right)\left[\sqrt{\frac{\pi}{2}}\left(\frac{\pi}{\pi - 4}\right)m_{23}\right]^{\frac{2}{3}},$$
$$\hat{\gamma} = \left(\frac{1}{\hat{\sigma}^2}\right)\left[\sqrt{\frac{\pi}{2}}\left(\frac{\pi}{\pi - 4}\right)m_3\right]^{\frac{2}{3}},$$
$$\hat{\alpha} = a + \sqrt{\frac{2\hat{\gamma}\hat{\sigma}^2}{2}}.$$

As before, the "wrong sign" on $m_3$ can derail estimation of $\gamma$, but in this instance, a convenient place to begin is with some small value; Coelli suggests

0.05. As noted above, there is no obvious method-of-moments estimator for $\mu$ in Stevenson's truncated-normal model.

The MOLS estimators for the exponential model are based on the moment equations $m_2 = \sigma_v^2 + \sigma_u^2$ and $m_3 = -2\sigma_u^3$. Thus,

$$\hat{\sigma}_u = [-m_3/2]^{\frac{1}{3}}, \hat{\sigma}_v^2 = m_2 - \hat{\sigma}_u^2, \hat{\alpha} = a + \hat{\sigma}_u.$$

For the gamma model, the MOLS estimators are

$$\hat{\sigma}_u = -(m_4 - 3m_2^2)/(3m_3), \hat{P} = -m_3/(2\hat{\sigma}_u^3), \hat{\sigma}_v^2 = m_2 - \hat{P}\hat{\sigma}_u^2, \hat{\alpha} = a + \hat{P}\hat{\sigma}_u.$$

Any of these can be used to obtain a full set of estimates for the stochastic frontier model parameters. They are all consistent. Thereafter, estimates of the efficiency distributions or of the individual coefficients, $-u_i$ or $\text{TE}_i$, can be computed just by adjusting the OLS residuals. There is a question of the *statistical* efficiency of these estimators. One specific result is given in Greene (1980a) for the gamma-distributed, deterministic frontier model, namely, that the ratio of the true variance of the MLE of any of the slope coefficients in the model to its OLS counterpart is $(P-2)/P$. Thus, the greater the asymmetry of the distribution—the gamma density tends to symmetry as $P$ increases—the greater is efficiency gain to using MLE (see Deprins and Simar, 1985, for further results). Of course, efficient estimation of the technical parameters is not necessarily the point of this exercise. Indeed, for many purposes, consistency is all that is desired. As noted, estimation of all of these models is fairly routine with contemporary software. The preceding are likely to be more useful for obtaining starting values for the iterations than as estimators in their own right.

### 2.4.9 Other specifications for stochastic frontier models

A number of other candidates have been proposed for the parametric forms of the stochastic frontier model. An early study by Lee (1983) proposed a four-parameter Pearson family of distributions for the purpose of testing the distributional assumptions of the model—the Pearson family nests a large number of familiar distributions. The model proved much too cumbersome for general usage, but it does suggest the possibility of alternatives to the familiar paradigm of normality coupled with a limited range of one-sided distributions for $u_i$. This section surveys a few of the alternative distributions that have been proposed for the stochastic frontier model.

The question of how to model inefficiency in a data set that spans several time periods is a major point in the analysis of panel data. In particular, researchers differ—and the data are inconsistent—on whether it is reasonable to model inefficiency as a time-invariant, firm-specific effect or as an effect that varies freely and randomly over time, or whether some intermediate formulation, in which $u_{i,t}$ (firm $i$ at time $t$) evolves systematically, is appropriate. This subject is revisited at length in section 2.7. Note at this point, however, a

proposal by Tsionas (2003) that could be used to analyze this issue, at least in part. He suggests the dynamic model

$$\ln u_{i,t}|\mathbf{z}_{it}, \gamma, \rho, \omega, u_{i,t-1} \sim N\left[\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}_{it} + \rho \ln u_{i,t-1}, \omega^2\right], \quad t = 2, \ldots, T,$$

$$\ln u_{i,1}|\mathbf{z}_{i1}, \gamma_1, \omega_1 \sim N[\boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{z}_{i1}, \omega_1^2],$$

where $\mathbf{z}_{i,t}$ is a vector of exogenous effects (not the inputs). The startup process (initial condition) is allowed to be different from the process governing the evolution of the inefficiency. Tsionas (2003) applies the technique to Bayesian estimation of a cost frontier estimated for a sample of 128 U.S. banks over 12 years. A multiple-output translog function is estimated. The estimated posterior mean of $\rho$ is 0.908, suggesting that, to some approximation, the measured inefficiency in his sample is close to constant over time. Note that this proposal employs a lognormal density for the inefficiency—a specification that has been used quite infrequently (see, e.g., Migon and Medici, 2001; Deprins and Simar, 1989b).

#### 2.4.9.1 Other parametric models

Migon and Medici (2001) also use Bayesian methods to estimate a stochastic frontier model with lognormal inefficiencies. Estimation is straightforward using the MCMC methods they employ. It would be more difficult to replicate this with orthodox classical methods, since forming the density for a normal minus a lognormal is an unsolved problem. The method of Misra and Greene and Misra (2003), shown below, however, which would approach the problem in essentially the same fashion as the Bayesian estimator, could easily be adapted to a lognormal distribution. The normal–lognormal model remains to be explored in this literature. As (possibly) a two-parameter density that resembles the attractive gamma model, I would surmise that this specification offers some interesting potential. Tsionas and Greene (2003) showed how the Bayesian approach outlined above for the normal–gamma model could be adapted to other functional forms. Motivated by the possibility that ordinary outliers in the data might distort the estimated model and ultimately end up expanding the range of variation of $u_i$ in the estimated model, they proposed a Student's $t$ for the symmetric distribution ($v_i$), that is, a distribution with much thicker tails than the normal. In their discussion of the MCMC procedure, they suggested that formulation of a tractable posterior is the only obstacle to any other distribution. (The half-normal and exponential were demonstrated, as well.) Whether other distributions would provide any benefit, or even substantively change the results, remains to be seen. [An application that considers the lognormal and Weibull distributions in addition to those considered here is Deprins and Simar (1989b).]

A similar consideration underlies the proposal by Greene and Misra (2003), who essentially followed Tsionas and Greene's (2003) suggestion, in a

classical estimator. Recall that the density for the observed data that underlies the log-likelihood is obtained as follows: First, $y_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + v_i - u_i$ and $\varepsilon_i = y_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i = v_i - u_i$. A symmetric density is assumed for $v_i$ and a one-sided one for $u_i$. Then, the unconditional density that enters the likelihood function is

$$f_\varepsilon(\varepsilon_i|\boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{x}_i) = \int_0^\infty f_v(\varepsilon_i + u_i) f_u(u_i) du_i,$$

where $\boldsymbol{\delta}$ is any parameters in the model other than $\alpha$ and $\boldsymbol{\beta}$, such as $\sigma_u$ and $\sigma_v$ in the half-normal and exponential models. The normal–half-normal and normal–exponential models are derived by obtaining a closed form for this integral. Since there is no closed form for the normal–gamma model, the relevant part of the log-likelihood is approximated by simulation. As observed at several pointsabove, the integral above is of the form of an expectation. In principle, it can be accurately approximated by simulation and averaging a number of draws from the appropriate underlying population. In order to apply the principle, the specification requires a distribution for $u_i$ from which a random sample of draws can be obtained, and an explicit specification for the density of $v_i$. With these in place, a generic formulation of the simulated log-likelihood for the stochastic frontier model would be

$$\log L_S(\alpha, \beta, \delta|\text{data}) = \sum_{i=1}^{N} \log \frac{1}{Q} \sum_{q=1}^{Q} f_v[y_i - \alpha - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + u_i, \delta].$$

This function is then maximized with respect to the underlying parameters. Like the normal–gamma model discussed above, it is smooth and continuous in the parameters. To take a specific example, the following shows an alternative way to estimate the normal–exponential model. The density (PDF) and CDF for the one-sided $u_i$ are

$$f_u(u_i) = (1/\sigma_u)\exp(-u_i/\sigma_u), F(u_i) = 1 - \exp(-u_i/\sigma_u), u_i \geq 0, \sigma_u > 0.$$

Inverting $F(u_i)$ for $u_i$ reveals the strategy for generating random draws on $u_i$:

$$u_{ir} = -\sigma_u \ln(1 - F_{ir}),$$

where $F_{ir}$ is a random draw from the standard uniform distribution, $U[0, 1]$, which one can do with any modern econometrics package. (For simplicity, the draw may simply be $F_{ir}$, since $1 - F_{ir}$ is also from the $U[0, 1]$ population.) The symmetric density is the normal distribution, so the simulated log-likelihood is

$$\text{Ln } L_S(\alpha, \beta, \sigma_v, \sigma_u|\text{data}) = \sum_{i=1}^{N} \ln \frac{1}{R} \sum_{r=1}^{R} \frac{1}{\sigma_v} \phi\left[\frac{y_i - \alpha - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + (-\sigma_u \log F_{ir})}{\sigma_v}\right]$$

This function and its derivatives are smooth and continuous in the parameters and can be maximized by conventional means (assuming one is able to fix the set of random draws—the same set of $R$ draws must be used each time the function is evaluated). The derivatives of this log-likelihood are as follows: For convenience, let the argument of the normal density be denoted $a_{ir} = y_i - \alpha - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \sigma_u \ln F_{ir}$, so that the bracketed function above is just $\phi(a_{ir}/\sigma_v)$. Let $\boldsymbol{\theta}$ denote the parameter vector $(\alpha, \boldsymbol{\beta}', \sigma_u)'$. Then,

$$\ln L_S(\theta, \sigma_v|\text{data}) = \sum_{i=1}^{N} \ln \frac{1}{R} \sum_{r=1}^{R} \frac{1}{\sigma_v} \phi\left(\frac{a_{ir}}{\sigma_v}\right),$$

$$\frac{\partial \ln L_S(\boldsymbol{\theta}, \sigma_v|\text{data})}{\partial \theta} = \sum_{i=1}^{N} \frac{\frac{1}{R}\sum_{r=1}^{R} \frac{1}{\sigma_v}\left[\left(\frac{a_{ir}}{\sigma_v}\right)\right] \phi\left(\frac{a_{ir}}{\sigma_v}\right) \frac{1}{\sigma_v} \begin{bmatrix} 1 \\ \mathbf{x}_i \\ \ln F_{ir} \end{bmatrix}}{\frac{1}{R}\sum_{r=1}^{R} \frac{1}{\sigma_v} \phi\left(\frac{a_{ir}}{\sigma_v}\right)},$$

$$\frac{\partial \ln L_S(\theta, \sigma_v|\text{data})}{\partial \sigma_v} = \sum_{i=1}^{N} \frac{\frac{1}{R}\sum_{r=1}^{R} \frac{1}{\sigma_v^2} \phi\left(\frac{a_{ir}}{\sigma_v}\right) \left[\left(\frac{a_{ir}}{\sigma_v}\right)^2 - 1\right]}{\frac{1}{R}\sum_{r=1}^{R} \frac{1}{\sigma_v} \phi\left(\frac{a_{ir}}{\sigma_v}\right)},$$

Simultaneous equation of the two gradients to zero produces the maximum simulated likelihood (MSL) estimators. Either the (moderately complicated) Hessian or the BHHH estimator can be used to estimate the asymptotic covariance matrix for the estimator.

In principle, this approach can be used with any pair of densities, $f_v(v_i)$, that has a tractable functional form and $f_u(u_i)$ from which a random sample of draws can be simulated. Greene and Misra (2003) worked out several pairs. Certainly there are others. (I noted the lognormal above, which was not considered by the authors.) There are two real questions yet to be considered in this setting: First, again, does the distribution really matter in terms of the estimates of $u_i$? (How those are computed remains to be derived. This is revisited below.) Second, in any event, it is unclear how one can choose among the various models. Likelihood ratio tests are inappropriate, because the models are not nested. Vuong's (1989) test for nonnested models probably is appropriate, but it is for pairs of competing models, and there may be more than two here.

Researchers in a number of areas (e.g., Cameron et al., 2004) in their analysis of health care) have suggested the copula method of formalizing bivariate relationships when the marginal distributions are known but the joint distribution remains to be determined. For the stochastic frontier model, the tool suggests a means to consider the possibility of specifying a model in which the inefficiency, $u_i$, might be correlated with the firm-specific idiosyncratic noise, $v_i$. The underlying economics may require a bit of investigation, but econometrically, this possibility points toward relaxing yet one more restriction in the stochastic frontier model. Smith (2004) has used the method to analyze

(yet again) Christensen and Greene's (1976) electricity generation cost data and the panel data on airlines listed in Greene (1997). Interestingly enough, the copula model applied to the electricity data produce some fairly substantial changes compared to the standard normal–half-normal model. The chi-squared test with one degree of freedom for the copula model against the null of the standard model is 5.32, while the 95% critical value is 3.84. As noted, the economic interpretation of the richer model specification needs to be solidified, but the empirical results suggest an intriguing possibility. This is a nascent literature, so I have no further empirical results to report.

#### 2.4.9.2 Semiparametric models

The stochastic frontier model considered thus far is fully parameterize—the production model is specified in full, and the full distributions of $v_i$ and $u_i$ are known up to the specific values of the parameters, which are estimated using either classical or Bayesian methods. Ongoing research has sought flexible specifications of the technology model and the distributions involved that relax the assumptions of the model. There have been many explorations in the production model and extensions of the distributions. The normal–gamma model, with its richer specification, for example, represents one such model extension. In addition, there have been numerous proposals to move away from specific distributional assumptions. The semiparametric approaches described here retain the essential framework of the stochastic frontier but relax the assumption of a specific distribution for $v_i$ or $u_i$, or both.

Fan, Li, and Weersink (1996) suggested modifying the production model:

$$y_i = g(x_i) + v_i - u_i,$$

where $g(x_i)$ remains to be specified. Since nearly all applications of the stochastic frontier model employ either the Cobb-Douglas or translog form, a semiparametric specification here represents relaxing one assumption restriction in the model, though it retains the fundamental stochastic (in their case, normal–exponential) specification. Huang and Fu (1999) continued this line of inquiry. In a similar analysis, Koop et al. (1994) specify a "semi-nonparametric" stochastic frontier cost function of the form

$$\ln C_i = H(y_i) + \ln c(\mathbf{w}_i) + v_i + u_i,$$

where $H(y)$ is specified semiparametrically, in terms of polynomials in the log of output and $\ln c(\mathbf{w})$ is a Cobb-Douglas or translog function of the input prices.

In a series of studies, Park and Simar (1994), Park et al. (1998), Adams et al. (1999), Sickles et al. (2002), and Sickles (2005) have explored the implications of a variety of distributional assumptions on estimation in the

panel-data model

$$y_{it} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + \alpha_i + \varepsilon_{it}.$$

Absent their much more general assumptions, this is a conventional fixed- or random-effects linear regression model. The various extensions involve different assumptions about $\varepsilon_{it}$, the relationships between $\alpha_i$ and $\mathbf{x}_{it}$, and so on. The stochastic frontier aspect of the model is embodied in the use of $\alpha_i - \max_j(\alpha_j)$ in the estimation of inefficiency, in the fashion of the deterministic frontier models discussed above. Instrumental variable, ML, generalized least squares (GLS), and generalized method of moments (GMM) estimation methods all appear in the different treatments. This body of results extends the assumptions behind the deterministic frontier models in a variety of directions but is not directed at the stochastic frontier model. The semiparametric nature of the model relates to the very loose specification of the effects and their relationship to the frontier. Section 2.7 returns to the discussion of panel models.

One way to extend the normal–half-normal stochastic frontier model (or others) with respect to the distribution of $v_i$ is the finite mixture approach suggested by Tsionas and Greene (2003). I return to the methodological aspects of the finite mixture model below; for the moment, let us examine only the model formulation. The frontier model is formulated in terms of $J$ "classes" so that, within a particular class,

$$f_\varepsilon(\varepsilon_i|\ \text{class} = j) = \frac{2}{\sqrt{2\pi(\sigma_u^2 + \sigma_{vj}^2)}}\left[\Phi\left(\frac{-\varepsilon_i(\sigma_u/\sigma_{vj})}{\sqrt{\sigma_u^2 + \sigma_{vj}^2}}\right)\right]\exp\left(\frac{-\varepsilon_i^2}{2(\sigma_u^2 + \sigma_{vj}^2)}\right),$$

$$\varepsilon_i = y_i - \alpha - \beta^{\mathrm{T}}\mathbf{x}_i.$$

(Note that the indexation over classes pertains to the variance of the symmetric component of $\varepsilon_i$, $\sigma_{v,j}$.) We thus have a class-specific stochastic frontier model. The unconditional model is a probability weighted mixture over the $J$ classes,

$$f_\varepsilon(\varepsilon_i) = \sum_j \pi_j f_\varepsilon(\varepsilon_i|\text{class} = j), 0 < \pi_j < 1, \sum_j \pi_j = 1.$$

The mixing probabilities are additional parameters to be estimated. The resulting unconditional model preserves the symmetry of the two-sided error component but provides a degree of flexibility that is somewhat greater than the simpler half-normal model. The mixture of normals is, with a finite number of classes, nonnormal.

This model lends itself well to either Bayesian (Tsionas and Greene, 2003) or classical (Orea and Kumbhakar, 2004; Greene, 2004a, 2005; Tsionas and Greene, 2003) estimation methods. The likelihood function is defined over $f_\varepsilon(\varepsilon_i)$ in the usual way and, with the one caveat about the number of classes

noted below, is not particularly difficult to maximize.[47] After estimation, a conditional (posterior) estimate of the class that applies to a particular observation can be deduced using Bayes theorem:

$$\text{prob}[\text{class} = j|y_i] = \frac{f(y_i|\text{class} = j)\text{prob}[\text{class} = j]}{\sum_{j=1}^{J} f(y_i|\text{class} = j)\text{prob}[\text{class} = j]} = \hat{\pi}_{j|i}$$

One would then assign an individual observation to the most likely class. Subsequent analysis, for example, efficiency estimation (see section 2.5), would then be based on the respective class for each observation.

Orea and Kumbhakar (2004), Tsionas and Greene (2003), and Greene (2004a, 2005) have extended this model in two directions. First, they allow the entire frontier model, not just the variance of the symmetric error term, to vary across classes. This represents a discrete change in the interpretation of the model. For the case above, the mixture model is essentially a way to generalize the distribution of one of the two error components. For the fully mixed models, we would reinterpret the formulation as representing a latent regime classification. In Greene (2004b), for example, the latent class model is proposed (ultimately with very limited success) as a means of accommodating heterogeneity in the extremely heterogeneous World Health Organization (Evans et al., 2000a, 2000b) data set. The implication of the more general model is that firms are classified into a set of different technologies and efficiency distributions. The specific classification is unknown to the analyst, hence the probabilistic mixing distribution. (This has a distinctly Bayesian flavor to it, as, in fact, the individual firm does reside in a specific class, but the analyst has only a set of priors, or mixing probabilities, to suggest which.) The second extension in the latter papers is to allow heterogeneity in the mixing probabilities:

$$\pi_{ij} = \frac{\exp(\boldsymbol{\theta}_j^{\mathrm{T}}\mathbf{z}_i)}{\sum_{j=1}^{J}\exp(\boldsymbol{\theta}_j^{\mathrm{T}}\mathbf{z}_i)}, \boldsymbol{\theta}_J = \mathbf{0}$$

The remainder of the model is a class-specific stochastic frontier model

$$f_\varepsilon(\varepsilon_i|\text{ class} = j) = \frac{2}{\sigma_j}\phi\left(\frac{\varepsilon_i|j}{\sigma_j}\right)\left[\Phi\left(\frac{-\lambda_j\varepsilon_i|j}{\sigma_j}\right)\right],$$
$$\varepsilon_i|j = y_i - \alpha_j - \beta_j^{\mathrm{T}}\mathbf{x}_i$$

This form of the model has all parameters varying by class. By suitable equality restrictions, however, subsets of the coefficients, such as the technology parameters, $\alpha$ and $\boldsymbol{\beta}$, can be made generic.

There remains a modeling loose end in this framework. The number of classes has been assumed to be known, but there is no reason to expect this. How to determine the appropriate number of classes is an ongoing problem

in this literature. In principle, one could use a likelihood ratio test to test down from a $J$ class model to a $J-1$ class model. However, the number of degrees of freedom for the test is ambiguous. If the model parameters are the same in two classes, then the number of classes is reduced by one whether or not the two probabilities are similarly restricted. One cannot test "up" from a $J-1$ class model to a $J$ class model, because if the correct model has $J$ classes, then the $J-1$ class model estimators will be inconsistent. A number of researchers have produced proposals for handling this problem, many of them involving information criteria such as the Akaike information criterion.

The latent class approach provides a means to build a large amount of cross-firm heterogeneity into the model. As discussed in section 2.6, this is a major, important extension of the model. With a sufficiently large number of classes, one can achieve quite a large amount of generality. As the number of classes grows, the model approximates a full random-parameters model, which is reconsidered in section 2.7.

The recent literature contains a number of studies of semiparametric approaches to frontier modeling. As discussed above, the "semiparametric" aspect of the model means different things in different studies. Sickles et al. (2002) and Sickles (2005) have loosened the assumptions about the "effects" in a deterministic frontier model. Orea, Kumbhakar, Greene, Tsionas, and others have relaxed the assumptions about all the model parameters through a finite mixture approach. Note, finally, two studies, Kopp and Mullahy (1989) and Griffin and Steel (2004), that have retained the essential structure of the stochastic frontier model but specifically focused on the specification of the inefficiency random variable, $u_i$. Kopp and Mullahy (1989) have derived GMM estimators for the stochastic frontier model that require only that the distribution of $v_i$ be symmetric, that the distribution of $u_i$ be defined over the positive half of the real line, and that moments of $u_i$ and $v_i$ up to order six be finite. This provides a high level of generality, but at the very high cost that the method produces no definable estimate of $u_i$, which ultimately is the point of the exercise. Under the assumptions made thus far, OLS estimates of the model with an adjusted constant term $(\alpha + E[u_i])$ satisfies the assumptions of the Gauss Markov theorem. Griffin and Steel (2004) explore what one might reluctantly call a "normal–Dirichlet model":

$$y_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_i,$$

where the model is all as above specified save for $u_i \sim F$, a "random probability measure generated by a Dirichlet process." A variety of parametric settings are explored, with the finding that the results (estimates of $E[u_i|\text{data}]$—a Bayesian estimator) are fairly strongly dependent on the assumptions. It does emerge that a fully parametric, normal–gamma model (estimated, again, using

MCMC procedures) fairly well resembles the much more general Dirichlet results.

#### 2.4.9.3 Nonparametric approaches

Kumbhakar, Park, Simar, and Tsionas (2005; see also Kumbhakar and Tsionas, 2002) suggested the following nonparametric approach. The global MLE of the parameters of the normal–half-normal model[48] are

$$\left[\hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\sigma}, \hat{\lambda}\right]_{\text{MLE}} = \arg\max \ln L(\alpha, \boldsymbol{\beta}, \sigma, \lambda | \text{ data})$$
$$= \sum_{i=1}^{N} \frac{2}{\sigma} \phi\left(\frac{\varepsilon_i}{\sigma}\right) \left[\Phi\left(\frac{-\varepsilon_i \lambda}{\sigma}\right)\right].$$

Local maximization of the log-likelihood for the nonparametric model involves the following: Choose a multivariate kernel function

$$K(\mathbf{d}) = (2\pi)^{-m/2} |\mathbf{H}|^{-1/2} \exp[-(1/2)\mathbf{d}^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{d}],$$

where $\mathbf{d}$ is a difference vector (defined below), $m$ is the number of parameters in $\boldsymbol{\beta}$, $\mathbf{H} = h\mathbf{S}$ where $\mathbf{S}$ is the sample covariance of the variables on the right-hand side, and $h$ is a bandwidth.[49] Then, for a particular value of $\mathbf{x}^*$, the local estimator is defined by

$$\left[\hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\sigma}, \hat{\lambda}\right](\mathbf{x}^*) = \arg\max \ln L_K(\alpha, \boldsymbol{\beta}, \sigma, \lambda | \text{ data})$$
$$= \sum_{i=1}^{N} \frac{2}{\sigma} \phi\left(\frac{\varepsilon_i}{\sigma}\right) \left[\Phi\left(\frac{-\varepsilon_i \lambda}{\sigma}\right)\right] K(\mathbf{x}_i - \mathbf{x}^*).$$

A full vector of parameters is defined for each vector $\mathbf{x}^*$ chosen. The authors suggest four reasons to prefer this approach: (1) There can be no functional form misspecification, since the full-parameter vector is a function of the data at every point. (2) The variances are also functions of $\mathbf{x}$, so the model allows for heteroskedasticity of unknown form. (I return to this issue below.) (3) In their truncation model, the mean of the underlying inefficiency distribution is also a function of $\mathbf{x}$, which represents a considerable generalization of the model. (4) This model generalizes Berger and Humphrey's (1991, 1992) thick frontier. While Berger and Humphrey's approach fits the model to specific quartiles of the data, this model fits the frontier at all points of the sample.

In a series of studies, Berger and Humphrey (e.g., 1991, 1992) analyze what they label the "thick frontier" approach to efficiency estimation. The analysis proceeds by first dividing the sample into classes by size and then within the size classes further subdividing the observations on the basis of average costs. "Best-practice" frontier models are then fit to the lowest quartiles of

the size classes using OLS or GLS. Berger and Humphrey (1991) analyze a three-output translog cost function. They argue that this approach combines the logic of the DEA "best practice," data-driven analysis and the appealing feature of the stochastic frontier model that combines both randomness in the frontier (its "thickness") with a formal model of inefficiency. However, the thick frontier approach is somewhat less parameterized than the stochastic frontier while at the same time having more structure than DEA. A number of authors (e.g., Mester, 1994; Wagenvoort and Schure, 2005) have used the thick frontier method to analyze cost inefficiency in the banking industry. Berger and Humphrey (1992) is a panel-data approach that adds exogenous heterogeneity to the basic model. (See section 2.6 for additional material on heterogeneity in efficiency analysis.) To the extent that it isolates inefficiency in the data, this technique is a nonparametric frontier estimator insofar as no distribution is assumed. A thoroughly detailed application of the thick frontier concept is given in Lang and Welzel (1998).

Note, finally, that the entire body of results on DEA can be viewed as a distribution-free, nonparametric approach to frontier estimation and efficiency analysis. Because DEA is treated in great detail in chapter 3, I do not pursue the subject here. Another concise, very readable introduction to the topic is given in Murillo-Zamorano (2004).

#### 2.4.9.4 Conclusion

All of these studies suggest that there is considerable scope for alternatives to the original normal–half-normal model of Aigner et al. All have appeared in applications in the literature. Nonetheless, the normal–half-normal model, along with some of the variants discussed below (e.g., the heteroskedastic model) has provided the most frequent specification for the recent research.

## 2.5 Stochastic Frontier Cost Functions, Multiple Outputs, and Distance and Profit Functions: Alternatives to the Production Frontier

This section discusses a variety of specifications that model production and (in)efficiency in functional forms that differ from the single-output production function examined up to this point.

### 2.5.1 Multiple-output production functions

The formal theory of production departs from the transformation function that links the vector of outputs, $\mathbf{y}$, to the vector of inputs, $\mathbf{x}$:

$$T(\mathbf{y}, \mathbf{x}) = 0$$

As it stands, some further assumptions are obviously needed to produce the framework for an empirical model. By assuming homothetic separability, the function may be written in the form

$$A(\mathbf{y}) = f(\mathbf{x})$$

(see Fernandez et al., 2000, for discussion of this assumption). The function $A(\mathbf{y})$ is an output aggregator that links the "aggregate output" to a familiar production function. The assumption is a fairly strong one, but with it in place, we have the platform for an analysis of (in)efficiency along the lines already considered. Fernandez et al. (2000) proposed the multiple-output production model,

$$\left(\sum_{m=1}^{M} \alpha_m^q y_{i,t,m}^q\right)^{1/q} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it}.$$

Inefficiency in this setting reflects the failure of the firm to achieve the maximum aggregate output attainable. Note that the model does not address the economic question of whether the chosen output mix is optimal with respect to the output prices and input costs. That would require a profit function approach. Fernandez et al. (2000) apply the method to a panel of U.S. banks—the 798-bank, 10-year panel analyzed by Berger (1993) and Adams et al. (1999).[50] Fernandez et al. (1999, 2000, 2002, 2005) have extended this model to allow for "bads," that is, undesirable inputs. Their model consists of parallel equations for the "goods" (dairy output of milk and other goods in Dutch dairy farms) and "bads" (nitrogen discharge). The two equations are treated as a Seemingly Unrelated Regressions system and are fit (as is the banking model) using Bayesian MCMC methods. The study of the electric power industry by Atkinson and Dorfman (2005) takes a similar approach, but fits more naturally in section 2.5.4, which examines it in a bit more detail.

### 2.5.2 Stochastic frontier cost functions

Under a set of regularity conditions (see Shephard, 1953; Nerlove, 1963), an alternative representation of the production technology is the *cost function*,

$$C(y, \mathbf{w}) = \min\{\mathbf{w}^{\mathrm{T}}\mathbf{x} : f(\mathbf{x}) \geq y\},$$

where $\mathbf{w}$ is the vector of exogenously determined input prices. The cost function gives the minimum expenditure needed to produce a given output, $y$. If a producer is technically inefficient, then its costs of production must exceed the theoretical minimum. It seems natural, then, to consider a frontier cost function as an alternative to the frontier production function model. The interpretation of the inefficiency terms in an empirical model is complicated a bit by the dual approach to estimation, however. Suppose that, on the production side of the model, the representation of a one-sided error term as reflective

purely of technical inefficiency is appropriate. The computation is conditional on the inputs chosen, so whether the choice of inputs is itself allocatively efficient is a side issue. On the cost side, however, *any* errors in optimization, technical *or* allocative, must show up as higher costs. As such, a producer that we might assess as operating technically efficiently by a production function measure might still appear inefficient viz-à-viz a cost function.

Similar arguments would apply to a profit function. This does not preclude either formulation, but one should bear in mind the possible ambiguities in interpretation in these alternative models. It might make more sense, then, to relabel the result on the cost side as "cost inefficiency." The strict interpretation of technical inefficiency in the sense of Farrell may be problematic, but it seems counterproductive to let this be a straightjacket. The argument that there is a cost frontier that would apply to any given producer would have no less validity. Deviations from the cost frontier could then be interpreted as the reflection of both technical and allocative inefficiency. At the same time, both inefficiencies have a behavioral interpretation, and whatever effect is carried over to the production side is induced, instead. The same logic would carry over to a profit function. The upshot of this argument is that estimation techniques that seek to decompose cost inefficiency into an allocative and a true Farrell measure of technical inefficiency may neglect to account for the direct influence of output itself on the residual inefficiency once allocative inefficiency is accounted for.

Let us begin by examining the costs of production of a single output conditioned on the actual input choices. That is, neglecting the first-order conditions for optimality of the input choices, we consider the implications for the costs of production of technical inefficiency. For simplicity, we assume constant returns to scale. The production function, $f(\mathbf{x})$, is linearly homogeneous and therefore homothetic. For homothetic production functions,[51]

$$y = F[f(\mathbf{x})],$$

where $F(t)$ is a continuous and monotonically increasing function when $t$ is positive. We have the fundamental result (from Shephard, 1953) that the corresponding cost function is

$$C(y, \mathbf{w}) = F^{-1}(y)c(\mathbf{w}),$$

where $c(\mathbf{w})$ is the unit cost function. For the stochastic frontier production function, then

$$y_i = f(\mathbf{x}_i)\mathrm{TE}_i e^{v_i},$$

so that the cost function is

$$C_i = F^{-1}(y)c(\mathbf{w}_i)\left[\frac{1}{\mathrm{TE}_i}\right]e^{-v_i}.$$

This corresponds to Farrell's (1957) original efficiency measure, that is, the cost savings that would be realized if output were produced efficiently. The

theoretical counterpart would be the input-based measure. In logs, then,

$$\ln C_i = \ln F^{-1}(y) + \ln c(\mathbf{w}_i) - \ln \mathrm{TE}_i - v_i.$$

In terms of our original model, then, the stochastic cost frontier is

$$\ln C_i = \ln F^{-1}(y) + \ln c(w_i) - v_i + u_i,$$

which is what might be expected. The sign on $v_i$ is inconsequential since its mean is zero and the distribution is symmetric (normal).

Now, suppose there are economies of scale in production. For the simplest case, we assume a Cobb-Douglas function with degree of homogeneity $\gamma$. The stochastic frontier cost function will be

$$\ln C_i = A' + \beta \ln w_{1i} + (1 - \beta) \ln w_{2i} + \frac{1}{\gamma} \ln y_i + \frac{1}{\gamma}(-v_i) + \frac{1}{\gamma} u_i.$$

Therefore, the composed disturbance on the cost frontier is

$$\varepsilon_i' = \frac{1}{\gamma}(-v_i + u_i).$$

The upshot is that the presence of economies of scale on the production side blurs somewhat the reflection of technical inefficiency on the cost side. The preceding result is general for a production function that exhibits a fixed degree of homogeneity.[52]

Evidently, the simple interpretation of the one-sided error on the cost side as a Farrell measure of inefficiency is inappropriate *unless the measure is redefined in terms of costs, rather than output.* That is, one might choose to make costs, rather than output, the standard against which efficiency is measured. At least in this context, this is nothing more than a matter of interpretation. It is equally clear that by some further manipulation, the estimated inefficiency obtained in the context of a cost function can be translated into a Farrell measure of technical inefficiency, that is, just by multiplying it by $\gamma$.

For the simple case above in which the production function is homogeneous, the effect of economies of scale can be removed by rescaling the estimated disturbance. A corresponding adjustment may be possible in more involved models such as the translog model. Suppose that the production function is homothetic, but not homogeneous. For convenience, let

$$G(y_i) = F^{-1}(y_i).$$

Then

$$\ln C_i = \ln c(\mathbf{w}_i) + \ln G(y_i).$$

The formulation above is clearly a special case. Unless $\ln G(.)$ is linear in $\ln y_i$, as it is when the production function is homogeneous, the technical inefficiency

may be carried over to the cost function in a very complicated manner.[53] The usual assumption that $u_i$ in the stochastic frontier cost function can vary independently of $y_i$ may be problematic.[54]

Any errors in production decisions would have to translate into costs of production higher than the theoretical norm. Likewise, in the context of a profit function, any errors of optimization would necessarily translate into lower profits for the producer. But, at the same time, the stochastic nature of the production frontier would imply that the theoretical minimum cost frontier would also be stochastic. Some recent applications that have been based on cost functions have made this explicit by further decomposing the stochastic term in the cost function to produce

$$\ln C_i = \ln C(y_i, \mathbf{w}_i) + v_i + u_i + A_i,$$

where $A_i$ is strictly attributable to allocative inefficiency (see, e.g., chapter 4 in Kumbhakar and Lovell, 2000).

The preceding describes the production and cost of the firm in long-run "equilibrium." (The concept must be qualified, because it is unclear whether it is appropriate to characterize an inefficient firm as being in equilibrium.) For the short term, in which there are fixed inputs, the variable cost function is

$$\ln C^F = \ln C(y, \mathbf{w}, \mathbf{x}^F).$$

As before, relative to optimal costs, any deviation from optimality must translate into higher costs. Thus, for example, with one output and one fixed input, one might analyze a translog variable cost function

$$\begin{aligned}\ln C^F = {} & \alpha + \sum_{k=1}^{K} \beta_k \ln w_k + \beta_F \ln F + \beta_y \ln y \\ & + \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K} \gamma_{kl} \ln w_k \ln w_l + \frac{1}{2}\gamma_{FF} \ln^2 F + \frac{1}{2}\gamma_{yy} \ln^2 y. \\ & + \sum_{k=1}^{K} \gamma_{kF} \ln w_k \ln F + \sum_{k=1}^{K} \gamma_{ky} \ln w_k \ln y + \gamma_{Fy} \ln F \ln y + v_i + u_i\end{aligned}$$

In their analysis of Swiss nursing homes, Farsi and Filippini (2003) specified a cost function with labor and capital treated as variable factors and number of beds treated as a fixed input. The variable cost function provides a useful datum; the shadow cost of a fixed input is $-\partial C^F / \partial x^F$. For the translog variable cost function, this would be

$$\frac{-\partial C^F}{\partial F} = \frac{-F}{C^F}(\beta_F + \gamma_{FF} \ln F + \sum_{k=1}^{K} \gamma_{kF} \ln w_k + \gamma_{Fy} \ln y).$$

### 2.5.3 Multiple-output cost functions

A significant advantage of analyzing efficiency on the cost side is the ease with which multiple outputs can be accommodated. Consider a transformation function

$$T(\mathbf{y}, \mathbf{x}) = 0,$$

where $\mathbf{y}$ is a vector of $M$ outputs and $\mathbf{x}$ is a vector of $K$ inputs. Assuming that production satisfies the necessary regularity conditions (including monotonicity, smoothness, and quasiconcavity), we may deduce that the cost function is of the form

$$\ln C_i = \ln C(y_1, \ldots, y_M, w_1, \ldots, w_K),$$

where the cost function is monotonic in outputs, monotonic in each input price, linearly homogeneous in the input prices, and so on. How we proceed from here, and how "inefficiency" enters the specification, depends crucially on the assumptions and will highlight the utility of not allowing the input versus output orientation discussed above to straightjacket the analysis.

Many analyses have proceeded directly to specification of a multiple-output translog cost function

$$\begin{aligned}\ln C_i = \alpha &+ \sum_{k=1}^{K} \beta \ln w_{ik} + \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K} \gamma_{kl} \ln w_{ik} \ln w_{il} \\ &+ \sum_{m=1}^{M} \delta \ln y_{im} + \frac{1}{2}\sum_{m=1}^{M}\sum_{r=1}^{M} \phi_{mr} \ln y_{im} \ln y_{ir} \\ &+ \sum_{m=1}^{M}\sum_{k=1}^{K} \kappa_{mk} \ln y_{im} \ln w_{ik} + v_i + u_i\end{aligned}.$$

(One could also analyze a multiple-output variable cost function, with one or more fixed factors.) Note that there is no necessary assumption of homotheticity or separability on the production side. Tsionas and Greene (2003) analyze a cost frontier for U.S. banks in which there are five outputs and five inputs. In this formulation, $u_i$ is interpreted as "economic inefficiency." Thus, the source of $u_i$ is either technical or allocative inefficiency, or both.

Analyses of two industries in particular, health care and banking, have yielded a rich crop of applications and development of new methods. Data in the banking industry are of particularly high quality. A few of the innovative studies in banking are as follows:[55]

- Lang and Welzel (1998) fit a translog, five-output, three-input cost function to German banking data. The study develops the thick frontier estimator.

- Ferrier and Lovell (1990) fit a multiple-output cost function to U.S. banking data. Among the innovations in this study were a decomposition of cost inefficiency into technical and allocative components and the inclusion of a large number of "environmental" variables in the cost function.
- Huang and Wang (2004) used the Fourier functional form in conjunction with a translog kernel to study a sample of Taiwanese banks.
- Tsionas and Greene (2003) fit a finite mixture of translog multiple-output cost functions to U.S. banks. Orea and Kumbhakar (2004) fit a similar mixture of translog functions using a panel of Spanish banks.

In each of these cases, the cost functions involved three or five outputs, multiple inputs, and a variety of model specifications. The methodologies span the range of techniques already listed, including both classical and Bayesian methods. The health care industry also provides a natural setting for multiple-output cost frontier analysis. In the banking industry studies, a challenging specification issue is how to identify the inputs and outputs and how to distinguish them—for example, are commercial loans, which produce loan interest income, an input or an output? A reading of the many received studies suggests that researchers have come to some agreement on these questions. In health care, there are difficult issues of identifying what the outputs are and, in some cases, measuring them. For example, the measurement of quality is a recurrent theme in this literature. Another question concerns residents in hospital cost studies—is training of residents an input or an output? Again, there are many questions in the literature, but there does seem to be at least broad agreement. A few studies that illustrate the analyses are as follows:

- Koop et al. (1997) use Bayesian methods to fit translog cost frontiers to a panel of U.S. hospitals. In their study, the outputs are number of discharges, number of inpatient days, number of beds, number of outpatient visits, and a case mix index. They also included a quasi-fixed input, capital in their cost function.
- Rosko (2001) analyzes a panel of U.S. hospitals. The translog cost function includes outputs inpatient discharges and outpatient visits. The mix of cases is also considered in this study but not as an output variable. A variety of panel-data techniques (Battese and Coelli, 1995) and models for heterogeneity in inefficiency are placed in the specification.
- Linna (1998) is similar to Rosko (2001) but also considers nonparametric (DEA) bases of inefficiency.
- Farsi and Filippini's (2003) analysis of Swiss nursing home costs analyzes a single output but includes two indicators of quality: a "dependency" index that reflects the intensity of care received by the facility's patients, and a nursing staff ratio.

### 2.5.4 Distance functions

The multiple-output cost frontier and the transformation function provide convenient vehicles for analyzing inefficiency in multiple-output contexts. Another approach that has proved useful in numerous empirical studies is based on the distance function. For output vector $\mathbf{y}$ and input vector $\mathbf{x}$, Shephard's (1953) *input distance function* is $D_I(\mathbf{y}, \mathbf{x}) = \max(\lambda : \mathbf{x}/\lambda$ is on the isoquant for $\mathbf{y})$. It is clear that $D_I(\mathbf{y}, \mathbf{x}) \geq 1$ and that the isoquant is the set of $\mathbf{x}$ values for which $D_I(\mathbf{y}, \mathbf{x}) = 1$. The corresponding output distance function would be $D_O(\mathbf{x}, \mathbf{y}) = \min(\lambda : \mathbf{y}/\lambda$ is producible with $\mathbf{x})$. In this instance, $D_O(\mathbf{y}, \mathbf{x}) \leq 1$. The definitions suggest efficiency measures, as noted earlier. Thus, the input distance suggests the degree to which $\mathbf{x}$ exceeds the input requirement for production of $\mathbf{y}$, which we would identify with cost, or "economic" inefficiency. Likewise, the output distance suggests the degree to which output falls short of what can be produced with a given input vector, $\mathbf{x}$, which relates to the technical inefficiency we have examined thus far.

To put these functions in the form of an econometric model, we use the restrictions implied by the underlying theory, namely, that the input distance function is linearly homogeneous in the inputs and the output distance function is linearly homogeneous in the outputs (see Kumbhakar et al., 2004). Thus, we normalize the input distance function on the (arbitrarily chosen) first input, $x_1$, and the output distance function on $y_1$ to write

$$x_1 D_I(x_2/x_1, x_3/x_1, \ldots, x_K/x_1, \mathbf{y})\text{TI} = 1,$$

where TI is the technical inefficiency index, $0 \leq \text{TI} \leq 1$. In similar fashion, we can formulate the output distance function,

$$y_1 D_O(\mathbf{x}, y_2/y_1, y_3/y_1, \ldots, y_M/y_1)\text{TO} = 1,$$

where TO is the economic inefficiency index, $\text{TO} \geq 1$. This formulation provides a natural framework for a stochastic frontier model. Most applications have used the translog form. Doing likewise, we write

$$0 = \ln x_1 + \ln D_I(x_2/x_1, x_3/x_1, \ldots, x_K/x_1, \mathbf{y}) + v + \ln[\exp(-u)],$$

where the deterministic part of the equation is formulated as the production model, $v$ captures the idiosyncratic part of the model as usual, and $u > 0$ produces $\text{TI} = \exp(-u)$. For the output distance function, a similar strategy produces

$$0 = \ln y_1 + \ln D_O(\mathbf{x}, y_2/y_1, y_3/y_1, \ldots, y_M/y_1) + v + \ln[\exp(u)].$$

Finally, in order to form a model that is amenable to familiar estimation techniques, we would shift the normalized variable to the left-hand side of the equation. Thus, the input distance stochastic frontier model would appear

$$-\ln x_1 = \ln D_I(x_2/x_1, x_3/x_1, \ldots, x_K/x_1, \mathbf{y}) + v - u,$$

and likewise for the output distance equation. Some methodological issues remain. As stated, it would seem that both input and output models would carry some type of simultaneous equations aspect, so that conventional estimators such as OLS would be persistently biased. Coelli (2000) and Cuesta and Orea (2002) consider these issues theoretically. Note that these methodologically oriented examinations come after the leading applications of the distance function technique (e.g., Sickles et al., 2002; Coelli and Perelman, 1996, 1999, 2000; all of which used the translog form as the modeling platform).

The distance function bears close resemblance to other specifications for studying efficiency. Thus, there have been comparisons of inefficiency estimates obtained from estimated distance functions to the counterparts obtained from DEA studies (see Coelli and Perelman, 1999; Sickles et al., 2002). Atkinson, Fare, and Primont (2003) used the concept of the distance function to derive a shadow cost function with which they studied allocative inefficiency. Finally, O'Donnell and Coelli (2005) forced the classical curvature (regulatory) conditions on their estimated distance function. They suggested their method of imposing restrictions on parameters in a Bayesian framework as an alternative to Kleit and Terrell (2001)—they used a Metropolis-Hastings procedure as opposed to Kleit and Terrell's accept/reject iteration.

Atkinson and Dorfman (2005) have extended the distance function method to include both desirable and undesirable outputs in the generation of electricity. The translog input distance function is of the form

$$0 = \gamma_0 + T(\mathbf{y}_g, \mathbf{y}_b, t, \mathbf{x}) + v_{it} - u_{it},$$

where $\mathbf{y}_g$ is a vector of "goods" (residential and commercial/industrial generation), $\mathbf{y}_b$ is a vector of "bads" (sulfur dioxide emissions), $t$ is a time trend, and $\mathbf{x}$ is a vector of inputs (fuel, labor, and capital). $T(\ldots)$ is a full translog function. The underlying theory imposes a large number of linear constraints on the (also large number of) model parameters. In this study, the "bad" is treated as a "technology shifter" (in contrast to Fernandez et al., 2000, who treated nitrogen runoff in dairy farming as an undesirable output). The estimator in this study is an elaborate form of Bayesian method of moments (see Kim, 2002; Zellner and Tobias, 2001).

### 2.5.5 Profit functions

The methodology described earlier can, in principle, be extended to revenue and profit functions. In terms of the received empirical literature, these two approaches have been less actively pursued than production, cost, and distance functions. Two explanations stand out. First, the estimation of a profit function would require a much greater range of assumptions about producer and market behavior. While production and cost functions are clearly reflective of individual firm optimization behavior, the profit function requires additional assumptions about market structure and price setting. Second, the data

demands for profit functions are considerably greater than those for cost and production functions.

A full implementation of a model for a profit frontier would include a production function and the first-order conditions for optimization (see Kumbhakar and Bhattacharyya, 1992; Kumbhakar and Lovell, 2000; Kumbhakar, 2001). For a multiple-output firm/industry, it would also require equations for the optimal mix and levels of the outputs. A full simultaneous equations framework (replete with many nonlinearities) is detailed in Kumbhakar and Lovell (2000; see also chapter 5). The authors also discuss the possibility of a "variable" profit function that takes some inputs as fixed. Again, the underlying assumptions behind such a model require much detail. The profit function framework shares a characteristic with the cost function; profit "inefficiency" would be a mix of both technical and allocative inefficiency. Moreover, there is a third layer that does not enter any of the frameworks considered thus far. For given output prices, any deviation from the optimal mix of outputs must reduce profits. Thus, this model presents yet another application of the "Greene problem" (discussed in greater detail below). Kumbhakar and Lovell (2000, p. 214) list a number of applications of different approaches to profit function estimation. Not surprisingly, because of the ready availability of very high-quality data, several of these studies (e.g., Akhavein et al., 1994; Berger and Mester, 1997; Humphrey and Pulley, 1997; Lozano-Vivas,1997) analyze (in)efficiency in the banking industry.

### 2.5.6 Output-oriented and input-oriented inefficiency

For output vector $\mathbf{y}$ and input vector $\mathbf{x}$, Shephard's (1953) *input distance function* is $\mathrm{D}_I(\mathbf{y},\mathbf{x}) = \max(\lambda : \mathbf{x}/\lambda$ is on the isoquant for $\mathbf{y})$; $D_I(\mathbf{y}, \mathbf{x}) \geq 1$. The corresponding output distance function would be $D_O(\mathbf{x},\mathbf{y}) = \min(\theta : \mathbf{y}/\theta$ is producible with $\mathbf{x})$; $D_O(\mathbf{y},\mathbf{x}) \leq 1$. The input distance suggests the degree to which $\mathbf{x}$ exceeds the input requirement for production of $\mathbf{y}$, which we would identify with cost, or "economic" inefficiency. The output distance suggests the degree to which output falls short of what can be produced with a given input vector, $\mathbf{x}$, which relates to the technical inefficiency examined thus far. The definitions suggest efficiency measures, as noted above. The translation of these notions into frontier models has produced the familiar modeling platforms for production of a single output. Skipping the obvious algebraic steps, we have the generic stochastic frontier model

$$y_i = f(\mathbf{x}_i)\theta_i \exp(v_i),$$

or

$$\ln y_i = \ln f(\mathbf{x}_i) + v_i + \ln \theta_I,$$

where $\theta_i = \exp(-u_i)$ in our model for output-oriented inefficiency. Taking logs produces our familiar stochastic frontier production model. For input-oriented inefficiency, we have the less commonly used formulation,

$$y_i = f(\lambda_i \mathbf{x}_i) \exp(v_i),$$

or

$$\ln y_i = \ln f(\lambda_i \mathbf{x}_i) + v_i.$$

In this formulation, the form of inefficiency in the production model is less clear. For example, moving to the usual Cobb-Douglas or translog model leaves a complicated function of $(\ln x_{ki} + \ln \lambda_i)$.

Most of the received applications have measured output-oriented inefficiency on the production side. On the cost side of the production model, the roles of the two measures are reversed. Neglecting $v_i$ for the moment purely for convenience), we have

$$y_i = \theta_i f(\mathbf{x}_i) \Leftrightarrow C_i = g(y_i/\theta_i, \mathbf{w}_i),$$

so unless $y_i$ enters the cost function (log)linearly, the form that $\theta_i$ takes in the cost function will be complicated. In contrast, for input-oriented technical inefficiency, we have

$$y_i = f(\lambda_i \mathbf{x}_i) \Leftrightarrow C_i = g(y_i, \mathbf{w}_i/\lambda_i).$$

For technologies that satisfy the regularity conditions for the dual relationships to exist, the cost function must be linearly homogeneous in the input prices. Thus, we must have

$$C_i = (1/\lambda_i) g(y_i, \mathbf{w}_i).$$

Taking logs here and using the usual interpretation of $\lambda_i$ produces

$$\begin{aligned} \ln C_i &= \ln g(y_i, \mathbf{w}_i) - \ln \lambda_{ii} \\ &= \ln g(y_i, \mathbf{w}_i) + u_i. \end{aligned}$$

Thus, we see that familiar applications of stochastic cost frontiers are based on a measure of input inefficiency. [I.e., unless it is assumed that the production function is homogeneous. If so, then $\ln C_i = (1/\gamma) \ln(y_i/\theta_i) c(\mathbf{w}_i)$, where $\gamma$ is the degree of homogeneity (see Christensen and Greene, 1976). In this case, input- and output-oriented inefficiency will be indistinguishable.]

Numerous applications have analyzed the distinction between input- and output-oriented inefficiency. Atkinson and Cornwell (1993), using panel data and a linear fixed-effects (deterministic frontier) model, found (perhaps not surprisingly) that the two assumptions produced different rankings of observations. As they point out, the distinction "matters." In similar kinds of analyses, Kumbhakar et al. (2004) and Alvarez et al. (2004) tested for the

presence of the two types of inefficiency. The latter study proceeded more or less on the lines of Atkinson and Cornwell (1993), using a panel-data set on Spanish dairy farms. Orea and Kumbhakar (2004). fit both input- and output-oriented models, and a hybrid that included both. They used a Vuong (1989) test was to test for the specification. Kurkalova and Carriquiry (2003) (using a technique suggested by Reinhard, Lovell, and Thijssen, 1999) estimated output-oriented inefficiency measures and then translated them ex post into input-oriented measures. Huang and Wang (2004) have also fit separate cost frontier models for input- and output-oriented inefficiency, in their case using the Fourier flexible form.

The preceding studies have steered around the inherent difficulty of the input orientation on the production side. Consider, in particular, a translog model, where we assume, as other authors have (looking ahead), panel data and time invariance for the inefficiency term. Thus,

$$\ln y_{it} = \alpha + \sum_{k=1}^{K} \beta_k (\ln x_{i,t,k} - u_i) + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \gamma_{kl} (\ln x_{i,t,k} - u_i)(\ln x_{i,t,l} - u_i) + v_{i,t},$$

where $u_i \geq 0$. Consistent with the received work, we would assume that $u_i$ has a half-normal or exponential (or gamma or lognormal) distribution. As usual, estimation of the parameters is complicated by the presence of the unobserved $u_i$. Consider the following approach based on MSL suggested by Kumbhakar and Tsionas (2004). (We have stripped their derivation down to its bare essentials here and changed notation a bit.) Note, first, that $u_i$ is the same for all $t$ and for all $k$, for a given $i$.

For convenience, write

$$z_{i,t,k}(u_i) = \ln x_{i,t,k} - u_i.$$

Conditioned on $u_i$, each term in the log-likelihood for $y_{it}$ is the log of the corresponding normal density (for $v_{i,t}$), so

$$\ln L|\mathbf{u} = \sum_{i=1}^{N} \left[ -\frac{T_i}{2} \ln 2\pi - \frac{1}{2\sigma^2} \sum_{t=1}^{T_i} \left( \ln y_{it} - T\left[ z_{i,t}(u_i) \right] \right)^2 \right].$$

where

$$T\left[\mathbf{z}_{i,t}(u_i)\right] = \alpha + \sum_{k=1}^{K} \beta_k z_{i,t,k}(u_i) + \tfrac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \gamma_{kl} z_{i,t,k}(u_i) z_{i,t,l}(u_i).$$

The inefficiency term must be integrated out of the log-likelihood before it can be maximized. The unconditional log-likelihood is

$$\ln L = \sum_{i=1}^{N} \int_{u_i} \left[ -\frac{T_i}{2} \ln 2\pi - \frac{1}{2\sigma^2} \sum_{t=1}^{T_i} \left( \ln y_{it} - T\left[\mathbf{z}_{i,t}(u_i)\right] \right)^2 \right] p(u_i)\, du_i.$$

The integrals cannot be expressed in closed form, so as it is above, this log-likelihood is not usable. However, for the distributions mentioned (half-normal, exponential), random draws on $u_i$ are easily obtainable. A usable simulated log-likelihood function is

$$\ln L^S = \sum_{i=1}^{N} \frac{1}{R} \sum_{r=1}^{R} \left[ -\frac{T_i}{2} \ln 2\pi - \frac{1}{2\sigma^2} \sum_{t=1}^{T_i} \left( \ln y_{it} - T\left[\mathbf{z}_{i,t}(u_{i,r})\right] \right)^2 \right].$$

Maximizing $\ln L^S$ produces estimates of all of the model parameters. [Tsionas (2004) shows how the Fourier transform produces an alternative, possibly simpler and faster algorithm for this optimization.] Ex post, it is useful to obtain an estimate of $u_i$—this was the purpose of the exercise to begin with. Kumbhakar and Tsionas (2004) suggest a method of approximating $E[u_i|\text{parameters, data}]$. I suggest a different (albeit similar) approach in section 2.7.

There is an element of ambiguity in the model as specified. Which form, input or output, is appropriate for a given setting? Alvarez et al. (2004) suggested that a given firm could be operating in either regime at any time. In their analysis of European railroads, they treated the input and output distance functions as two latent regimes in a finite mixture model. In essence, their model allows the data to sort themselves into the two regimes rather than arbitrarily assuming that all observations obey one or the other at the outset.

## 2.6 Heterogeneity in Stochastic Frontier Function Models

This section is devoted to the issue of between firm heterogeneity in stochastic frontier modeling. We depart from a "pure" production model,

$$\ln y_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_{it} + v_{it} - u_{it},$$

or cost model,

$$\ln C_{it} = C(y_{it}, \mathbf{w}_{it}; \boldsymbol{\beta}) + v_{it} + u_{it},$$

in which $v_{it} \sim N[0, \sigma_v^2]$ and $u_{it}$ has some distribution characterized by a constant mean, $\mu$ and constant variance, $\sigma_u^2$—sometimes both embodied in a single parameter, as in the exponential model. At this departure point, we say that the technology and the inefficiency distributions across individuals and

time are homogeneous. They have the same parameters both in the production or cost function and in the inefficiency distribution. Of course, even at this point, that is not quite true, since the "stochastic" part of the *stochastic frontier* model specifically models the production technology as having a firm-specific (and time-specific) shift factor, $v_{it}$. Thus, at the outset, what we mean by homogeneity in the model is that firms differ only with respect to this random, noisy shift factor. We now wish to incorporate other forms of heterogeneity in the model. This includes, among other features, heteroskedasticity in the random parts of the model and shifts in the technology that are explainable in terms of variables that are neither inputs nor outputs. We begin by defining more precisely what we have in mind by heterogeneity.

### 2.6.1 Heterogeneity

One important way to categorize heterogeneity is between observable and unobservable heterogeneity. By observable heterogeneity, we mean as reflected in measured variables. This would include specific shift factors that operate on the production or cost function (or elsewhere in the model). For example, in his study of hospital costs, Linna (1998) has an "exogenous" variable reflecting case mix in the cost function. How such variables should enter the model is an important question. (In brainstorming sessions on frontier modeling with my colleagues, we call this "where do we put the $z$'s?") They might shift the production function or the inefficiency distribution (i.e., enter the regression functions) or scale them (i.e., enter in the form of heteroskedasticity), or some combination of both (see Alvarez, Amsler, Orea and Schmidt, 2006, on the "scaling property") All of these possibilities fall in the category of observable heterogeneity (as I see it).

Unobserved heterogeneity, in contrast, enters the model in the form of "effects." This is usually viewed fundamentally as an issue of panel data, though I don't necessarily see it that way. Unobserved heterogeneity might (in principle, perhaps, always) reflect missing variables in the model. When these are not missing factors of production, or their unit prices, they have to be labeled as something different, however. Unobserved heterogeneity enters our model as characteristics, usually time invariant, that may or may not be related to the variables already in the model. We submit that unobserved heterogeneity should be considered as distinct from the unobservable object of most of our study, technical or cost inefficiency. For example, Greene (2004b) analyzes the problem of distinguishing the two in the World Health Organization's (WHO, 2000) vastly heterogeneous panel-data set on world health care attainment that includes 191 countries—virtually all of the world's population. I examine the issue in some detail below.

A related issue is parameter, or technology heterogeneity. Several studies to be discussed below have analyzed models with some type of shifting or cross-firm variation in the structural parameters of the model. Many of these are the sort of "random-parameter" models that are, again, usually associated

with modeling in panel-data sets. I digress at this point to pin down a possibly misleading part of the modeling vernacular. In the numerous Bayesian treatments of frontier modeling, the parameters of the model are treated as "random," but the randomness in this context is not what I mean by parameter heterogeneity. In this discussion, what I intend by random parameters [e.g., in Huang (2004) or Orea and Kumbhakar's (2004) latent class model] is random difference across firms or individuals. The "randomness" of the parameters in a Bayesian treatment reflects "uncertainty" of the analyst, not heterogeneity across firms. I suggest the following litmus test: The parameter vector in a "random-parameters model" will contain an observation subscript "$i$," as in

$$\ln y_{it} = \alpha_i + \boldsymbol{\beta}_i^{\mathrm{T}} \mathbf{x}_{it} + v_{it} - u_{it}.$$

The Bayesian counterpart to this is the "hierarchical model," which adds to the preceding priors that might appear as $\boldsymbol{\beta}_i \sim N[\boldsymbol{\beta}, a\boldsymbol{\Omega}]; \boldsymbol{\beta} \sim N[\mathbf{0}, \boldsymbol{\Omega}]$ (see, e.g., Tsionas, 2002; Huang, 2004). Variation in parameters is an important element of many studies. It can also be partly observable, for example, as in Kurkalova and Carriquiry (2003), in which parameters are allowed to vary systematically over time.[56]

A second, very important issue is the distinction between heterogeneity (latent or otherwise) in the production model and heterogeneity in the inefficiency model. These two have quite different implications for modeling and for estimation. Most of the literature on heterogeneity is focused on the latter, although to the extent that omitted heterogeneity in the production or cost model always shows up somewhere else (i.e., in the estimated features of $u_{it}$), they are not unrelated.

### 2.6.2 One-step and two-step models

In cases in which heterogeneity is observable, we are sometimes interested in models in which those observables enter in the form of parameterized functions of "exogenous variables." The leading case is in which these variables are believed to affect the distribution of inefficiency. For example, in Greene (2004b), it is suggested that in the provision of health care, per capita income, and the distribution of income are relevant determinants of the efficiency of health care delivery. In such cases, researchers have often analyzed (in)efficiency in two steps. In the first, conventional estimates of inefficiency are obtained without accounting for these exogenous influences (see section 2.8 for estimation of $u_i$). In the second step, these estimates are regressed or otherwise correlated with the exogenous factors (see, e.g., Greene, 2004b, table 6; Annaert et al., 2001).[57] It is easy to make a convincing argument that not accounting for the exogenous influences at the first step will induce a persistent bias in the estimates that are carried forward into the second. This is analyzed at length in Wang and Schmidt (2002), who argue that this is akin to an omitted variable problem in the linear regression model.

The biases in estimated coefficients will be propagated in subsidiary estimates computed using those coefficients. Caudill and Ford (1993) and Caudill et al. (1995) provide evidence of such first-level biases in estimated technology coefficients that result from neglected heteroskedasticity. Wang and Schmidt (2002) take the analysis another step to consider how this bias affects estimates of "inefficiency."[58] In their model, the neglected heterogeneity "scales" both the mean and variance of the inefficiency distribution. Ultimately, the case made by these authors is that when heterogeneity in the model is parameterized in terms of observables, those features should all appear in the model at the first step. In what follows, I will assume this is the case—the various model extensions noted below all presume "full information" (usually ML) estimation at the first step.

### 2.6.3 Shifting the production and cost function

I have mentioned numerous applications in which exogenous variables that are not outputs, inputs, or input prices enter the model. Among the examples are time changes that likely reflect technological change [e.g., Berger and Mester (1997), the case mix variables in Linna's (1998) hospital cost study, and exogenous country effects such as form of government and climate in Greene (2004b)]. Little is changed in the model by adding exogenous shifts, environment variables, technical change, and so on, to the production, cost, or distance function, as in

$$\ln y_{it} = f(\mathbf{x}_{it}, \beta) + g(\mathbf{z}_{it}, \delta) + h(t) + v_{it} - u_{it};$$

however, it must be noted that there is a potential identification issue. The model is obviously indistinguishable from an otherwise "pure" model in which the inefficiency component is $u_{it}^{*} = g(\mathbf{z}_{it}, \boldsymbol{\delta}) + h(t) - u_{it}$. It is up to the model builder to resolve at the outset whether the exogenous factors are part of the technology heterogeneity or whether they are elements of the inefficiency distribution.

The more pernicious identification problem arises in panel-data models in which there is unobservable, time-invariant heterogeneity. A perennial issue in the analysis of efficiency is whether inefficiency is time invariant or varies through time (systematically or haphazardly). I examine several models that relate to this question in this section. In the WHO health care model (Evans et al., 2000a, 2000b), technical inefficiency is deduced from a fixed-effects model (see Schmidt and Sickles, 1984),

$$\ln y_{it} = a_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - [\max_{j}(a_j) - a_i].$$

In this application (and others of the same type), any unobserved time-invariant heterogeneity must be captured in the estimated "inefficiency," $[\max_j(a_j) - a_i]$. For the WHO data, this component is potentially enormous,

because these are country-level data. A random-effects–style model (see, e.g., Pitt and Lee, 1981; Koop et al., 1997),

$$\ln y_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_i,$$

fares no better—it simply layers on the additional assumption that both inefficiency and heterogeneity are uncorrelated with $\mathbf{x}_{it}$. To accommodate this undesirable feature of both treatments, Greene (2004a, 2004b, 2005) proposes "true" fixed- and random-effects models,

$$\ln y_{it} = a_i + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it}$$

and

$$\ln y_{it} = (\alpha + w_i) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it}.^{59}$$

In both cases, the assumptions of the stochastic frontier model are maintained, so the estimators are ML—in the former case by including the dummy variables in the model and in the latter case by MSL. Note that these models substantively change the assumptions about the time-invariant effects. In the prior specifications, the time-invariant term is entirely time-invariant inefficiency, and time-invariant heterogeneity is either assumed away or inadvertently buried in it. In the "true" effects model, all time-invariant effects are treated as unobserved heterogeneity, and the inefficiency component varies freely through time. Doubtless, the "truth" is somewhere between the two extremes. Unfortunately, there is an identification issue that is only resolved through nonsample information (i.e., additional assumptions). Farsi et al. (2003) have studied the impact of the different assumptions in a model of nursing home costs and found, perhaps not surprisingly, that the differences are quite noticeable. Kotzian (2005) extends the notion a bit to full-parameter vector heterogeneity and finds, likewise, that accounting for heterogeneity has substantial impacts on measured inefficiency.

### 2.6.4 Parameter variation and heterogeneous technologies

In the frontiers context, cross-firm parameter variation would be viewed as heterogeneity in the technology being employed (see Huang, 2004, for discussion). The idea of parameter variability in regression models was proposed by Hildreth and Houck (1968), among others, who applied the idea to linear regression models. The guiding wisdom in many treatments is still provided by the linear model. Textbook treatments of random-parameter models thus often analyze the generalized regression model and methods of "mixing" group-specific least squares estimates—essentially a GLS estimator. Kalirajan and Obwona (1994) is an early application in the frontiers literature. More contemporary treatments have couched parameter variation in terms of parameter heterogeneity, generally in panel-data models. In general, such

models, both classical and Bayesian, are handled through likelihood-based and often simulation methods.[60]

When the parameter variation reflects observable heterogeneity, it is straightforward to build it directly in the model. Thus, Kotzian (2005) uses interactions with group-specific dummy variables to accommodate group differences in a model of health care attainment. Kurklova and Carriquiry (2003) do similarly with time variation in a production model for farm production.

A number of recent treatments have modeled technology heterogeneity with less systematic variation. In Orea and Kumbhakar (2004), Greene (2005), and O'Donnell and Griffiths (2004), a latent class specification is suggested to accommodate heterogeneity across firms in the sample. In the first two of these, the formulation captures differences in groups of firms within the sample. O'Donnell and Griffiths (2004), in contrast, use the latent class formulation to capture the effects of different weather "regimes" on rice farming. The latent class model, in general, is a stochastic frontier model,

$$\ln y_{it}|q = f_q(\mathbf{x}_{it}, \boldsymbol{\beta}_q) + v_{it}|q - u_{it}|q,$$

where $q$ indicates the class or regime. Class membership is unknown, so the model proceeds to add the sorting probabilities,

$$\text{prob}[\text{class} = q|\mathbf{z}_i] = p(q|\mathbf{z}_i).$$

Note how exogenous factors may (but need not) enter the class probabilities. O'Donnell and Griffiths (2004) document a Bayesian MCMC method of estimating the model. Greene (2005) and Orea and Kumbhakar (2004) use ML methods instead.

Tsionas (2002) and Huang (2004) proposed a hierarchical Bayesian approach to frontier modeling with heterogeneous technologies. Tsionas's stochastic frontier model [applied to the Christensen and Greene (1976) electricity generation data] is

$$\ln y_{it} = \alpha + \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it},$$

$$f(v_{it}) = N[0, \sigma^2],\ p(\sigma) = \text{inverted gamma}(s, M) \propto \exp(-s/(2\sigma^2))(\sigma^2)^{-(M+1)/2},$$

$$f(u_{it}) = \theta \exp(-\theta u_{it}),\ p(\theta) = \text{gamma}(q, N) = \theta^{N-1}\exp(-q\theta)[q^N/\Gamma(N)],$$

$$p(\boldsymbol{\beta}_i) = N[\boldsymbol{\beta}, \boldsymbol{\Omega}],\ p(\boldsymbol{\beta}) = \text{"flat"} \propto 1,\ p(\alpha) \propto 1,$$

$$p(\boldsymbol{\Omega}) = \text{inverted Wishart } \propto |\boldsymbol{\Omega}|^{-(K+\nu+1)/2}\exp(-\text{tr}\boldsymbol{\Omega}^{-1}\mathbf{W}/2).$$

Assumed values for the elements of the priors, $s$, $M$, $q$, $N$, $\nu$, and $\mathbf{W}$ are discussed. The prior for $\theta$ is crucial. As Tsionas (2002) notes, $q = -\ln r^*$ is a crucial element in many studies (e.g., Koop et al., 1997). In most of these, the researchers use $r^* = 0.875$, implying a prior median efficiency of 87.5% when $N = 1$ (exponential). Tsionas reports that he used $N = 1$ (exponential prior for $\theta$), but $q = 10^{-6}$, which implies essentially a flat prior for $\theta$ over

the entire positive half line. For the other parameters, he reports prior values $s = 10^{-6}$ and $M = 1$, so $p(\sigma) \propto 1/\sigma$ (approximately), which is a Jeffrey's (noninformative) prior; $\nu = 1$, and $\mathbf{W} = 10^{-6}\mathbf{I}$, so $p(\mathbf{\Omega})$ is almost flat also. An MCMC-based Gibbs sampler is described for estimation. The parameter estimates (posterior means) are reasonable, but the estimated posterior mean for $\theta$ in the full model is 75.12, implying an inefficiency distribution concentrated almost entirely at zero ("near perfect efficiency"—as he notes, estimated efficiencies are almost uniformly above 0.99). Using the same data, van den Broeck et al. (1994) found values ranging from 0.83 to 0.91 when $r^* = 0.875$ and even lower when $r^* = 0.50$. The difference is attributed to the extension of the model to individual-specific parameters. (The classical MLE of $\theta$ is approximately 10, which implies median efficiency of roughly 93%.)

Tsionas (2002) hints at relaxing the exponential distribution assumption for $f(u_{it})$ to allow the more flexible gamma distribution (see Greene, 1993, 2004b), but only suggests how to fit the Erlang form (integer $P$) of the model for $P = 1$ (as above), 2, and 3. Huang (2004) presents a full extension of the model to the gamma distribution for $u_{it}$,

$$f(u_{it}) = u_{it}^{P-1} \exp(-\theta u_{it})[\theta^P / \Gamma(P)].$$

A second extension allows a subset of the parameters to remain equal across firms—Huang (2004) uses this to force the constant term to be the same for all firms, while the other parameters remain firm specific. The author is (ironically) vague about what prior is used for the parameters of $f(u_{it})$. $P$ is taken to have gamma prior with parameters (1, 1)—that is, exponential with mean 1. But, for $q$ in $p(\theta)$, he suggests that he is following van den Broeck et al. (1994) and Tsionas (2002), who use completely different values. A footnote suggests something in the neighborhood of 0.8 is used for $-\ln r^* = q$. Huang's final results do not differ much from Tsionas's. The posterior means for $\theta$ and $P$ are 77.433 (Tsionas found 77.12) and 0.9063 (Tsionas forced the latter to equal 1). Huang (2004) likewise finds posterior estimates of mean efficiency that differ only trivially from 0.99. The values that he finds when he assumes homogeneous technologies are more in line with van den Broeck et al. (1994, their figure 2).

These results are not plausible. I surmise that they result from fitting a separate parameter vector to every observation in a cross section, something that cannot be done with classical, MSL procedures. The Gibbs sampler (MCMC) method has no built-in mechanism that will break down when one attempts to do so. (One could trace the Gibbs samples from different starting points in the chain and look for failures to converge. That does not appear to have been done here.) Consider a classical alternative. In Greene (2005), the random-parameters model

$$\ln y_{it} = \alpha_i + \boldsymbol{\beta}_i^{\mathrm{T}} \mathbf{x}_{it} + v_{it} - u_i,$$
$$(\alpha_i, \boldsymbol{\beta}_i) \sim N[(\alpha, \boldsymbol{\beta}), \Sigma],$$

$$v_{it} \sim N[0, \sigma_v^2],$$
$$u_{it} \sim N[0, \sigma_u^2]$$

is estimated by MSL. (An extension is proposed that allows the mean of the normal distribution to include a term $\Delta \mathbf{z}_i$ which produces a two-level model and adds an additional layer of heterogeneity in the model.) As a general rule, the classical estimator of this (any) random-parameters model does not work very well in a cross section. For the same data used in the preceding two studies, the MSL estimates appear quite reasonable, with the exception of the estimate of $\sigma_v$, which goes nearly to zero. All of the variation in $v_{it}$ is soaked up by the firm-specific parameters, leaving nothing for the idiosyncratic disturbance. (In contrast, in the hierarchical Bayes model, all the variation in $u$ is absorbed elsewhere in the estimated model.) The estimated efficiency values from this model (discussed further in the next section) are 0.984 (the result of a numerical problem in manipulating the near zero value of $\sigma_v$), for every firm in the sample—equally implausible. If the normal–gamma model discussed above, with nonrandom (homogeneous) coefficients, is fit by MSL, the estimated efficiencies from that model (EFFG) produce the kernel density plot shown in figure 2.7. This figure is virtually identical to Huang's (2004) figure 2, which does likewise for the homogeneous technologies model, even including the small second mode near 0.67. To explore the idea suggested above, I divided the sample into 20 size classes and fit the random-parameters model with these 20 groups treated as a panel. The results corresponding to figure 2.7 appear in figure 2.8.

These results are strikingly at odds with the Bayesian estimates. To return to the issue of parameter heterogeneity, note that these are firm-level, not plant-level, data and that most of these firms are fairly large multiplant utilities.

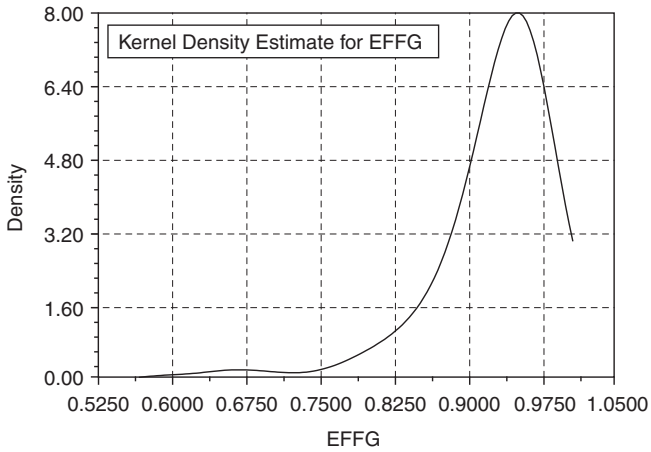


**Figure 2.7.** Estimated Efficiencies for Electric Power Generation

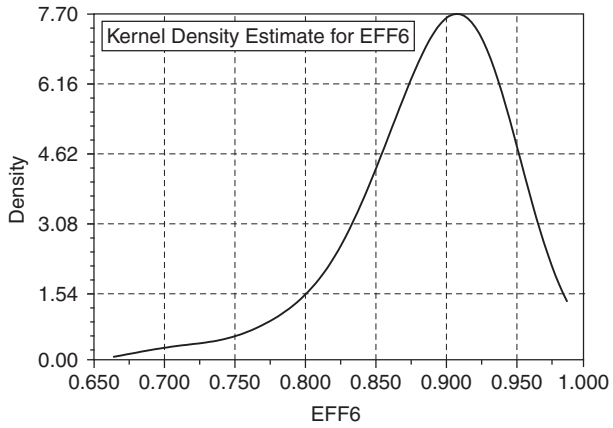


**Figure 2.8.** Efficiencies for Heterogeneous Technologies Model

The proposition that there are the very large differences in technology across firms suggested by the large parameter variances estimated in the heterogeneous parameter models seems dubious. The statistical issue of computing individual-specific coefficients in a cross section and the underlying economics suggest that these results need a much closer look.

### 2.6.5 Location effects on the inefficiency model

Thus far, we have analyzed different approaches to introducing heterogeneity in the technology into the stochastic frontier while retaining the simple additive homogeneous inefficiency term. Let us now turn attention to models that consider the location and scale of the inefficiency itself. Heterogeneity of the sort examined above is also a natural place to focus the discussion. A central issue in this specification search is how to handle the possibility of time variation in inefficiency in a model for panel data. This is considered in section 2.7.

Kumbhakar's (1993) "production risk model,"

$$\ln y_{it} = \alpha + T(\ln \mathbf{x}_{it}, \boldsymbol{\beta}) + g(\mathbf{x}_{it}, \boldsymbol{\delta})\varepsilon_{it},$$

where $g(\mathbf{x}_{it}, \boldsymbol{\delta}) = \sum_k \delta_k x_{itk}$ in a translog model (log-quadratic) and $\varepsilon_{it} = \tau_i + \lambda_t + v_{it}$, is similar. In this case, inefficiency is estimated with $g(\mathbf{x}_{it}, \hat{\delta})(\hat{\alpha} + \hat{\tau}_i) - \max_j[g(\mathbf{x}_{it}, \hat{\delta})(\hat{\alpha} + \hat{\tau}_j)]$.

Whether inefficiency can be appropriately modeled in the preceding fashion is the subject of some debate. Forcing a pattern of any sort on all firms in the sample is restrictive. (Of course, it is less so than assuming there is no variation at all.) Another approach to capturing variation in inefficiency is the addition of a nonnegative effect directly to the production function. Deprins

and Simar (1989b) suggested $E[u|\mathbf{z}_i] = \exp(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}_i)$, which produces the model

$$\ln y_i = \ln f(\mathbf{x}_i, \boldsymbol{\beta}) - \exp(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}_i) + \varepsilon_I,$$

where $E[\varepsilon_i] = 0$. A formal specification for the distribution of $\varepsilon$ completes the model. This approach is somewhat cumbersome analytically, because it loses the essential nature of the nonnegative inefficiency. Kumbhakar, Ghosh, and McGuckin (1991) suggested a similar specification,

$$u_i = \boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}_i + \varepsilon_I,$$

with similar shortcomings. Constraining the sign of $u_i$ is difficult in the specifications considered thus far. Kumbhakar et al.'s (1991) solution has been used in many recent applications:

$$u_i = |N[\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}_i, \sigma_u^2]|$$

Reifschneider and Stevenson's (1991) proposal to address the issue is

$$\ln y_i = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - d(\boldsymbol{\delta}, \mathbf{z}_i) - u_i^* + v_i,$$

where both $d(\boldsymbol{\delta}, \mathbf{z}_i)$ and $u_i^*$ are positive. Specifying $d(\boldsymbol{\delta}, \mathbf{z}_i) = \exp(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}_i)$ and $u_i^* \sim N^+[0, \sigma_u^2]$ satisfies the requirement. This is, of course, similar to Kumbhakar et al.'s (1991) model, but with an additional constraint satisfied. Reifschneider and Stevenson (1991) apply the model with truncated normal, exponential, and Erlang distributions assumed for $u_i$. Actually, the model is overspecified, since it is not necessary for both $\exp(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}_i)$ and $u_i^*$ to be positive for $u_i = u_i^* + \exp(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}_i)$ to be positive. Huang and Liu (1994) complete the specification by formulating the model so that only $u_i^* \geq -\exp(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{z}_i)$ is built into the model. This is done by using a truncated normal rather than a half-normal distribution for $u_i^*$. [In Huang and Liu's formulation, the shift function also includes the levels of the inputs. This has implications for the elasticities as well as some ambiguous implications for whether inefficiency is input or output oriented. Battese and Coelli (1995) propose a specification that is generally similar to that of Huang and Liu (1994).] Since the truncation point enters the log-likelihood function nonlinearly, even for a linear function for $d(\boldsymbol{\delta}, \mathbf{z}_i)$, this substantially complicates the estimation. On the other hand, by manipulating the model a bit, we can show that the Huang and Liu model can be obtained from Stevenson's (1980) truncated-normal model just by replacing $\varepsilon_i$ with $\varepsilon_i + d(\boldsymbol{\delta}, \mathbf{z}_i)$ and $v_i$ with $v_i + d(\boldsymbol{\delta}, \mathbf{z}_i)$—Huang and Liu specified $d_i = \boldsymbol{\delta}'z_i$ for a set of variables that need not appear in the production function. The model proposed by Kumbhakar et al. (1991) is, likewise, equivalent to that of Huang and Liu (1994).

A number of other variations on this theme have been suggested. Battese, Rambaldi, and Wan (1994) use

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + d(\boldsymbol{\delta}, \mathbf{z}_i)(u_i + v_i).$$

Note the use of an additive as opposed to multiplicative disturbance in this model. Battese et al. were interested specifically in modeling $y_i$ in "natural units" (the authors" term). Technical efficiency in this model, defined as

$$\text{TE}_i = \frac{E[y_i|u_i, \mathbf{x}_i]}{E[y_i|u_i = 0, \mathbf{x}_i]} = 1 - \frac{d_i}{f_i}u_i,$$

clearly depends on $d_i$. [Note that since $y_i$ is positive, $\text{TE}_i \in (0, 1)$.] Battese et al. present the log-likelihood function for this model with Cobb-Douglas specifications for $f(\cdot)$ and $d(\cdot)$ in an appendix. Estimation is, at least in principle, straightforward, though the use of the additive form of the disturbance probably unnecessarily complicates the function and the maximization process. There is, however, a purpose to doing so; the main subject of the paper is *production risk*, defined for the $k$th input as

$$\eta_k = \frac{\partial \text{var}[y_i|u_i, \mathbf{x}_i]}{\partial x_{ki}} = 2\beta_k \frac{\text{var}[y_i|u_i, \mathbf{x}_i]}{x_{ki}}$$

for their model.

Last, an intuitively appealing modification of Battese et al.'s (1994) formulation is

$$\ln y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + v_i - d_i u_i,$$

where, as before, $d_i$ is a nonnegative function and $u_i$ has one of the distributions specified above. Suppose that we assume Stevenson's truncated-normal model for $u_i$. Then, by using the change of variable formula, it is easy to show that $d_i u_i$ has a truncated normal distribution, as well; when $r_i = d_i u_i$,

$$h(r_i) = \left[\frac{1}{d_i\sigma_u}\right] \frac{\phi[(r_i - d_i\mu)/d_i\sigma_u]}{\Phi\left[d_i\mu/d_i\sigma_u\right]}.$$

Therefore, the log-likelihood function and all of the subsequent results needed for estimating the technical efficiency values for the observations can be obtained for this model just by replacing $\mu$ with $\mu_i = d_i\mu$ and $\sigma_u$ with $\sigma_{ui} = d_i\sigma_u$ in Stevenson's model. This implies that the transformed parameters, $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$ and $\lambda = \sigma_u/\sigma_v$, will now be functions of $d_i$. An application of this model is Caudill and Ford (1993), who use this formulation with $\mu = 0$ and $d_i = [f(x_i, \boldsymbol{\beta})]^\delta$. This adds a single new parameter to the model, $\delta$. Since the authors are interested in the effects of the heteroskedasticity on the parameter estimates, it remains for subsequent researchers to establish how, if at all, this (and, for that matter, any of the aforementioned models) changes the estimates of $u_i$ and $\text{TE}_i$.

### 2.6.6 Shifting the underlying mean of $u_i$

The discussion thus far [with the exception of Huang and Liu's (1994) model] has treated the distributions of the stochastic component of the frontier, $v_i$, and

the inefficiency component, $u_i$, as homogeneous populations, with constant mean and variance and fixed distribution. Heterogeneity in the model arises only in the inputs (and any other control variables) in the production or cost functions. But, there is ample evidence that both of these characteristics can vary widely across firms, as well.

A natural starting point for accommodating heterogeneity in the inefficiency model is in the location of the distribution. Figure 2.9 shows the form of the density for a *truncated-normal* model for three values of $\mu : -0.5, 0.0$ (the half-normal model), and 0.5. Clearly, the value of $\mu$ makes a considerable difference in the shape of the distribution. Firm-specific heterogeneity can easily be incorporated into the model as follows:

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + v_i - u_i,$$
$$v_i \sim N[0, \sigma_v^2],$$
$$u_i = |U_i|,$$

where $U_i \sim N[\mu_i, \sigma_u^2], \mu_i = \mu_0 + \boldsymbol{\mu}_1'\mathbf{z}_i$.

As noted, this is the same as the reduced form of the model proposed by Huang and Liu (1994). The difference is that, here, the heterogeneity is specifically designed as the location parameter in the underlying distribution. One might include in $\mathbf{z}_i$ industry-specific effects or other technological attributes. For example, an analysis of production in the airline industry might include load factor (the proportion of seat-miles flown that are also passenger-miles, a number that has often varied around 0.75 in this industry). This brings a relatively minor change in the estimator (in principle), though in practice, the numerical properties of the estimator do change considerably. The modified

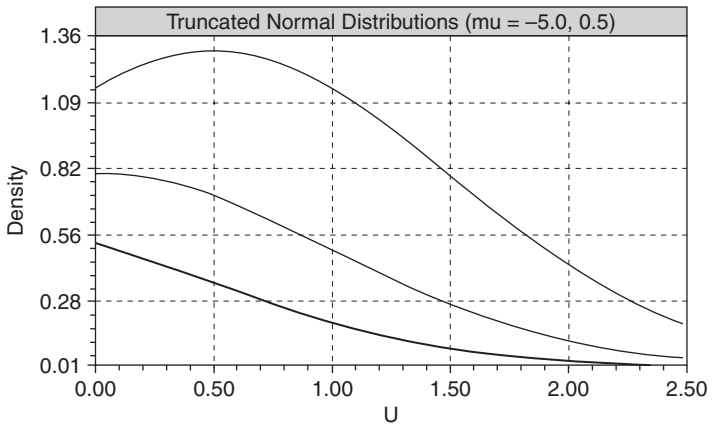


**Figure 2.9.** Truncated Normal Distributions

log-likelihood is now

$$\text{Ln } L(\alpha, \beta, \sigma, \lambda, \mu^0, \mu_1) = -N\left[\ln\sigma + \frac{1}{2}\ln 2\pi + \ln\Phi(\mu_i/\sigma_u)\right] + \sum_{i=1}^{N}\left[-\frac{1}{2}\left(\frac{\varepsilon_i + \mu_i}{\sigma}\right)^2 + \ln\Phi\left(\frac{\mu_i}{\sigma\lambda} - \frac{\varepsilon_i\lambda}{\sigma}\right)\right],$$

where $\lambda = \sigma_u/\sigma_v, \sigma^2 = \sigma_u^2 + \sigma_v^2$ and $\sigma_u = \lambda\sigma/\sqrt{1+\lambda^2}$. The sign of $\varepsilon_i$ is reversed in the two appearances for estimation of a cost or output distance frontier. This is a relatively minor modification of the original normal–half-normal, though the interpretation has changed substantively.

### 2.6.7 Heteroskedasticity

As in other settings, there is no reason to assume that heterogeneity would be limited to the mean of the inefficiency. A model that allows heteroskedasticity in $u_i$ or $v_i$ is a straightforward extension. A convenient generic form would be

$$\text{var}[v_i|\mathbf{h}_i] = \sigma_v^2 g_v(\mathbf{h}_i, \boldsymbol{\delta}), g_v(\mathbf{h}_i, 0) = 1,$$
$$\text{var}[U_i|\mathbf{h}_i] = \sigma_u^2 g_u(\mathbf{h}_i, \boldsymbol{\tau}), g_u(\mathbf{h}_i, 0) = 1$$

(see Reifschneider and Stevenson, 1991; Simar, Lovell, and Eeckhaut, 1994). We have assumed that the same variables appear in both functions, although with suitably placed zeros in the parameter vectors, this can be relaxed. The normalization $g_v(\mathbf{h}_i, \mathbf{0}) = 1$ is used to nest the homoskedastic model within the broader one. The two functions are assumed to be strictly continuous and differentiable in their arguments. In applications, linear functions (e.g., $1 + \boldsymbol{\delta}^{\mathrm{T}}\mathbf{h}_i$) are likely to be inadequate, because they would allow negative variances. Likewise, a function of the form $\sigma_{ui}^2 = \sigma_u^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\delta}$ (Caudill and Fort, 1993; Caudill et al., 1995) does not prevent invalid computations.[61] Reifschneider and Stevenson also suggested $\sigma_{ui}^2 = \sigma_u^2 + g_u(\mathbf{h}_i, \boldsymbol{\tau})$, which requires $g_u(\mathbf{h}_i, \boldsymbol{\tau}) \geq 0$ and $g_u(\mathbf{h}_i, 0) = 0$. A more convenient form is the exponential,

$$g_v(\mathbf{h}_i, \boldsymbol{\delta}) = [\exp(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{h}_i)]^2 \text{ and } g_u(\mathbf{h}_i, \boldsymbol{\tau}) = [\exp(\boldsymbol{\tau}^{\mathrm{T}}\mathbf{h}_i)]^2$$

(Hadri, 1999).[62] For estimation, it is necessary to revert to the parameterization that is explicit in the two variances, $\sigma_{vi}^2$ and $\sigma_{ui}^2$, rather than the form in $\lambda$ and $\sigma^2$, because if either of the underlying variances is heterogeneous, then both of these reduced-form parameters must be heterogeneous, but not in an obvious fashion. The resulting log-likelihood is somewhat cumbersome, but quite manageable computationally. The complication that arises still results

from the heterogeneity in the mean:

$$\begin{aligned}
&\text{Ln } L(\alpha, \beta, \sigma_u, \sigma_v, \delta, \tau, \mu_0, \mu_1) \\
&\quad = -\frac{N}{2}\ln 2\pi + \sum_{i=1}^{N}[\ln \sigma_i] + \sum_{i=1}^{N} \ln \Phi\left[\frac{\mu_i}{\sigma_{ui}}\right] \\
&\qquad + \sum_{i=1}^{N}\left[-\frac{1}{2}\left(\frac{\varepsilon_i + \mu_i}{\sigma_i}\right)^2 + \ln \Phi\left(\frac{\mu_i}{\sigma_i(\sigma_{ui}/\sigma_{vi})} - \frac{\varepsilon_i(\sigma_{ui}/\sigma_{vi})}{\sigma_i}\right)\right]
\end{aligned}$$

$$\begin{aligned}
\sigma_i &= \sqrt{\sigma_v^2[\exp(\delta^{\mathrm{T}}\mathbf{h}_i)]^2 + \sigma_u^2[\exp(\tau^{\mathrm{T}}\mathbf{h}_i)]^2} \\
&= \sqrt{\sigma_{vi}^2 + \sigma_{ui}^2}
\end{aligned}$$

There are numerous applications, including Caudill and Ford (1993), Caudill et al. (1995), Hadri (1999), and Hadri et al. (2003a, 2003b), that specify a model with heteroskedasticity in both error components.

### 2.6.8 The scaling property

Wang and Schmidt (2002) and Alvarez, Amsler, Orea, and Schmidt (2006) suggest a semiparametric approach to accounting for exogenous influences that they label the "scaling property" proposed by Simar et al. (1994). Thus,

$$u_i = u(\delta, \mathbf{z}_i) = h(\delta, \mathbf{z}_i) \times u_i^*,$$

where $u_i^*$ is a nonnegative random variable whose distribution does not involve $\mathbf{z_i}$ and $h(\delta, \mathbf{z_i})$ is a nonnegative function. [For our purposes, we will also require that $h(\delta, \mathbf{z_i})$ be continuous in $\delta$, though this is not strictly required for the model to be internally consistent.] The extension goes beyond heteroskedasticity, because it implies $E[u_i] = h(\delta, \mathbf{z}_i)E[u_i^*]$. Simar et al. (1994) exploit this to develop a nonlinear least squares estimator that does not require a distributional assumption. The likelihood function for the half- or truncated-normal or exponential model is not particularly difficult, however, though it does carry some testable restrictions that relate the mean and variance of the underlying distribution of $u_i$ (for further explorations of the scaling property, see Alvarez, Amsler, Orea, and Schmidt, 2006; see also related results in Bhattacharyya, Kumbhakar, and Bhattacharyya, 1995). Candidates for $h(\delta, \mathbf{z}_i)$ are the usual ones, linear, $\delta^{\mathrm{T}}\mathbf{z}_i$ and exponential, $\exp(\delta'\mathbf{z}_i)$. Simar et al. suggest the truncated normal $N[\mu, \sigma_u^2]^+$ as a convenient specification that carries all of their assumptions. Wang and Schmidt (2002) then provide a lengthy argument why conventional estimators of the production parameters and the JLMS (Jondrow, Lovell, Materov, and Schmidt, 1982) estimates of $u_i$ will be seriously biased. The same arguments apply to estimates of $\text{TE}_i = \exp(-u_i)$.

Several parts of the earlier literature predate Wang and Schmidt (2002). Kumbhakaret al. (1991), Huang and Liu (1994), and Battese and Coelli (1995) have all considered normal–truncated-normal models in which $\mu_i = \delta' \mathbf{z}_i$. Reifschneider and Stevenson (1991), Caudill and Ford (1993), and Caudill, Ford, and Gropper (1995) suggested different functional forms for the variance, such as $\sigma_{ui} = \sigma_u \times \exp(\delta^{\mathrm{T}} \mathbf{z}_i)$. None of these formulations satisfies the scaling property, though a combination does. Let $h(\delta, \mathbf{z}_i) = \exp(\delta^{\mathrm{T}} \mathbf{z}_i)$ and assume the truncated-normal model. Then it follows that $u_i = |U_i|$, $U_i \sim N\{\mu \times \exp(\delta^{\mathrm{T}} \mathbf{z}_i), [\sigma_u \times \exp(\delta^{\mathrm{T}} \mathbf{z}_i)]^2\}$. Wang (2002) proposes a model that specifically violates the scaling property, $\sigma_{ui}^2 = \exp(\delta^{\mathrm{T}} \mathbf{z}_i)$ and $\mu_i = \delta^{\mathrm{T}} \mathbf{z}_i$, to examine nonneutral shifts of the production function. Alvarez, Amsler, Orea and Schmidt (2006) examine the various specifications that are encompassed by this formulation. We note, in all these specifications, the underlying mean of $u_i$ is functionally related to the variance. This is a testable restriction.

The scaling property adds some useful dimensions to the stochastic frontier model. First, it allows firm heterogeneity to show up by shrinking or inflating the inefficiency distribution without changing its basic shape. Second, if $u_i = h(\delta, \mathbf{z}_i) \times u_i^*$, then $\partial \ln u_i / \partial \mathbf{z}_i = \partial \ln h(\delta, \mathbf{z}_i) / \partial \mathbf{z}_i$ irrespective of the underlying distribution of $u_i{}^*$. Third, in principle, the model parameters can be estimated by nonlinear least squares without a distributional assumption (for discussion, see Alvarez et al., 2006; Kumbhakar and Lovell, 2000). Given the robustness of estimates of $u_i$ explored in the preceding section, we suspect that this is a minor virtue. Moreover, the full model with this assumption built into it is not a major modification of the normal–truncated-normal model already considered, though it does have one built in ambiguity that we now explore. We will allow both $u_i$ and $v_i$ to be exponentially heteroskedastic. The log-likelihood for Wang and Schmidt's (2002) model is then

$$\operatorname{Ln} L(\alpha, \beta, \delta, \gamma, \mu^0) = -(N/2) \ln 2\pi - \sum_{i=1}^{N} [\ln \sigma_i + \ln \Phi(\mu_i / \sigma_{ui})] + \sum_{i=1}^{N} \left[ -\frac{1}{2} \left( \frac{\varepsilon_i + \mu_i}{\sigma_i} \right)^2 + \ln \Phi \left( \frac{\mu_i}{\sigma_i \lambda_i} - \frac{\varepsilon_i \lambda_i}{\sigma_i} \right) \right]$$

where

$\mu_i = \mu \exp(\delta^{\mathrm{T}} z_i)$, $\sigma_{ui} = \sigma_u \exp(\delta^{\mathrm{T}} z_i)$, $\sigma_{vi} = \sigma_v \exp(\gamma^{\mathrm{T}} z_i)$,
$\lambda_i = \sigma_{ui} / \sigma_{vi}$, and $\sigma_i = \sqrt{\sigma_{vi}^2 + \sigma_{ui}^2}$.

We allow for $\sigma_{ui}$ and $\sigma_{vi}$ to differ by placing zeros in the parameter vectors where needed to allow different variables to appear in the functions. Note that there is a set of equality restrictions built into the model, across $\mu_i$ and $\sigma_{ui}$. Also, though $\sigma_u$ and $\sigma_v$ must be positive [they could be written as $\exp(\delta_0)$ and $\exp(\gamma_0)$ to impose this], $\mu$ must be allowed to have either sign.

Wang and Schmidt (2002) provide a Monte Carlo analysis of the biases that result if the scaling property is ignored. The point is well taken. It is also useful, however, to view the scaling property as an alternative specification of the stochastic frontier model that may or may not be consistent with the data (i.e., the underlying population). The assumption is easy to test, because we can simply relax the equality restriction that links the mean and the standard deviation of the distribution. (Note, however, that there remains what is probably a minor restriction in the Wang and Schmidt model, that with or without the scaling property imposed, the specification of the mean does not allow for a linear shift of $\mu_i$ independent of $\mathbf{z}_i$; there is no free constant term in the equation for $\mu_i$. For this reason, even with the exponential specification for the variances, the truncation model is usually specified with $\mu_i = \mu_0 + \delta^{\mathrm{T}}\mathbf{z}_i$.)

## 2.7 Panel-Data Models

When producers are observed at several points in time, three shortcomings in the foregoing analysis can be handled explicitly.[63] In the stochastic frontier model, it is necessary to assume that the firm-specific level of inefficiency is uncorrelated with the input levels. This may be unwarranted. Some of the specifications discussed above (e.g., Huang and Liu, 1994) reconstructed the inefficiency term in the model to include functions of $\mathbf{x}$. The assumption of normality for the noise term and half- or truncated normality for the inefficiency, while probably relatively benign, is yet another assumption that one might prefer not to make. A few alternatives are noted in the preceding. Nonetheless, under certain assumptions, more robust panel-data treatments are likely to bring improvements in the estimates. A fundamental question concerns whether inefficiency is properly modeled as fixed over time. The point is moot in a cross section, of course. However, it is very relevant in the analysis of panel data. Intuition should suggest that the longer the panel, the "better" will be the estimator of time-invariant inefficiency in the model, however computed. But, at the same time, the longer the time period of the observation, the less tenable the assumption becomes. This is a perennial issue in this literature, without a simple solution.

In this section, we will detail several models for inefficiency that are amenable to analysis in panel-data sets. Actual computation of estimators of $u_i$ or $u_{it}$ are considered in section 2.8. Treatments of firm and time variation in inefficiency are usefully distinguished in two dimensions. The first, as mentioned, is whether we wish to assume that it is time varying or not. Second, we consider models that make only minimal distributional assumptions about inefficiency ("fixed-effects" models) and models that make specific distributional assumptions such as those made above: half-normal, exponential, and so forth. The former have a virtue of robustness, but this comes at a cost of a downward bias (see Kim and Schmidt, 2000). The latter make possibly restrictive assumptions but bring the benefit of increased precision.

There are $N$ firms and $T_i$ observations on each. (It is customary to assume that $T_i$ is constant across firms, but this is never actually necessary.) If observations on $u_{it}$ and $v_{it}$ are independent over time as well as across individuals, then the panel nature of the data set is irrelevant and the models discussed above will apply to the pooled data set. But, if one is willing to make further assumptions about the nature of the inefficiency, a number of new possibilities arise. We consider several of them here.

### 2.7.1 Time variation in inefficiency

A large proportion of the research on panel-data applications analyzes (essentially) a deterministic frontier in the context of "fixed-effects" models:

$$\ln y_{it} = \alpha + \beta^{\mathrm{T}}\mathbf{x_{it}} + \mathbf{a_i} + \mathbf{v_{it}},$$

where $a_i$ is the fixed effect normalized in some way to accommodate the nonzero constant. This regression style model is typically identified with the fixed-effects linear model. If the $u_i$ values are treated as firm-specific constants, the model may be estimated by OLS, as a "fixed-effects" model (using the "within-groups" transformation if the number of firms is too large to accommodate with simple OLS).[64] It is more useful here to think in terms of the specification being distribution free, particularly in view of the Bayesian treatments in which the distinction between "fixed" and "random" effects is ambiguous. Development of this model begins with Schmidt and Sickles (1984), who propose essentially a deterministic frontier treatment of estimated inefficiencies

$$\hat{u}_i = \max_j(\hat{a}_j) - \hat{a}_i.$$

Sickles (2005) presents a wide variety of approaches and interpretations of this model. Note the assumption of time invariance in this treatment. One individual in the sample is fully efficient ($u_i = 0$), and others are compared to it, rather than to an absolute standard.

This fixed-effects approach has the distinct advantage of dispensing with the assumption that the firm inefficiencies are uncorrelated with the input levels. Moreover, no assumption of normality is needed. Finally, this approach shares the consistency property of the deterministic frontier model in the estimation of $u_i$. This estimate is consistent in $T_i$, which may, in practice, be quite small. But, as noted above, this consistency may make no economic sense—the longer the time interval, the less tenable the time invariance assumption.

An extension suggested by Kumbhakar (1990, 1993) is to add a "time" effect, $\gamma_t$, to the fixed-effects model. Strictly in terms of estimation, the statistical properties of $c_t = \hat{\gamma}_t$ depend on $N$, which is more likely to be amenable to conventional analyses. This can be treated as a fixed- or random-effects model, of course. In either case, it is necessary to compensate for the presence of the time effect in the model. However, since the time effect is the same for all

firms in each period, the earlier expression for $\hat{u}_i$ would now define $\hat{u}_{it}$ rather than $\hat{u}_i$. This does relax the assumption of time invariance of the production function, but it does not add to the specification of the inefficiency term in the model. The modification considered next does.

Kumbhakar and Hjalmarsson (1995) also suggested a precursor to Greene's (2004a) true fixed- and random-effects models. They proposed

$$u_{it} = \tau_i + a_{it}$$

where $a_{it} \sim N^+[0, \sigma^2]$. They suggested a two-step estimation procedure that begins with either OLS/dummy variables or feasible GLS and proceeds to a second-step analysis to estimate $\tau_i$. (Greene's estimators are full information MLEs that use only a single step.) Heshmati and Kumbhakar (1994) and Kumbhakar and Heshmati (1995) consider methodological aspects of these models, including how to accommodate technical change.

Cornwell et al. (1990) propose to accommodate systematic variation in inefficiency, by replacing $a_i$ with

$$a_{it} = \alpha_{i0} + \alpha_{i1} t + \alpha_{i2} t^2.$$

Inefficiency is still modeled using $u_{it} = \max(a_{it}) - a_{it}$. With this modified specification, the most efficient firm can change from period to period. Also, since the maximum (or minimum, for a cost frontier) is period specific and need not apply to the same firm in every period, this will interrupt the quadratic relationship between time and the inefficiencies. [Kumbhakar (1990) proposes some modifications of this model and examines the implications of several kinds of restrictions on the parameters.] The signature feature of this formulation is that inefficiency for a given firm evolves systematically over time. Cornwell et al. (1990) analyze the estimation problem in this model at some length (see also Hausman and Taylor, 1981). For large $N$, this presents a fairly cumbersome problem of estimation (note 65 notwithstanding). But, for data sets of the size in many applications, this enhanced fixed-effects model can comfortably be estimated by simple, unadorned least squares.[65] Alternative approaches are suggested by Kumbhakar (1991a), Kumbhakar and Heshmati (1995), and Kumbhakar and Hjalmarsson (1995). Applications are given by, for example, Cornwell et al. (1990), Schmidt and Sickles (1984), and Gong and Sickles (1989), who apply the preceding to a cost frontier model; and Good, Roller, and Sickles (1993, 1995), Good and Sickles (1995), and Good, Nadiri, Roller, and Sickles (1993), who use various models to analyze the airline industry. Some semiparametric variations on the Cornwell et al. (1990) approach are discussed in Park, Sickles, and Simar (1998) and Park and Simar (1992).

Lee and Schmidt (1993) proposed a less heavily parameterized fixed-effects frontier model,

$$\ln y_{it} = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_{it} + a_{it} + v_{it},$$

where $a_{it} = \theta_t a_i, \theta_t$ is a set of time dummy variable coefficients, and, as before, $\hat{u}_{it} = \max_i(\hat{\theta}_t \hat{a}_i) - \hat{\theta}_t \hat{a}_i$. This differs from the familiar fixed-effects model, $\alpha_{it} = \theta_t + \delta_i$, in two ways. First, the model is nonlinear and requires a more complicated estimator. Second, it allows the unobserved firm effect to vary over time. A benefit is that time-invariant firm effects may now be accommodated. The authors describe both fixed- and random-effects treatments of $\theta_t$.

Numerous similarly motivated specifications have been proposed for the stochastic frontier model

$$\ln y_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_{it} + v_{it} - u_{it}.$$

Two that have proved useful in applications are Kumbhakar's (1990) model,

$$u_{it} = u_i / [1 + \exp(\gamma_1 t + \gamma_2 t^2)],$$

and Battese and Coelli's (1992) formulation (which is the model of choice in many recent applications),

$$u_{it} = u_i \times \exp[-\eta(t - T)].$$

An alternative formulation that allows the variance effect to be nonmonotonic is

$$u_{it} = u_i \times \exp[\eta_1(t - T) + \eta_2(t - T)^2].$$

In all formulations, $u_i = |U_i| \sim N^+[0, \sigma_u^2]$. The Battese and Coelli model has been extended to the truncated-normal model, as well, in Kumbhakar and Orea (2004) and Greene (2004a). Cuesta (2000) also proposed a modification, with firm-specific scale factors, $\eta_i$, in the scaling function. The authors present full details on the log-likelihood, its derivatives, and the computation of $E[u|\varepsilon]$ and $E[\mathrm{e}^{-u}|\varepsilon]$.

Tsionas (2003) proposed an autoregressive model in which inefficiency evolves via an autoregressive process:

$$\ln u_{it} = \gamma^{\mathrm{T}} \mathbf{z}_{it} + \rho \ln u_{i,t-1} + w_{it}$$

(Specific assumptions are also made for the initial value, which would be important here because the typical panel in this setting is very short.) The autoregressive process embodies "new sources of inefficiency." In Tsionas's Bayesian MCMC treatment, the Gibbs sampler is used to draw observations directly from the posterior for $u_{it}$ rather than using the JLMS (Jondrow, Lovell, Materov, and Schmidt, 1982) estimator ex post to estimate firm- and time-specific inefficiency. In an application to U.S. banks, he finds the posterior mean $\rho = 0.91$, which implies that the process is almost static. The implied short- and long-run efficiencies for the whole sample range around 60%.

### 2.7.2 Distributional assumptions

If the assumption of independence of the inefficiencies and input levels can be maintained, then a random-effects approach brings some benefits in precision.

One advantage of the random-effects model is that it allows time-invariant firm-specific attributes, such as the capital stock of a firm that is not growing, to enter the model. The familiar random-effects regression model is easily adapted to the stochastic frontier model. There are also some interesting variants in the recently received literature. We consider the basic model first.

The relevant log-likelihood for a random-effects model with a half-normal distribution has been derived by Lee and Tyler (1978) and Pitt and Lee (1981) and is discussed further by Battese and Coelli (1988).[66] The truncated-normal model of Stevenson (1980) is adopted here for generality and in order to maintain consistency with Battese and Coelli (1988), who provide a number of interesting results for this model. The half-normal model can be produced simply by restricting $\mu$ to equal 0. We also allow the underlying mean to be heterogeneous. Note, as well, that the extension of the model to include multiplicative heteroskedasticity in $v_{it}$ and/or $u_{it}$ is straightforward. I omit this detail in the interest of brevity. The structural model is, then,

$$
\begin{aligned}
y_{it} &= \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it},\\
u_i &\sim |N[\mu_i, \sigma_u^2]|\\
v_{it} &\sim N[0, \sigma_v^2].
\end{aligned}
$$

As before, there are $T_i$ observations on the $i$th firm. Battese and Coelli (1988) and Battese, Coelli, and Colby (1989) have extensively analyzed this model, with Stevenson's extension to the truncated normal distribution for $u_i$. They also provide the counterpart to the JLMS (Jondrow et al., 1982) estimator of $u_i$. With our reparameterization, their result is

$$
\begin{aligned}
E[u_i|\varepsilon_{i,1}, \varepsilon_{i,2}, \ldots, \varepsilon_{i,T_i}] &= \mu_i^* + \sigma_{i^*}\left[\frac{\phi(\mu_i^*/\sigma_{i^*})}{\Phi(-\mu_i^*/\sigma_{i^*})}\right],\\
\mu_i^* &= \gamma_i\mu + (1-\gamma_i)(-\bar{\varepsilon}_i),\\
\varepsilon_{it} &= y_{it} - \alpha - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it},\\
\gamma_i &= 1/(1+\lambda T_i),\\
\lambda &= \sigma_u^2/\sigma_v^2,\\
\sigma_{i^*}^2 &= \gamma_i\,\sigma_u^2\,.
\end{aligned}
$$

As $T_i \to \infty, \gamma_i \to 0$, and the entire expression collapses to $-\bar{\varepsilon}_i$, which in turn converges to $u_i$, as might be expected. As Schmidt and Sickles (1984) observe, this can be interpreted as the advantage of having observed $u_i N$ times. Taking the mean averages out the noise contained in $v_{it}$, which only occurs once. It is worth noting that the preceding, perhaps with the simplifying assumption that $\mu = 0$, could be employed after estimation of the random-effects model by GLS, rather than ML. The aforementioned corrections to the moment-based variance estimators would be required, of course. Battese and Coelli (1988,

1992) have also derived the panel-data counterpart to $E[e^{-u_i|\varepsilon_i}]$,

$$E[\exp(-u_i)|\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT_i}] = \left[\frac{\Phi[(\mu_i^*/\sigma_{*i}) - \sigma_{*i}]}{\Phi(\mu_i^*/\sigma_{*i})}\right] \exp[-\mu_{i^*} + \tfrac{1}{2}\sigma_{i^*}^2].$$

### 2.7.3 Fixed-effects, random-effects, and Bayesian approaches

An obstacle to the fixed-effects approach is the presence of time-invariant attributes of the firms. If the model is conditioned on firm attributes such as the capital stock, location, or some other characteristics, and if these do not vary over time, then the Least Squares Dummy Variable estimator cannot be computed as shown above. Worse yet, if these effects are simply omitted from the model, then they will reappear in the fixed effects, masquerading as inefficiency (or lack of), when obviously they should be classified otherwise. The question is one of identification. Hausman and Taylor (1981) have discussed conditions for identification of such effects and methods of estimation that might prove useful. However, the economic content of this discussion may be at odds with the algebraic and statistical content. Regardless of how fixed effects enter the model discussed above, they will reappear in the estimates of inefficiency and thereby create some ambiguity in the interpretation of these estimates. As such, alternative treatments, such as the random-effects model, may be preferable.

Consider the base case model,

$$\ln y_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_i,$$

where either the fixed-effects interpretation described above or the random-effects model described in the next subsection is adopted. The time-invariant element in the model, $u_i$, is intended to capture all and only the firm-specific inefficiency. If there are time-invariant effects, such as heterogeneity, in the data, they must also appear in $u_i$ whether they belong there or not. [This is exactly the opposite of the argument made by Kumbhakar and Hjalmarsson (1995), who argued that the time-varying $v_{it}$ would inappropriately capture *time-varying* inefficiency.] In analyzing the WHO panel data on 191 countries, Greene (2004b) argued that under either interpretation, $u_i$ would be absorbing a large amount of cross-country heterogeneity that would inappropriately be measured as inefficiency. He proposed a "true" fixed-effects model,

$$\begin{aligned}
\ln y_{it} &= \alpha_i + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it},\\
v_{it} &\sim N[0, \sigma_v^2],\\
u_{it} &\sim |N[0, \sigma^2]|,
\end{aligned}$$

which is precisely a normal–half-normal (or truncated-normal or exponential) stochastic frontier model with the firm dummy variables included. This model is a very minor extension of existing models that nonetheless has seen little

use. Most panel-data sets in this literature are quite small, so the number of dummy variable coefficients is usually manageable—Greene (2003a, 2004a, 2004b, 2005) shows how to fit the model with large numbers of firms, but in point of fact, in common applications where there are only a few dozen firms or so, this is trivial. The formulation does assume that inefficiency varies randomly across time, however, which is the substantive change in the model. As noted above, the "truth" is probably somewhere between these two strong assumptions. Greene (2005) also considers a "true random-effects" model that modifies the parameterized models in the next section, in this case, a random-parameters (random-constant) model

$$\begin{aligned}
\ln y_{it} &= (\alpha + w_i) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it},\\
v_{it} &\sim N[0, \sigma_v^2],\\
u_{it} &\sim |N[0, \sigma_u^2]|,\\
w_i &\sim \text{with mean 0 and finite variance.}
\end{aligned}$$

[Note this is the complement to Huang's (2004) model, which contained a homogeneous constant term and heterogeneous slope vectors.] This model is fit by MSL methods. Farsi et al. (2003) examined the relative behavior of these four models in a study of Swiss nursing homes and noted a preference for the true random-effects specification.

The preceding sections have noted a variety of applications of Bayesian methods to estimation of stochastic frontier models. As the next section turns specifically to the estimation of inefficiencies, $u_{it}$, or efficiency, $\exp(-u_{it})$, note the essential component of the Bayesian approach. Koop, Osiewalski, and Steel (1997; KOS) and numerous references cited there and above lay out this method. For the "fixed-effects" approach,[67] the model is simply a linear regression model with firm dummy variables. The Bayesian inference problem is simply that of the linear model with normal disturbances and $K + N$ regressors. The posterior mean estimator of the slopes and constants are the least squares coefficients and

$$\hat{\alpha}_i = \bar{y}_i - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\bar{\mathbf{x}}_i.$$

Estimation of $u_i$ is done in precisely the same fashion as it classical counterpart:

$$\hat{u}_i = (\max_j \hat{\alpha}_j) - \hat{\alpha}_i$$

(Posterior variances and statistical inference are considered in the next section.) Thus, with noninformative priors, the Bayesian estimators are identical to their classical counterparts (of course). For the "random-effects" approach, $u_i$ values are treated as missing data. The technique of data augmentation is used for estimation. The simplest model specification, based on

KOS (see also Tsionas, 2002), would be

$$\begin{aligned}
\ln y_{it} &= \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - \mathbf{u}_i,\\
p(v_{it}) &= N[0,\sigma^2], p(\sigma) = \text{inverted gamma}(s,M)\\
&\propto \exp\left[-s/(2\sigma^2)\right](\sigma^2)^{-(M+1)/2},\\
p(u_{it}) &= (1/\lambda)\exp(-u_i/\lambda), p(\lambda) \text{ to be determined},\\
p(\boldsymbol{\beta}) &= \text{“flat”} \propto 1, p(\alpha) \propto 1.
\end{aligned}$$

The Gibbs sampler for this model is quite simple—see KOS for details. The controversial part is the necessary informative prior density for $\lambda$. Kim and Schmidt (2000) describe several alternative approaches.

We note that this formulation has the same two shortcomings as its classical counterpart. The “fixed-effects” form cannot accommodate any time-invariant covariates. Neither the fixed- nor the random-effects form has any provision for unmeasured time-invariant heterogeneity. The true fixed- and random-effects models could also be accommodated in this framework. For the true fixed-effects model,

$$\begin{aligned}
\ln y_{it} &= \alpha_i + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it},\\
p(v_{it}) &= N[0,\sigma^2], p(\sigma) = \text{inverted gamma}(s,M),\\
p(u_{it}) &= (1/\lambda)\exp(-u_i/\lambda), p(\lambda) \text{ to be determined},\\
p(\boldsymbol{\beta}) &= \text{“flat”} \propto 1, p(\alpha_i) \propto 1.
\end{aligned}$$

This is KOS’s random-effects form with a complete set of firm-specific constants and inefficiency both firm and time varying. The joint posterior would involve the $N + K$ regression parameters, $\sigma$, $\lambda$, and all $NT$ missing values $u_{it}$. With an ample data set, this is essentially the same as KOS’s random-effects model—the dimensionality of the parameter space increases (dramatically so), but the computations are the same. One advantage of this formulation is that the difficult inference problem associated with $\hat{u}_i = (\max_j \hat{\alpha}_j) - \hat{\alpha}_i$ is avoided. For the true random-effects model, we would have

$$\begin{aligned}
\ln y_{it} &= \alpha_i + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} + w_i - u_{it},\\
p(v_{it}) &= N[0,\sigma^2], p(\sigma) = \text{inverted gamma}(s,M),\\
p(w_i) &= N[0,\tau^2], p(\tau) = \text{inverted gamma}(r,T),\\
p(u_{it}) &= (1/\lambda)\exp(-u_{it}/\lambda), p(\lambda) \text{ to be determined},\\
p(\boldsymbol{\beta}) &= \propto 1, p(\alpha) \propto 1.
\end{aligned}$$

Note that $w_i$ is time invariant. Once again, this is the same as KOS’s random-effects model. Now, the data augmentation problem is over $N+NT$ dimensions for the values of $w_i$ and $u_{it}$.

## 2.8 Estimation of Technical Inefficiency

Arguably, the main purpose of the preceding is to lay the groundwork for estimation of inefficiency, that is, $u_i$ or $\text{TE}_i = \exp(-u_i)$. For example, along with an abundance of DEA applications, regulators in many fields have begun to employ efficiency analyses such as those discussed here to compare and analyze regulated firms and to set policies (see, e.g., the discussion in chapter 1). Bauer et al. (1998) sounded some cautions for regulators who might use the results of the analysis described here in their oversight of financial institutions. Among others, regulators are keenly interested in these techniques. This section describes methods of using the estimated models to estimate technical inefficiency in the stochastic frontier setting.

The core model we have used to reach this point is

$$\ln y_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it},$$

where we allow for observations across firms and time. The various model forms, including normal–half-normal, truncation, exponential, and gamma models with heterogeneity of various sorts, panel-data and cross-section treatments, and Bayesian and classical frameworks, have all provided platforms on which the main objective of the estimation is pursued: analysis of inefficiency, $u_{it}$.

The estimation effort in the preceding sections is all prelude to estimation of the inefficiency term in the equation, $u_{it}$, or some function of it. Note, before continuing, a variety of obstacles to that objective. Foremost is the fundamental result that the inefficiency component of the model, $u_{it}$, must be observed indirectly. In the model as stated, the data and estimates provide only estimates of $\varepsilon_{it} = v_{it} - u_{it}$. We will have to devise a strategy for disentangling these, if possible. Second, regardless of how we proceed, we must cope not only with a noisy signal ($v_{it}$ with $u_{it}$), but we must acknowledge estimation error in our estimate—$\alpha$ and $\boldsymbol{\beta}$ are not known with certainty. For the "fixed-effects" estimators, estimation of technical inefficiency is only relative. Recall in this setting, the model is

$$\ln y_{it} = \alpha_i + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it},$$

and the estimator is $u_i = \max(\alpha_j) - \alpha_i$. Finally, adding to the difficult estimation problem is the complication of devising a method of recognizing a degree of uncertainty in the estimate. A point estimate of $u_{it}$ may be insufficient. It is one thing to suggest that inefficiency is on the order of 10%, but quite another to suggest that one's best guess is from 0 to 35% with a mean of 10%, which conveys considerably less information.

This section details some known results about estimation of technical inefficiency. [Kim and Schmidt (2000) is a useful source for more complete presentation.] I consider both fixed-effects and stochastic frontier estimators, and briefly visit Bayesian as well as the classical estimator. The focus of the

discussion is the workhorse of this branch of the literature, Jondrow et al.'s (1982) conditional mean estimator.

### 2.8.1 Estimators of technical inefficiency in the stochastic frontier model

However the parameters of the model are computed, the residual, $y_{it} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_{it}$ estimates $\varepsilon_{it}$, not $u_{it}$. The standard estimator of $u_{it}$, is the conditional mean function, $E[u_{it}|\varepsilon_{it}]$. Recall

$$f(u_{it}|\varepsilon_i) = \frac{f(u_{it},\varepsilon_{it})}{f(\varepsilon_{it})} = \frac{f(u_{it})f(\varepsilon_{it}|u_{it})}{f(\varepsilon_{it})}$$
$$= \frac{f_u(u_{it})f_v(\varepsilon_{it}+u_{it})}{\int_0^\infty f_u(u_{it})f_v(\varepsilon_{it}+u_{it})du_{it}}.$$

We will use as the estimator the conditional mean from the conditional distribution,

$$E(u_{it}|\varepsilon_{it}) = \frac{\int_0^\infty u_{it}f_u(u_{it})f_v(\varepsilon_{it}+u_{it})du_{it}}{\int_0^\infty f_u(u_{it})f_v(\varepsilon_{it}+u_{it})du_{it}}.$$

In several leading cases, this function has a known form.[68] JLMS (Jondrow, Lovell, Materov, and Schmidt, 1982) present the explicit result for the half-normal model,

$$E[u_{it}|\varepsilon_{it}] = \left[\frac{\sigma\lambda}{1+\lambda^2}\right]\left[\tilde{\mu}_{it} + \frac{\phi(\tilde{\mu}_{it})}{\Phi(\tilde{\mu}_{it})}\right], \tilde{\mu}_{it} = \frac{-\lambda\varepsilon_{it}}{\sigma},$$

where $\phi(.)$ and $\Phi(.)$ are the density and CDF of the standard normal distribution. For the truncated-normal model, the result is obtained by replacing $\tilde{\mu}_{it}$ with $\tilde{\mu}_{it} + \mu\sigma_u^2/\sigma^2$. The corresponding expressions for the exponential and gamma models are

$$E\left[u_{it}|\varepsilon_{it}\right] = z_{it} + \sigma_v\frac{\phi(z_{it}/\sigma_v)}{\Phi(z_{it}/\sigma_v)}, z_{it} = \varepsilon_{it} - \sigma_v^2/\sigma_u$$

and

$$E[u_{it}|\varepsilon_{it}] = \frac{q(P,\varepsilon_{it})}{q(P-1,\varepsilon_{it})},$$

respectively.[69] Note, for the gamma model, that this must be computed by simulation (see section 2.4.6). The estimator of $\mathrm{u}_{it}$ in the random-parameters model (Greene, 2005) is also computed by simulation.

Battese and Coelli (1988) suggested the alternative estimator

$$E[\mathrm{TE}t_i|\varepsilon_{it}] = E[\exp(-u_{it})|\varepsilon_{it}].$$

For the truncated-normal model (which includes the half-normal case), this is

$$E[\exp(-u_{it})|\varepsilon_{it}] = \frac{\Phi[(\mu^*_{it}/\sigma_*) - \sigma_*]}{\Phi[(\mu^*_{it}/\sigma_*)]} \exp\left[-\mu^*_{it} + \frac{1}{2}\sigma^2_*\right],$$

where
$\mu^*_{it} = \tilde{\mu}_{it} + \mu\sigma^2_u/\sigma^2$.

Recalling the approximation $u_{it} \approx 1 - \text{TE}_{it}$, Battese and Coelli (1988) observed that the difference between the two estimates reflects the inaccuracy of the approximation $1 - u_{it}$ to $\exp(-u_{it})$. For the JLMS results, Battese and Coelli report the alternative estimate based on the preceding as 8.9% as opposed to 9.6% for the conditional mean. The difference is fairly small in this instance, but many studies report technical efficiency values considerably less than the 90% they reported.

Some additional results that are useful for the practitioner are, first, that estimation of cost inefficiency based on the preceding results can be accomplished simply by changing the sign of $\varepsilon$ where it appears in any expression. Second, for analysis of a panel with time-invariant effects, Kim and Schmidt (2000) point out that one need only replace $\varepsilon$ with the group mean $< \varepsilon\text{-overbar} >_i$, and $\sigma^2_v$ with $\sigma^2_v/T$ to obtain the appropriate results. Finally, in all these expressions, the heterogeneous or heteroskedastic models may be imposed simply by employing the firm-specific formulations for $\mu$, $\sigma_u$, and/or $\sigma_v$. In what follows, I limit attention to a simple model and leave the extensions in these directions to be employed by the practitioner as needed.

### 2.8.2 Characteristics of the estimator

I digress at this point to note some possible misperceptions in the literature (abetted, alas, by the Greene [1993] survey). The JLMS estimator is unbiased as an estimator of $u_{it}$ only in the sense that it has the same expectation that $u_{it}$ does. It does not estimate $u_{it}$ unbiasedly in the sense that in repeated sampling, the mean of a set of observations on $E[u_{it}|\varepsilon_{it}]$ would equal $u_{it}$. They would not. First, the notion of repeated sampling is itself inconsistent with the definition, since the estimator is conditioned on a specific set of data. This does not mean that it is conditioned on the data for a particular firm—it is conditioned on a specific $y_{it}$ and $\mathbf{x}_{it}$. To see the logic of this objection, consider that there is nothing inconsistent in doing the analysis on a sample that contains two firms that have identical $y_{it}$ and $\mathbf{x}_{it}$, but different $u_{it}$. Indeed, that is precisely the point of the stochastic frontier model. Thus, the estimator $E[u_{it}|\varepsilon_{it}]$ is an estimator of the mean of the distribution that produces these two observations with this particular $y_{it}$ and $\mathbf{x}_{it}$. There is nothing in its construction that makes us expect the JLMS estimator to be an *unbiased* estimator of either one of the two hypothetical draws on $u_{it}$. It is an estimator of the mean of

this conditional distribution from which both are drawn. The empirical estimator based on the ML or Bayesian parameter estimates is not unbiased for $E[u_{it}|\varepsilon_{it}]$ either, by the way, since it is a nonlinear function of the parameter estimators.

In general, the empirical estimator of $E[u_{it}|\varepsilon_{it}]$ is a consistent estimator, for the usual reasons. Again, it is a conditioned on a particular $(y_{it}, \mathbf{x}_{it})$, so the JLMS estimator is not based on a sample of one at all. The estimator, computed using the MLEs or random draws from the posterior, converges to the true conditional mean function. That is,

$$\text{plim}\left\{\left[\frac{\hat{\sigma}\hat{\lambda}}{1+\hat{\lambda}^2}\right]\left(\hat{\mu}_i + \frac{\varphi(\hat{\mu}_{it})}{\Phi(\hat{\mu}_{it})}\right)\middle|\hat{\mu}_{it} = \frac{-\left(y_{it} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_{it}\right)\hat{\lambda}}{\hat{\sigma}}\right\} = E[u_{it}|\varepsilon_{it}]\,.$$

The JLMS estimator is not a consistent estimator of $u_{it}$ either, again for the reasons noted above. No amount of data can reveal $u_{it}$ perfectly in the stochastic frontier model. It can in a panel-data model in which it is assumed that $u_{it}$ is time invariant, since then, like any common-effects model, a method-of-moments estimator of the "effect" is consistent in $T$. But, absent this possibility, the JLMS estimator does not converge to $u_{it}$. (And, note once again, the idea that $u_i$ would remain the same long enough for asymptotics with respect to $T$ to apply would require some difficult economic justification.) On the other hand, it does converge to something: $E[u_{it}|y_{it}, \mathbf{x}_{it}, \boldsymbol{\beta}, \ldots]$. But, again, this is not $u_{it}$; it is the mean of the distribution from which $u_{it}$ is generated.

The foregoing extends to Bayesian estimation, as well, notwithstanding the reflexive assertion that Bayesian inference is "exact" (while classical inference is inexact) that now fashionably precedes every Bayesian analysis. Indeed, a closer look at the Bayesian estimator of $u_{it}$ in the "standard model" is revealing. In the Gibbs sampling MCMC iterations in the standard, normal–exponential model, we have

$$p(u_{it}|\boldsymbol{\beta}, \sigma_v^2, \sigma_u, y_{it}, \mathbf{x}_{it}) = \text{truncated at zero } N\left[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} - \mathbf{y}_{it} - \sigma_v^2/\sigma_u, \sigma_v^2\right]$$

(see Koop et al., 1995, p. 357). The Gibbs sampler draws observations from this distribution in the cycle. At any particular draw, a close look at this distribution shows that its conditional mean is precisely the JLMS estimator for the exponential distribution, for the specific values of the parameters at that cycle. (It would seem pathological if anything else emerged, since the estimator is, ultimately, the posterior mean in the conditional distribution.) Does this estimator "converge" to something? Ultimately, we hope to sample from the marginal posterior $p(u_{it}|\mathbf{x}_{it}, \mathbf{y}_{it})$, but clearly at every step on the way, this is going to be computed using draws that employ some model parameters. So, where does this end up? With noninformative priors, the Bayesian posterior means of the parameters are going to converge to the same point that the MLEs converge to. (This is merely the well-known result that, with noninformative

priors, the likelihood function must eventually dominate the posterior, and the mode of the likelihood converges to the posterior mean as the sample size grows without bound.) The end result of this argument is that the Bayesian estimators of $u_{it}$ are based on the same estimator as the classical estimator. The former estimates $E[u_{it}|\varepsilon_{it}]$ by sampling from $p(u_{it}|y_{it}, \mathbf{x}_{it}, E[\boldsymbol{\beta}|\text{data}], \text{etc.})$, while the latter computes the function directly using the MLEs. Therefore, the Bayesian estimator, like the classical one, is not a consistent estimator of $u_{it}$. Nor is it unbiased, again, for the same reasons.

Aside from what are likely to be minor numerical differences, the Bayesian and classical estimators differ in two other respects. [See Kim and Schmidt (2000) and the above study of the gamma model for examples.] The sampling variance of the MLE-based estimator will be larger than its Bayesian counterpart, for the usual reasons. However, this raises another important similarity. The difference in the sampling behavior of the statistics speaks to the behavior of the statistic, not to a difference in the quantity being estimated. That is, we have yet to examine $\text{var}[u_{it}|\varepsilon_{it}]$, which is the same regardless of how we choose to estimate it. Kim and Schmidt (2000) have examined this in detail. I briefly note their results below. A remaining aspect of this comparison, however, concerns how "confidence intervals" for $u_{it}$ are constructed. A natural (at least intuitively appealing) strategy is to use the (admittedly flawed) estimator $E[u_{it}|\varepsilon_{it}]$ and bracket it with two lengths of an estimate of $(\text{var}[u_{it}|\varepsilon_{it}])^{1/2}$. This is essentially what the classical estimators discussed below attempt to do. Indeed, it is becoming customary in both settings to report both the point and variance estimators. But, for the Bayesian estimator, this is a demonstrably suboptimal strategy. Since, for example, for the exponential model, it is known exactly that the posterior density is truncated normal, which is asymmetric, an "HPD interval" containing the usual 95% of the distribution will be less than four standard deviations wide. Bayesian estimators typically include pictures of the posterior density for a few observations. Delineation of the HPD intervals in these densities might be a useful addition to the reported results.

### 2.8.3 Does the distribution matter?

Many authors have pursued the question in the section title, in cross-section and panel-data sets, across a variety of platforms. Most of the received results suggest the encouraging finding that the estimates of inefficiency are reasonably robust to the model specification. Since all results are application specific, however, the question does not have an analytical answer. In order to continue that thread of discussion, we consider a small example, following along the presentation in Kumbhakar and Lovell (2000). The authors report estimates based on the cost frontier estimated in Greene (1990) using the Christensen and Greene (1976) electricity data (described further below). Kumbhakar and Lovell obtained rank correlations for estimates of inefficiencies from the four distributions examined above that ranged from a low of 0.7467 (exponential and gamma) to a high of 0.9803 (half-normal and truncated normal). The

results below based on these same data are considerably stronger. I suspect that at least two reasons lie behind this: First, the results below are based on a full translog model, which is probably a more appropriate functional form—Christensen and Greene (1976) found likewise; second, the simulation-based estimator for the gamma model appears to be a considerably better algorithm than the brute force method used in the above studies. We also fit a production function, rather than a cost function.

The first application (there are several others to follow in later sections) is an extension of Christensen and Greene's (1976) estimates of a translog cost function for U.S. electricity generation in 1970. The sample consists of data on 123 American (fossil-fueled) electric-generating companies. The data set contains the variables described in table 2.1. The authors (and Kumbhakar and Lovell, 2000) used these data to fit a translog cost function in a single output (generation) based on three inputs: capital, labor, and fuel. I obtained physical input figures from the cost, factor shares, and input prices and then used these data to fit a translog production function.

Data on physical units are obtained by computing $x_k, = \text{cost} \times \text{share}_k/\text{price}_k$. The translog production function is then

$$\ln y = \alpha + \sum_{k=K,L,F} \beta_k \ln x_k + \sum_{k=K,L,F} \sum_{m=K,L,F} \gamma_{km} \ln x_k \ln x_m + v_i - u_i.$$

Estimates of the parameters of the stochastic frontier functions under various assumptions are shown in table 2.2. (The data are not normalized.) The OLS and method-of-moments estimators for the variance parameters are given in the first column. For brevity, standard errors and indicators of significance are omitted, because at this point the purpose is only to compare the coefficient estimates. Based on the differences in the parameter estimates in table 2.2,

**Table 2.1**
Descriptive Statistics for Christensen and Greene Electricity Data (123 Observations)

| Variable | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|
| Cost | 48.467 | 64.0636 | 0.1304 | 282.9401 |
| Output | 9501.146 | 12512.82 | 4.0 | 72247.0 |
| Capital price | 72.895 | 9.516264 | 39.127 | 92.65 |
| Capital share | 0.22776 | 0.060103 | 0.0981 | 0.4571 |
| Labor price | 7988.560 | 1252.83 | 5063.49 | 10963.9 |
| Labor share | 0.14286 | 0.0563198 | 0.0527 | 0.03291 |
| Fuel price | 30.80661 | 7.928241 | 9.0 | 50.4516 |
| Fuel share | 0.62783 | 0.088789 | 0.02435 | 0.08136 |
| Capital | 0.14397 | 0.19558 | 0.000179168 | 1.28401 |
| Labor | 0.074440 | 0.00098891 | 0.000004341821 | 0.00490297 |
| Fuel | 1.00465 | 1.28670 | 0.002641465 | 6.9757 |

**Table 2.2**
Estimated Stochastic Frontier Production Functions

| Parameter | OLS | Half-Normal | Truncated | Exponential | Gamma |
|---|---|---|---|---|---|
| $\alpha$ | 5.381 | 6.199 | 7.137 | 7.136 | 7.037 |
| $\beta K$ | 1.364 | 1.272 | 1.296 | 1.299 | 1.335 |
| $\beta L$ | −1.403 | −1.174 | −0.910 | −0.911 | −0.942 |
| $\beta F$ | 1.327 | 1.224 | 0.978 | 0.976 | 0.964 |
| $\gamma KK$ | 0.479 | 0.469 | 0.397 | 0.394 | 0.346 |
| $\gamma LL$ | −0.204 | −0.170 | −0.139 | −0.139 | −0.148 |
| $\gamma FF$ | 0.319 | 0.316 | 0.301 | 0.300 | 0.276 |
| $\gamma KL$ | 0.051 | 0.041 | 0.065 | 0.066 | 0.084 |
| $\gamma KF$ | −0.581 | −0.562 | −0.502 | −0.500 | −0.463 |
| $\gamma LF$ | 0.204 | 0.185 | 0.133 | 0.132 | 0.120 |
| $\lambda$ | 0.218 | 0.956 | 15.791 | 0.806 | 1.071 |
| $\sigma$ | 0.127 | 0.144 | 1.481 | 0.120 | 0.133 |
| $\mu$ | NA | NA | −29.084 | NA | NA |
| $P$ | NA | NA | NA | NA | 0.674 |
| $\sigma_u$ | 0.02714 | 0.0995 | 1.477 | 0.0750 | 0.097 |
| $\sigma_v$ | 0.1244 | 0.1041 | 0.0936 | 0.0931 | 0.0906 |
| **Ln L** | 85.996 | 86.292 | 88.186 | 88.209 | 88.849 |

it does appear that the functional form matters considerably. However, the estimates of $E[u_i|\varepsilon_i]$ tell a quite different story. Figure 2.10 and tables 2.3 and 2.4 show that the JLMS estimates of $u_i$ are almost identical. The agreement between the exponential and truncated-normal model is striking. Note that the correlation matrix shows both raw correlations among the estimates and rank correlations based on the ranks of the inefficiencies. These results are considerably closer than those found by Kumbhakar and Lovell (2000). The parameter estimates for the truncated-normal model do look extreme. In fact, judging from the estimate of $\sigma^2$, the truncated-normal model has considerably altered the results. The estimate of $\sigma$ is 1.481 compared to 0.144 for the half-normal model, a 10-fold increase. The very large estimate of $\mu$ suggests, in turn, that the inefficiency estimates should be drastically affected, but this turns out not to be the case. The argument of the function $E[u|\varepsilon]$ can be written as $[a(-\varepsilon) + (1 - a)\mu]$, where $a = \sigma_u^2/\sigma^2$. For these data, $a$ is roughly 0.9945, so, in fact, $\mu$ hardly figures into the estimated inefficiencies at all. The kernel density estimate in figure 2.11 based on the estimates of $u_i$ for the truncated-normal model is essentially identical to that obtained for the other three sets of estimates. The estimated residuals, $e_i$, from the truncation model look essentially the same as for the other distributions, as well. We conclude, based on this example, as noted above, that the estimated inefficiencies seem quite robust to the model specification.

We note, finally, a caution about figure 2.11 (and counterparts in many other studies, such as the nearly identical figure 2 in Huang, 2004): The density

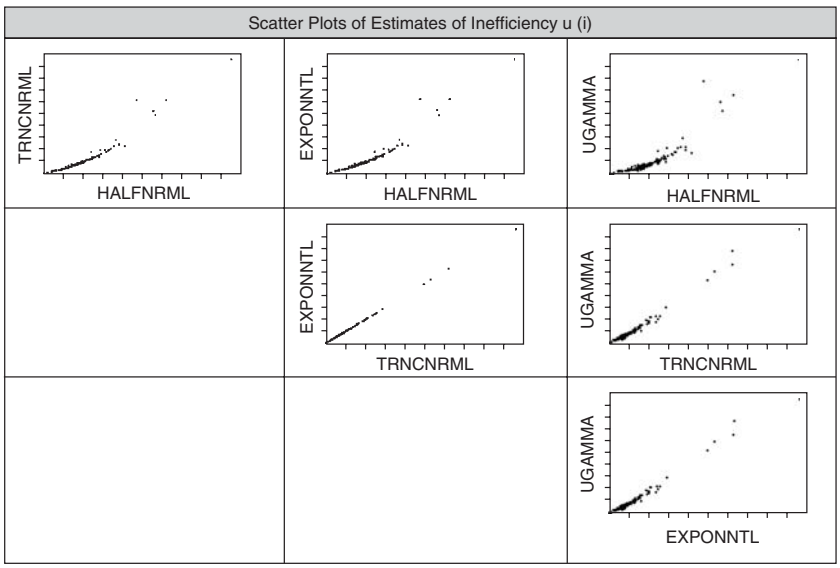


**Figure 2.10.** Scatter Plots of Inefficiencies from Various Specifications

**Table 2.3**
Descriptive Statistics for Estimates of $E[u_i|\varepsilon_i]$ (123 Observations)

| Distribution | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|
| Half-normal | 0.07902 | 0.03246 | 0.02630 | 0.27446 |
| Truncated normal | 0.07462 | 0.05936 | 0.01824 | 0.47040 |
| Exponential | 0.07480 | 0.06001 | 0.01810 | 0.47324 |
| Gamma | 0.06530 | 0.06967 | 0.01136 | 0.49552 |

**Table 2.4**
Pearson and Spearman Rank Correlations for Estimates of $E[u_i|\varepsilon_i]$[a]

| | Half-Normal | Truncated Normal | Exponential | Gamma |
|---|---|---|---|---|
| Half-normal | 1.00000 | 0.99280 | 0.99248 | 0.95540 |
| Truncated normal | 0.95291 | 1.00000 | 0.99994 | 0.96864 |
| Exponential | 0.95158 | 0.99998 | 1.00000 | 0.96897 |
| Gamma | 0.91163 | 0.98940 | 0.99019 | 1.00000 |

[a] Pearson correlations below diagonal; Spearman rank correlations above diagonal.

estimator above shows the distribution of a sample of estimates of $E[u_i|\varepsilon_i]$, not a sample of estimates of $u_i$. (Huang's figures are correctly identified as such.) The mean of the estimators is correctly suggested by the figure. However, the spread of the distribution of $u_i$ is understated by this figure. In this bivariate distribution, $\text{var}(E[u_i|\varepsilon_i]) = \text{var}[u_i] - E(\text{var}[u_i|\varepsilon_i])$. There

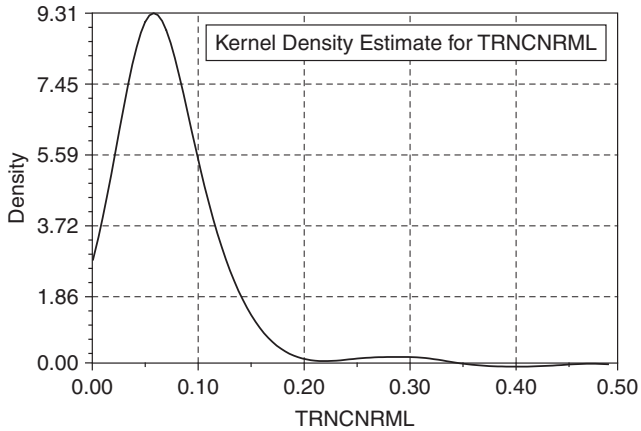


**Figure 2.11.** Kernel Density Estimator for Mean Efficiency

is no reason to expect the latter term in this expression to be small. We can get a quick suggestion of the extent of this error with a small numerical experiment. We take the normal–half-normal results as the basis. We take the estimates in table 2.2 as if they were the true parameters. Thus, $\sigma_u = 0.0995$, $\sigma_v = 0.1041$, $\sigma = 0.144$, $\lambda = 0.9558$. Derived from these, we have $E[u] = \sigma_u\phi(0)/\Phi(0) = 0.07939$, $\text{var}[u] = \sigma_u^2(\pi - 2)/\pi = 0.003598$, $\text{var}[\varepsilon] = \sigma_v^2 + \text{var}[u] = 0.0144$. Now, using $E[u|\varepsilon] = \text{JLMS}(\varepsilon)$ as given above, a function of $\varepsilon$, we use the delta method to approximate the variance of $E[u|\varepsilon]$. This value based on the results above is 0.008067, so the standard deviation is 0.0284 which is roughly the standard deviation of the data shown in the kernel density estimator above. (This value approximates the 0.03246 in the table.) However, the unconditional standard deviation of $u$, which is what we actually desire, is the square root of 0.003598, or about 0.05998. The upshot is that, as this example demonstrates, descriptive statistics and kernel density estimators based on the JLMS estimators correctly show the expectation of $u$ but underestimate the variation. Ultimately, a quick indication of the extent is suggested by $\lambda$; the smaller is $\lambda$, the worse the distortion will be.[70]

### 2.8.4 Confidence intervals for inefficiency

Horrace and Schmidt (1996, 2000) suggest a useful extension of the JLMS result. Jondrow et al. (1982) have shown that the distribution of $u_i|\varepsilon_i$ is that of an $N[\mu_i^*, \sigma^*]$ random variable, truncated from the left at zero, where $\mu_i^* = -\varepsilon_i\lambda^2/(1 + \lambda^2)$ and $\sigma^* = \sigma\lambda/(1 + \lambda^2)$. This result and standard results for the truncated normal distribution (see, e.g., Greene, 2003a) can be used to obtain the conditional mean and variance of $u_i|\varepsilon_i$. With these in hand, one can construct some of the features of the distribution of $u_i|\varepsilon_i$ or

$E[\text{TE}_i|\varepsilon_i] = E[\exp(-u_i)|\varepsilon_i]$. The literature on this subject, including the important contributions of Bera and Sharma (1999) and Kim and Schmidt (2000), refers generally to "confidence intervals" for $u_i|\varepsilon_i$. For reasons that will be clear shortly, we will not use that term—at least not yet, until we have made more precise what we are estimating.

For locating $100(1-\alpha)\%$ of the conditional distribution of $u_i|\varepsilon_i$, we use the following system of equations:

$$\begin{aligned}
\sigma^2 &= \sigma_v^2 + \sigma_u^2 \\
\lambda &= \sigma_u/\sigma_v \\
\mu_i^* &= -\varepsilon_i\sigma_u^2/\sigma^2 = -\varepsilon_i\lambda^2/(1+\lambda^2) \\
\sigma^* &= \sigma_u\sigma_v/\sigma = \sigma\lambda/(1+\lambda^2) \\
LB_i &= \mu_i^* + \sigma^*\Phi^{-1}\left[1 - \left(1 - \frac{\alpha}{2}\right)\Phi\left(\mu_i^*/\sigma^*\right)\right] \\
UB_i &= \mu_i^* + \sigma^*\Phi^{-1}\left[1 - \frac{\alpha}{2}\Phi\left(\mu_i^*/\sigma^*\right)\right]
\end{aligned}$$

Then, if the elements were the true parameters, the region $[\text{LB}_i, \text{UB}_i]$ would encompass $100(1-\alpha)\%$ of the distribution of $u_i|\varepsilon_i$. Although the received papers based on classical methods have labeled this a *confidence interval* for $u_i$, I emphatically disagree. It is a range that encompasses $100(1-\alpha)\%$ of the probability in the conditional distribution of $u_i|\varepsilon_i$. The range is based on $E[u_i|\varepsilon_i]$, not $u_i$ itself. This is not a semantic fine point. Once again, note that, in a sample that contains two identical observations in terms of $y_i$, $\mathbf{x_i}$, they could have quite different $u_i$, yet this construction produces the same "interval" for both. The interval is "centered" at the estimator of the conditional mean, $E[u_i|\varepsilon_i]$, not the estimator of $u_i$ itself, as a conventional "confidence interval" would be. The distinction is more transparent in a Bayesian context. In drawing the Gibbs sample, the Bayesian estimator is explicitly sampling from, and characterizing the conditional distribution of, $u_i|\varepsilon_i$, not constructing any kind of interval that brackets a particular $u_i$—that is not possible.[71] For constructing "confidence intervals" for $\text{TE}_i|\varepsilon_i$, it is necessary only to compute $\text{TE UB}_i = \exp(-\text{LB}_i)$ and $\text{TE LB}_i = \exp(-\text{UB}_i)$.

These limits are conditioned on known values of the parameters, so they ignore any variation in the parameter estimates used to construct them. Thus, we regard this as a minimal width interval.[72]

### 2.8.5 Fixed-effects estimators

Numerous recent studies have approached efficiency analysis with a fixed-effects model. This frees the analyst from having to layer a distributional

assumption on the inefficiency element of the model. The model departs from the usual assumptions,

$$y_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_i.$$

We define $\alpha_i = \alpha - u_i$, then

$$y_{it} = \alpha_i + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it},$$

then $u_i = \alpha - \alpha_i$. Estimation proceeds via least squares, Interest in this setting focuses on the relative inefficiencies, which are estimated via

$$\begin{aligned}
\hat{\alpha}_i &= \bar{y}_i - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\bar{\mathbf{x}}_i \\
&= \alpha_i + \bar{\mathbf{v}}_i - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\bar{\mathbf{x}}_i; \\
\hat{\alpha} &= \max_j(\hat{\alpha}_j) \\
\hat{\mathbf{u}}_i &= \hat{\alpha} - \hat{\alpha}_i.
\end{aligned}$$

Technically, efficiency is estimated in the usual fashion via $\mathrm{TE}_i = \exp(-u_i)$. By construction, these estimates are relative to the most efficient (at least estimated as such) firm in the sample, whereas the estimates of $u_i$ in the stochastic frontier model are absolute—relative to zero, that is. Horrace and Schmidt (1996, 2000) and Kim and Schmidt (2000) describe methods of computing asymptotic variances and doing statistical inference for these MCB ("multiple comparisons with the best") estimators.

The MCB computations are reasonably complicated. Kim and Schmidt (2000) suggest an alternative bootstrapping procedure (see also Horrace and Richards, 2005). The procedure departs a bit from familiar applications of the bootstrap. The model is fit using the full observed sample as specified above. Bootstrap iterations are constructed by then resampling from the estimated normal distribution of $v_{it}$ and computing the bootstrap sample, $y_{it}^{(b)} = \hat{\alpha}_i + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_{it} + \hat{v}_{it}^{(b)}$. The sample of sets of estimates of $u_i$ are used to make inference about the distributions of inefficiency. They also note that these estimators are prone to a downward bias. Kim and Schmidt (2000) propose an adjustment to deal with the bias.

There is a fundamental difference between this approach and the one detailed above for the stochastic frontier model. In this case, the estimator is $\hat{u}_i = \hat{\alpha} - \hat{\alpha}_i$, not $E[u_i|\varepsilon_i]$. There is no "noise" in this estimator. Thus, the "confidence interval" in this setting is for $u_i$, not for the mean of the distribution that generates $u_i$. But, it must be borne in mind that the $u_i$ underlying the computations is only relative to the (estimate of the) minimum $u_i$ in the sample.

### 2.8.6 The Bayesian estimators

Koop et al. (1997) describe procedures for Bayesian estimation of both "fixed-effects" and "random-effects" models for technical inefficiency. We have detailed both of these above. In the fixed-effects approach, they merely add the firm-specific intercepts to the classical normal regression model; the posterior means are the usual within-groups estimators. The posterior distribution is, however, multivariate $t$, rather than multivariate normal. Since the number of degrees of freedom in any reasonable data set will be sufficiently large to render the posterior essentially normal, it follows that the Bayesian estimators of $\alpha_i$ are the same as the classical ones, as will be the confidence intervals. For the comparisons to the best, $\hat{u}_i = \max_j(\hat{\alpha}_j) - \hat{\alpha}_i$, "exact" inference will be difficult, because the precise distribution will remain complicated even though the marginal posterior is known. However, simple Monte Carlo analysis can be used to reveal characteristics such as the percentiles of the distribution for each $\hat{u}_i$. The authors do note that although a similar analysis can be used for $\text{TE}_i = \exp(-\hat{u}_i)$, this estimator will have a built in downward bias. No simple solution is proposed. For the "random-effects," that is, stochastic frontier model, the Gibbs sampler with data augmentation described above is used both for point estimation of $E[u_i|\varepsilon_i]$ and for interval estimation—both mean and variance (and quantiles) of the conditional distribution are computed during the MCMC iterations, so no post estimation processing, other than arranging the sample data, is necessary.

### 2.8.7 A comparison

Kim and Schmidt (2000) compared the several estimators discussed above in four applications. Consistent with the experiment in section 2.8.3, they found that the different estimators do tell very similar stories. The overall conclusions are that Bayesian and classical estimators of comparable models give comparable results, and by and large, fixed-effects estimators produce greater inefficiency estimates than random effects. Based on the above results and Greene (2004a, 2004b), I would conjecture that at least some of this is due to the role of latent heterogeneity that is not otherwise accommodated in the model. This would be, of course, subject to further investigation.

## 2.9 Allocative Inefficiency and the Greene Problem

A classic application of the theory of the preceding discussion is the Averch and Johnson (1955) hypothesis that rate-of-return regulation of electric utilities in the United States in the 1950s led to "gold plating" of physical facilities. Utilities were alleged (by economists, at least) to be wasting money on excessively capitalized facilities. This type of inefficiency would clearly fall under what we have labeled "economic inefficiency" and would fall outside the scope of technical inefficiency. *Allocative inefficiency* refers to the extent to which the

input choices fail to satisfy the marginal equivalences for cost minimization or profit maximization. The essential element of a complete model would be a cost or profit function with a demand system in which failure to satisfy the optimization conditions for profit maximization or cost minimization, irrespective of what happens on the production side, translates into lower profits or higher costs. The vexing problem, which has come to be known as the "Greene problem" (see the first edition of this survey, Greene, 1993), is the formulation of a complete system in which the demands are derived from the cost function by Shephard's lemma, or a profit function by Hotelling's (1932) lemma, and in which deviations from the optimality conditions in any direction translate to lower profits or higher costs.

Several approaches to measuring allocative inefficiency based on cost functions and demand systems have been suggested. See Greene (1993), for details on some of these early specifications. Lovell and Sickles (1983), for example, analyze a system of output supply and input demand equations. Unfortunately, no method is yet in the mainstream that allows convenient analysis of this type of inefficiency in the context of a fully integrated frontier model. [See Kumbhakar and Tsionas (2004, 2005a) for some significant progress in this direction. Kumbhakar and Lovell (2000, chapter 6) also discuss some of the elements of what would be a complete model.] Some of these models are based on the notion of shadow prices and shadow costs—the nonoptimality of the input choices is taken to reflect "optimality" with respect to the "wrong" or "shadow" prices. Some of the early work in this direction is detailed in the 1993 edition of this survey. A recent study that takes this approach is Atkinson, Fare, and Primont (2003).

Research in this area that would lead to a convenient mainstream methodology remains at an early stage (note the aforementioned), so I leave for the next version of this survey to lay out the details of the emerging research.

## 2.10 Applications

In order to illustrate the techniques described above, this section presents some limited studies based on several widely traveled data sets. The Christensen and Greene (1976) electricity generation cross-section data have been used by many researchers, particularly those studying Bayesian methods. A small panel-data set from the pre-deregulation U.S. domestic airline industry (admittedly now quite outdated) that was previously used (e.g., in Kumbhakar, 1991a, 1991b) provides a convenient device for illustrating some of the more straightforward fixed- and random-effects techniques.[73] The banking data set used by Kumbhakar and Tsionas (2002) and by Tsionas and Greene (2003) provides a larger, more homogeneous panel that we can use to study some of the more recently proposed panel-data techniques. Finally, WHO (2000) panel-data set on health care attainment has been used by numerous researchers

for studying different approaches to efficiency modeling (e.g., Evans et al., 2000a, 2000b; Greene, 2004b; Gravelle et al., 2002a, 2002b; Hollingsworth and Wildman, 2002). For our purposes, these data are a well-focused example of a heterogeneous panel.

As noted in the introduction to this chapter, the body of literature on stochastic frontier estimation is very large and growing rapidly. There have been many methodological innovations in the techniques, some of which have "stuck" and are now in the broad range of tools familiar to practitioners, and others of which have proved less popular. The range of computations in this section is very far from exhaustive. The purpose here is only to illustrate some of the most commonly used methods, not to apply the complete catalogue of the empirical tools that appear in the literature. This section begins with a description of computer programs that are currently used for frontier estimation. Subsequent subsections provide some additional details on the data sets and the series of applications.

As noted above, the applications are based on four well-known data sets. This section provides some additional details on the data. The actual data sets are available from my home page (http://http://www.stern.nyu.edu/∼wgreene) in the form of generic *Excel* spreadsheet (.xls) files and *LIMDEP* project (.lpj) files. The *LIMDEP* command sets used to generate the results are also posted so that interested readers can replicate the empirical results.[74] Each of the four applications illustrates some particular aspect or feature of stochastic frontier estimation. We begin with a basic stochastic cost frontier model estimated for U.S. electric power generators.

### 2.10.1 Computer software

The analysis presented below is carried out using version 8 of *LIMDEP* (Econometric Software, Inc., 2000). Some of the techniques are available in other packages. Of course, least squares and variants thereof can be handled with any econometrics program. Basic panel-data operations for linear regression models (linear fixed- and random-effects estimation) can be carried out with most econometrics packages, such as *SAS* (SAS Institute, Inc., 2005), *TSP* (TSP International, 2005), *RATS* (Estima, 2005), *Stata* (Stata, Inc., 2005), *LIMDEP*, *EViews* (QMS, 2005), or *Gauss* (Aptech Systems, Inc., 2005). Low-level languages such as *Matlab*, *Gauss*, *S-plus*, *Fortran*, and *C++* can be used to carry out most if not all of the computations described here, but contemporary, commercially available software obviates the extensive programming that would be needed for most techniques.[75]

In specific terms, the contemporary software offers essentially the following: *TSP* supports the basic cross-section version of the normal–half-normal stochastic frontier model. The cross-sectional version of the stochastic frontier model is actually quite straightforward and, for example, is easily programmed with *Matlab*, *Gauss*, *R*, or *Fortran*, or even with the command languages in *Stata*, *LIMDEP*, or *TSP*. Coelli's (1996) *Frontier 4.1* also handles

a few additional cross-section and panel-data variants of the stochastic frontier model.[76] To date, only two general econometrics programs, Stata and *LIMDEP/NLOGIT*, contain as supported procedures more than the basic stochastic frontier estimator. *Stata* provides estimators for the half- and truncated-normal and exponential models for cross sections (with heteroskedasticity in the distribution of $u_i$), and panel-data variants for the Battese and Coelli (1992, 1995) specifications. *LIMDEP* and *NLOGIT* include all of these and a variety of additional specifications for heteroskedasticity and heterogeneity for cross sections and numerous additional panel-data specifications for fixed-effects, random-effects, random-parameters, and latent class models.

The preceding are all single-equation methods and estimators. Simultaneous estimation of a cost function and a demand system based on a multivariate normal distribution for all disturbances presents no particular obstacle with modern software (once again, *TSP*, *LIMDEP*, *RATS*, *Gauss*). But, there is no general-purpose program yet available for handling a properly specified system of cost and demand equations that would estimate both technical and allocative inefficiency (i.e., solve the Greene problem). Kumbhakar and Tsionas (2004, 2005a, 2005b, 2005c) used *Gauss* for Bayesian and classical estimation of their technical/allocative inefficiency analysis systems.

There seems to be no general-purpose software for Bayesian estimation of stochastic frontier models. A few authors have offered downloadable code; for example, Griffin and Steel (2004) provide in "zipped" format some *C++* code that users will have to compile on their own computer systems. O'Donnell and Griffiths (2004) offer their *Matlab* code. Other Bayesian estimators appear to have been based on Gauss or Matlab code and the freely distributed *WinBugs* (MRC, 2005) package. As a general proposition, there seem to be a great variety of ad hoc strategies adopted more or less "on the fly" (e.g., Metropolis Hastings algorithms for intractable integrals, experimentation with different priors to see how they affect the results, different strategies for using specified draws from the Markov chain). The lack of a general-purpose program such as *Frontier* seems natural.[77] I did find a reference to "BSFM: a Computer Program for Bayesian Stochastic Frontier Models" by Arickx et al. (1997), but no later reference seems to appear. Since nearly all of the development of Bayesian estimators for stochastic frontier model has occurred after 1997, this is of limited usefulness unless it has been further developed. For better or worse, practitioners who opt for the Bayesian approach to modeling will likely be using their own custom-written computer code.

### 2.10.2 The stochastic frontier model: electricity generation

The Christensen and Greene (1976) data have been used by many researchers to develop stochastic frontier estimators, both classical and Bayesian. The data are a 1970 cross section of 123 American electric utilities.[78] The main outcome variables are generation (output) in billions of kilowatt hours and

total cost ($million) consisting of total capital, labor, and fuel cost incurred at the generation stage of production. Other variables in the data set are the prices and cost shares for the three inputs. Remaining variables, including logs and squares and products of logs, are derived. The basic data set is described in table 2.1.

#### 2.10.2.1 Cost frontier model specification

The original Christensen and Greene (1976) study was centered on a translog cost function. Many recent applications use the Cobb-Douglas model for the goal function, though the translog function is common, as well. (Variables are provided in the data set for the full translog model for the interested reader.) In order to provide a comparison with numerous other estimates that appear in the recent literature, we will analyze a homothetic, but not homogeneous, version of the cost function, constructed by adding a quadratic term in log output to the Cobb-Douglas cost function:

$$\ln(C/P_F) = \beta_1 + \beta_2 \ln(P_K/P_F) + \beta_3 \ln(P_L/P_F) + \beta_4 \ln y + \beta_5(1/2 \ln^2 y) + \varepsilon$$

This is the form used in the Bayesian analyses discussed below and the applications discussed above.

#### 2.10.2.2 Corrected and modified least squares estimators

OLS estimates of the cost function are presented in the first column of table 2.5. Linear homogeneity in prices has been imposed by normalizing cost, $P_K$ and $P_L$, by the price of fuel, $P_F$. The functional form conforms to a homothetic but

**Table 2.5**
Estimated Stochastic Cost Frontiers (Standard Errors in Parentheses)

| Variable | Least Squares | Half-Normal | Exponential | Gamma |
|---|---|---|---|---|
| Constant | −7.294 (0.344) | −7.494 (0.330) | −7.634 (0.327) | −7.652 (0.355) |
| Ln $P_K/P_F$ | 0.0748 (0.0616) | 0.0553 (0.0600) | 0.0332 (0.0586) | 0.0293 (0.0656) |
| Ln $P_L/P_F$ | 0.2608 (0.0681) | 0.2606 (0.0655) | 0.2701 (0.0632) | 0.2727 (0.0678) |
| Ln $y$ | 0.3909 (0.0370) | 0.4110 (0.0360) | 0.4398 (0.0383) | 0.4458 (0.0462) |
| $1/2 \ln^2 y$ | 0.0624 (0.00515) | 0.0606 (0.00493) | 0.0575 (0.00506) | 0.0568 (0.00604) |
| $\lambda = \sigma_u/\sigma_v$ | NA | 1.373 | NA | NA |
| $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$ | 0.1439 | 0.1849 | NA | NA |
| $\theta$ | NA | NA | 10.263 | 8.425 |
| $P$ | NA | NA | 1.000 | 0.6702 |
| $\sigma_u$ (= $1/\theta$ for exp.) | 0.1439 | 0.1494 | 0.09742 | 0.09716 |
| $\sigma_v$ | NA | 0.1088 | 0.1044 | 0.1060 |
| Log-likelihood | 66.4736 | 66.8650 | 67.9610 | 68.1542 |

not homogeneous production function (see Christensen and Greene, 1976). Economies of scale ES in this model are measured by the scale elasticity:

$$\text{ES} = \{1/[\beta_4 + \beta_5 \ln y]\} - 1$$

The estimated coefficients are in line with expectations. The theoretical values for the capital, labor, and fuel coefficients are the factor shares. The sample averages of 0.23, 0.14, and 0.63 are reasonably consistent with the coefficients of 0.07, 0.26, and 0.67. Scale economies range from 100% for the smallest firms to minus 10% for the very largest—the mean is 0.13, which is roughly what was observed in the original study based on a translog model.

The least squares residuals, $e_i$, provide implied estimates of inefficiencies. Under the assumption of a *deterministic* frontier, we can obtain estimates of $u_i$ by adjusting the intercept of the estimated production or cost function until all residuals (save one that will be zero) have the correct sign. Thus, for the production frontier, $\hat{u}_i = \max_i(e_i) - e_i$ while for a cost frontier, $\hat{u}_i = e_i - \min_i(e_i)$. Under the deterministic frontier interpretation, the mean, $\bar{u}$, and variance, $s^2$, of the derived estimates of $u_i$ can be used to compute method-of-moments estimators for the underlying parameters. The moment equations and estimates of the various parameters are reported in table 2.6. The sample mean and variance provide conflicting estimates of $\theta$ for the exponential distribution. We use a GMM estimator to reconcile them. The sample mean $\bar{u}$ was used as an initial consistent estimator, $\gamma^0$, of $E[u] = \gamma = 1/\theta$. [Note that here, $\bar{u} = -\min_i(e_i)$.] The weighting matrix was then $\mathbf{W} = \mathbf{1}/N^2$ times the $2 \times 2$ moment matrix for $m_{i1} = (\bar{u}_i - \gamma_0)$ and $m_{i2} = [(\hat{u}_i - \bar{u})^2 - \gamma_0^2]$. We then

**Table 2.6**
Method of Moments Estimators for Efficiency Distribution for Deterministic Frontier Model Based on OLS Residuals

| Estimator | Exponential | Gamma | Half-Normal |
|---|---|---|---|
| | Population Moments | | |
| $E[u_i]$ | $1/\theta$ | $P/\theta$ | $(2/\pi)^{1/2}\sigma_u$ |
| $\text{Var}[u_i]$ | $1/\theta^2$ | $P/\theta^2$ | $[(\pi - 2)/\pi]\sigma_u^2$ |
| $E[\exp(-u_i)]$ | $[\theta/(1+\theta)]$ | $[\theta/(1+\theta)]^P$ | $2\Phi(-\sigma_u)\exp(\sigma_u^2/2)$ |
| | Implied Estimates[a] | | |
| $\sigma_u$ | 0.3930 | 0.1415 | 0.2348 |
| $\theta$ | 2.544 | 2.258 | NA |
| $P$ | 1.000 | 1.021 | NA |
| $E[\exp(-u_i)]$ | 0.7179 | 0.6424 | 0.8371 |
| Sample mean efficiency[b] | 0.6425 | 0.6425 | 0.6425 |

[a] $\sigma_u = 1/\theta$ for exponential, $\sigma_u = P^{1/2}/\theta$ for gamma.
[b] Sample average of $\exp(-u_i)$.

minimized with respect to $\gamma$ the quadratic form

$$q = [(\bar{u} - \gamma), (s_u^2 - \gamma^2)]^{\mathrm{T}} \mathbf{W}^{-1} [(\bar{u} - \gamma), (s_u^2 - \gamma^2)].^{79}$$

We use the sample variance to form the estimator for the half-normal frontier. The estimated mean technical efficiency, $E[\exp(-u_i)]$, can be computed using the sample moment or by using the estimated functions of the underlying parameters for the specific model. Overall, the three models produce estimates of mean cost efficiency ranging from 64% to 84%. This is considerably lower than other estimates produced from these data for these models (e.g., see the discussion of the Bayesian estimators below).

Under the assumption of a *stochastic* frontier model, each raw residual $e_i$ is an estimate of

$$y_i - (\alpha - E[u_i]) - \boldsymbol{\beta}^T \mathbf{x}_i = v_i - (u_i - E[u_i]).$$

Moments of the disturbances now involve $\sigma_v^2$, as well as the parameters of the distribution of $u$. The moment equations for the parameters of the distribution of $u_i$ in the models are as follows:

Exponential: $\theta = (-2/m_3)^{1/3}$, $P = 1.0$, $\sigma_v = (m_2 - 1/\theta^2)^{1/2}$,
$\alpha = a + 1/\theta$
Gamma: $\theta = -3m_3/(m_4 - 3m_2^2)$, $P = (-1/2)\theta^3 m_3$,
$\sigma_v = (m_2 - P/\theta^2)^{1/2}$, $\alpha = a + P/\theta$
Half-normal: $\sigma_u = \{m_3/[(2/\pi)^{1/2}(1 - [4/\pi])]\}^{1/3}$,
$\sigma_v = (m_2 - [(\pi - 2)/\pi]\sigma_u^2)^{1/2}$, $\alpha = a + \sigma_u(2/\pi)^{1/2}$

Counterparts to the results just given for the normal–gamma, normal–exponential, and normal–half-normal models based on the first four central moments (the first is zero) are given in table 2.7. I emphasize that all of these estimators are consistent. They are demonstrably less efficient than the MLEs, but the extent to which this affects the end results remains to be shown. The estimates do seem somewhat erratic—particularly compared to the MLEs given further below. However, the estimates clearly show that allowing for

**Table 2.7**
Method of Moments Estimates for Stochastic Frontier Models Based on OLS Residuals

| | Exponential | Gamma | Half-Normal |
|---|---|---|---|
| $\theta$ | 23.62 | 1.467 | NA |
| $P$ | 1.000 | 0.0002399 | NA |
| $\sigma_u$ | 0.0424 | 0.0106 | 0.08864 |
| $\sigma_v$ | 0.1344 | 0.1406 | 0.1304 |
| $\alpha$ | −7.256 | −7.294 | −7.223 |
| $E[\exp(-u)]$ | 0.9594 | 0.9999 | 0.9330 |

the firm-specific stochastic effect $v_i$ considerably reduces the estimated coefficients of inefficiency—the average efficiency rises from about 70% to more than 90%. Clearly, what is estimated to be a considerable amount of random noise in the stochastic frontier model is identified as inefficiency in the deterministic frontier model.

#### 2.10.2.3 MLE of the stochastic cost frontier model

Table 2.5 contains the MLEs of the half-normal, exponential, and gamma stochastic cost frontier models. Though the structural parameters are still fairly variable, the estimated distributions of $u_i$ implied by these estimates are much more stable than the method-of-moments estimators based on OLS. Estimates of the firm-specific inefficiencies, $E[u_i|\varepsilon_i]$, were computed using the JLMS method. Descriptive statistics appear in table 2.8. Figure 2.12 displays the estimated distribution for the efficiency terms from the gamma model.

**Table 2.8**
Descriptive Statistics for JLMS Estimates of $E[u_i|\varepsilon_i]$ Based on MLEs of Stochastic Frontier Models

| Model | Mean | Standard Dev. | Minimum | Maximum |
|---|---|---|---|---|
| Normal | 0.11867 | 0.060984 | 0.029822 | 0.37860 |
| Exponential | 0.097438 | 0.076407 | 0.022822 | 0.51387 |
| Gamma | 0.081423 | 0.077979 | 0.016044 | 0.52984 |

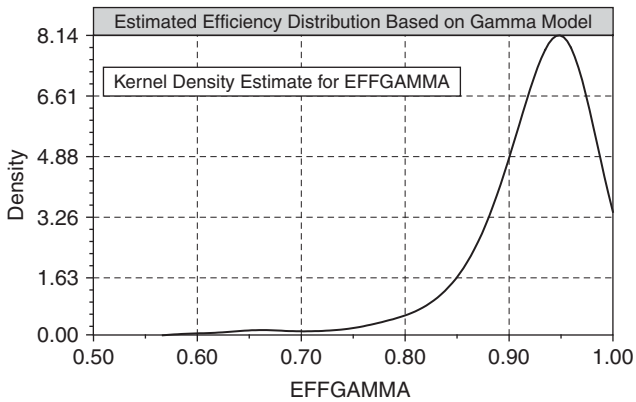


**Figure 2.12.** Kernel Density Estimate for Estimated Mean Efficiencies Based on Normal–Gamma Stochastic Frontier Model

The three estimates are very similar: The correlations are 0.968 for (normal, exponential), 0.944 for (normal, gamma), and 0.994 for (exponential, gamma). Figure 2.13 shows the three sets of estimates. The observations are sorted by output, so the figure also suggests that the large estimates for $u_i$ mostly correspond to very small outputs, which is to be expected. Finally, figure 2.14 shows the upper and lower confidence bounds for the total efficiency estimates using the Horrace and Schmidt (1996) and Bera and Sharma (1999) procedures described in section 2.8 for the normal–half-normal results.

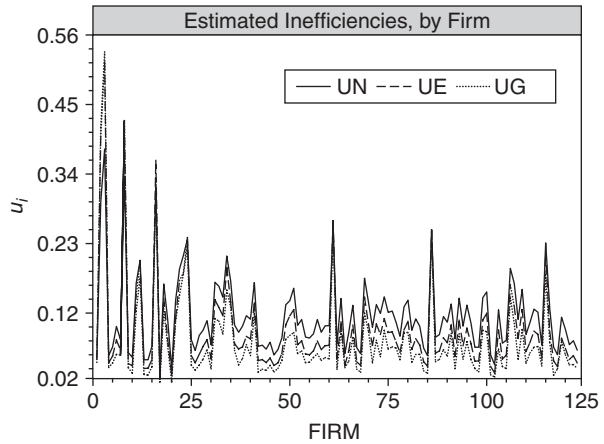


**Figure 2.13.** Estimates of $E[u_i|E_i]$

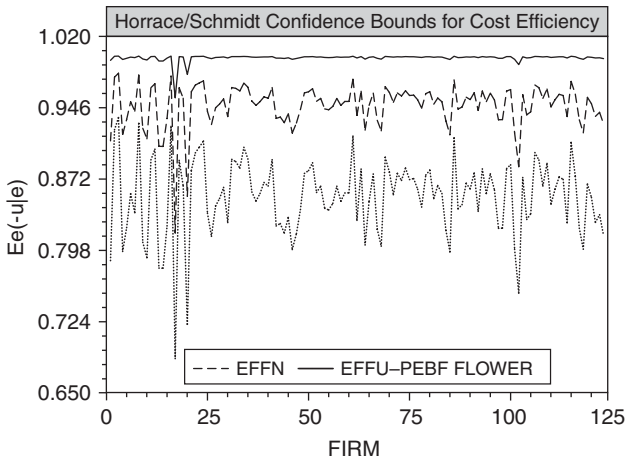


**Figure 2.14.** Confidence Limits for $E[u_i|E_i]$

#### 2.10.2.4 Bayesian and classical estimates of the normal–gamma frontier model

The Aigner-Lovell-Schmidt normal–half-normal model has provided the centerpiece of the stochastic frontier model specification in the large majority of studies. However, the literature contains a long succession of attempts to generalize the model, as discussed above. One of the enduring strands of research, beginning with Greene (1980a) and most recently augmented by Huang (2004), has been the development of the normal–gamma model. The original proposal in Greene (1980) suggested a deterministic frontier model with gamma distributed inefficiencies: $\varepsilon_i = u_i$ and $f(u_i) = [\lambda^P/\Gamma(P)]\exp(-\lambda u_i)u_i^{P-1}$. The deterministic frontier approach in general, and this specification in particular, has since taken a back seat in the evolution of the model. Beckers and Hammond (1987) and Greene (1990) proposed a stochastic frontier form of the gamma model. The normal–gamma model was not successfully implemented in either study. Its title notwithstanding, the complexity of the former seems to have prevented implementation. The latter presented a potentially simpler approach, but numerical difficulties examined closely by van den Broeck at al. (1994) and Ritter and Simar (1997) suggested that the classical MLEs suggested for the model in Greene (1990) were not accurately computed. Bayesian estimators were suggested in van den Broeck et al. (1994, 1995), which demonstrated the feasibility of the Bayesian method. Greene (2003b) has proposed a simulation-based MLE that appears to surmount the numerical problems. Concurrently, Huang's (2004) extension of Tsionas's (2002) and van den Broeck et al.'s (1994) Bayesian estimator brought considerable progress in allowing full variation in the crucial shape parameter in the gamma model.[80]

There have been numerous Bayesian applications of the stochastic frontier model since about 1995 (see, e.g., Koop et al., 1994; Kim and Schmidt, 2000). Owing to the mathematical convenience of the exponential and gamma densities, most of these have relied on the normal–exponential and normal–gamma specification. The theoretical development of the Bayesian approach has often applied the normal–gamma model, and in this strand of the literature, the Christensen and Greene (1976) data used here have provided a convenient common ground. First in the line are van den Broeck et al. (1994) and Koop et al. (1995), who fit the same quadratic model used above. The primary innovation in their study was to propose a data augmentation algorithm that produced posterior estimates of the inefficiency terms, $u_i$, along with the technology parameters.[81] As has been observed elsewhere (see Koop et al., 1997), estimation of the counterpart of a fixed-effects model in the Bayesian paradigm requires an informative prior. In their case, they equivalently assumed that $u_i$ were drawn from a gamma prior with prior median efficiency $r^* = 0.875$. As shown below, their findings were actually quite similar to those presented here.[82]

Following on van den Broeck et al.'s (1994) model, which assumes the Erlang (integer $P$) form of the normal–gamma model, Tsionas (2002) shows how the assumptions can be relaxed and the algorithm updated. Unlike Koop et al. (1995), who used importance sampling, Tsionas used a Gibbs sampler to draw observations from the posterior. In addition, his proposed method produces a hierarchical Bayesian estimator (ostensibly suggested for panel data) that yields firm-specific estimates for the technology parameters in the cost function as well as the firm-specific inefficiencies, $u_i$. It is useful to lay out Tsionas's specification in some detail. The cost model is

$$y_{it} = \alpha + \boldsymbol{\beta}_i^{\mathrm{T}} \mathbf{x}_{it} + v_{it} + u_{it},$$

where, initially, $f(u_{it}) = \theta \exp(-\theta u_{it})$. Later, this is extended to the two-parameter gamma model given above. Independent priors for the model are specified: $\alpha$, $\boldsymbol{\beta}_i \sim N[(a, \mathbf{b}), \boldsymbol{\Omega}]$, $i = 1, \ldots, N$; $(a, \mathbf{b}) \sim N[(0, \mathbf{0}), \mathbf{W}]$; $\boldsymbol{\Omega} \sim$ inverted Wishart; $\theta \sim$ two-parameter gamma; $\sigma_v \sim$ inverted gamma. Under his specification, $u_{it}$ values are draws from an exponential population with parameter $\theta$, where the prior mean for $\theta$ is, in turn, $q = -\ln r^*$. [Lengthy details on this specification are given in Tsionas's (2002) paper.] The Gibbs sampler for Tsionas's method for exponentially distributed $u_{it}$ ($P = 1$) is as follows:

1. Draw $\boldsymbol{\beta}_i$ from a conditional normal distribution.
2. Draw $\sigma$ from a conditional gamma distribution.
3. Draw $(a, \mathbf{b})$ from a conditional normal distribution.
4. Draw $\boldsymbol{\Omega}$ from a conditional inverted Wishart distribution.
5. Draw $u_{it}$ from a conditional truncated normal distribution.
6. Draw $\theta$ from a conditional gamma distribution.

The samples from the posterior distributions are obtained by cycling through these steps. The slightly more general cases of $P = 2$ and $P = 3$ are also considered. These cases greatly complicate step 5—direct sampling of random draws from the conditional distribution of $u_{it}$ becomes impossible when $P$ is not equal to one. Tsionas proposes a separate algorithm developed in a separate paper (Tsionas, 2000a), whereas Huang (2004) suggests a Metropolis Hastings strategy in his study.

Huang (2004) notes incorrectly that Tsionas allows noninteger values of $P$ in his implementation. Huang, in contrast, does and specifies continuous gamma priors for both $\theta$ and $P$ in his model. He goes on to discuss sampling from the posterior distribution of $u_{it}$ under the fully general specification. Thus, arguably, Huang brings full generality to the normal–gamma model. In comparison to Greene's approach, the extension would be the hierarchical Bayes extension of the model to allow separate coefficient vectors *even in a cross section*. Whether this is actually a feasible extension remains for ongoing research to establish. It is worth noting that the Tsionas and Huang methods have established a method of obtaining posterior means and variances (indeed,

entire distributions) for 761 parameters, $[(\alpha_i, \boldsymbol{\beta}_i, u_i), i = 1, \ldots, 123]a, \mathbf{b}, \boldsymbol{\Omega}, P, \theta, \sigma_u)$ based on a sample that contained 615 values in total on cost, log output and its square, and the logs of the two price ratios.

Table 2.9 displays the sets of estimates of the gamma frontier models obtained by classical MLE methods and the preferred set of Bayesian estimates (posterior means) from each of the three studies. The structural parameter estimates are somewhat similar.[83] The striking aspect of the results is in the estimated inefficiencies. van den Broeck et al.'s estimates are quite close to those here. Tsionas reports an implausible set of estimates that imply that every firm is at least 99.9% efficient. Huang's results for the heterogeneous translog model (firm-specific parameters) are essentially the same as Tsionas's, but those for his homogeneous parameters model are almost identical to those presented here. Indeed, figure 2 in Huang (2004, p. 286) is indistinguishable from figure 2.12, even including the slight second mode around abscissa of 0.67. Moreover, figure 1 in van den Broeck et al. (1994, p. 290) is likewise strikingly similar to figure 2.12 and Huang's figure 2.

With Tsionas (2002), Huang (2004), and Greene (2003b), it does seem that the technological problem of the normal–gamma model has largely been solved. The extension to "random" parameters yielded by the former two in cross-section data does seem overly optimistic. The random-parameters form has been extended to classical "mixed-model" estimation in, for example, Train (2003) and Greene (2003a, 2004b), with attendant "estimators" of the conditional means of individual specific parameter vectors. In both the classical and Bayesian frameworks, it seems at this juncture an interesting to pursue the question of what the implications are of extracting more posterior estimates of parameter distributions than there are numbers in the sample. In

**Table 2.9**
Estimates of the Normal–Gamma Stochastic Frontier Model (Coefficients or Posterior Means Only)

| | Greene | | van den Broeck | Tsionas | Huang | |
|---|---|---|---|---|---|---|
| | Exponential | Gamma[a] | Gamma | Gamma | Random | Fixed |
| Constant | −7.6336 | −7.652 | −7.479 | −7.416 | −7.217 | −7.4784 |
| Ln$y$ | 0.4398 | 0.4458 | 0.4276 | 0.445 | 0.3668 | 0.4447 |
| $\text{Ln}^2\text{y}$ | 0.02875 | 0.02839 | 0.0295 | 0.023 | 0.0335 | 0.0284 |
| $\text{Ln}P_L/P_F$ | 0.2701 | 0.2727 | 0.2492 | 0.247 | 0.2517 | 0.2346 |
| $\text{Ln}P_K/P_F$ | 0.03319 | 0.02933 | 0.0449 | 0.043 | 0.0695 | 0.0590 |
| $\theta$ | 10.263 | 8.425 | 11.273 | 75.12 | 77.4337 | 9.9025 |
| $P$ | 1.0000 | 0.6702 | 1.0000 | 1.000 | 0.9063 | 0.9575 |
| $\sum_v$ | 0.1044 | 0.1060 | 0.1136 | 0.0781 | 0.0374 | 0.1114 |
| Ln $L$ | 67.961 | 68.154 | NA | NA | NA | NA |
| Mean effect | 0.9072 | 0.9276 | 0.91 | 0.999 | 0.9891 | 0.9103 |

[a]Simulations for maximum simulated likelihood are computed using 200 Halton draws.

the end, however, as others have also observed, there appears to be notably little difference between the Bayesian posterior mean and classical MLEs of the stochastic frontier model.

#### 2.10.2.5 Duality between production and cost functions

Christensen and Greene (1976) studied, among other aspects of costs, the appropriate form of the production and cost function. Their specification search ultimately led to a translog cost function that was dual to a nonhomothetic production function. With linear homogeneity in the factor prices imposed, the full cost specification is

$$\ln\left(\frac{C}{PF}\right) = \alpha + \beta_K \ln\left(\frac{PK}{PF}\right) + \beta_L \ln\left(\frac{PL}{PF}\right) + \delta_y \ln y + \delta_{yy}\frac{1}{2}\ln^2 y + \gamma_{KK}\frac{1}{2}\ln^2\left(\frac{PK}{PF}\right) + \gamma_{LL}\frac{1}{2}\ln^2\left(\frac{PL}{PF}\right) + \gamma_{KL}\ln\left(\frac{PK}{PF}\right)\ln\left(\frac{PL}{PF}\right) + \theta_{yK}\ln y \ln\left(\frac{PK}{PF}\right) + \theta_{yL}\ln y \ln\left(\frac{PL}{PF}\right) + v + u.$$

Likelihood ratio tests firmly favored the full translog model over the restricted, homothetic technology that results if the final two terms are omitted. In translating this estimation exercise to the present stochastic frontier exercise, we find that in the nonhomothetic version, the estimate of $\lambda$ (and with it, any evidence of inefficiency) virtually disappears. With a homothetic function, the estimates of $\sigma_u$ and $\sigma_v$ are 0.16826 and 0.09831; thus, $u$ accounts for about 52% of the total variance of $[(\pi - 2)/\pi]\sigma_u^2 + \sigma_v^2$. With the full nonhomothetic form, these fall to 0.0000 and 0.13742, respectively. That is, the nonhomotheticity terms have picked up some of the idiosyncratic variation (in $v$) and all of the variation in $u$. Since the previous frontier analyses of these data have all used restricted versions of the cost function, this raises some interesting questions as to what our predecessors have actually found. It does seem reasonable to guess that estimates of inefficiency all equal to zero are themselves implausible as well, so some intriguing possibilities remain for future researchers to sort out.

With the preceding as a backdrop, we recompute the cost function using the homothetic form ($\theta_{yK} = \theta_{yL} = 0$) and recompute the JLMS estimators of cost inefficiency, $E[u|\varepsilon]$. We also compute a translog production frontier, without restricting it to homogeneity. The full unrestricted translog frontier would be

$$\ln y = \alpha_0 + \sum_{j=K,L,F} \beta_j \ln X_j + \frac{1}{2}\sum_{j=K,L,F}\sum_{m=K,L,F} \gamma_{jm} \ln X_j \ln X_m + v - u.$$

(Estimates of the frontier parameters are omitted in the interest of brevity.) We do impose the symmetry restrictions on the second-order terms. (I did

not check or impose the second-order monotonicity or quasiconcavity conditions.) As noted above, the relationship between the "inefficiency" in production and that from the cost function is confounded by at least two factors. In the absence of allocative inefficiency, if there are economies of scale, then $u_i$ in the cost function would equal $1/r$ times its counterpart on the production side where $r$ is the degree of homogeneity. The second source of ambiguity is the allocative inefficiency that will enter the cost inefficiency but not the technical (production) inefficiency. Acknowledging this possibility, we compute the JLMS estimates of $u_i$ from both functions. Using the production function parameters, we then compute the implied (local) scale elasticity,

$$\text{ES} = \sum_{j=K,L,F} \frac{\partial \ln y}{\partial \ln X_j} = \sum_{j=K,L,F} \left[ \beta_j + \frac{1}{2} \sum_{m=K,L,F} \gamma_{jm} \ln X_m \right].$$

We then divide each estimated inefficiency from the cost function by this elasticity computed from the production frontier. Figure 2.15 displays the scatter plot of these two sets of inefficiency estimates, which, with the expected variation, are clearly indicating the same "story" about inefficiency in these data. (The correlation between the two sets of estimates is 0.579.) Note that the mean inefficiency on the cost side of 0.124 (oddly close to the standard Bayesian prior value of 0.125) is noticeably larger than its counterpart on the production side of 0.080. It is tempting to attribute the difference to allocative inefficiency, which would not appear on the production side, as well as to a small distortion that results from the effect of economies of scale.

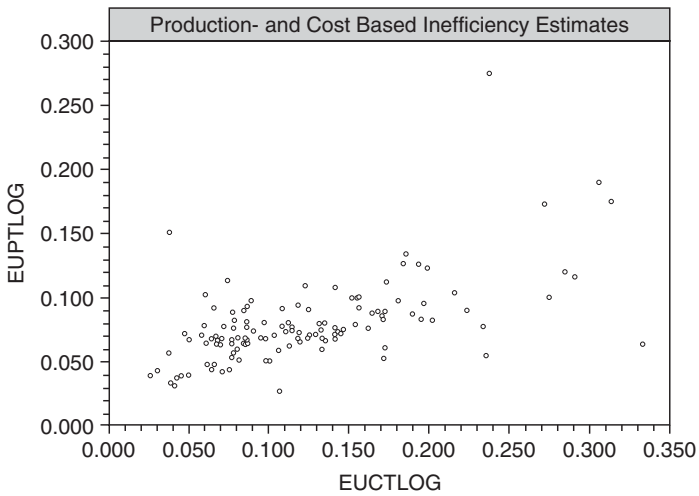


**Figure 2.15.** Cost and Production Inefficiencies

### 2.10.3 Time-invariant and time-varying inefficiency: airlines panel data

These data are from the pre-deregulation days of the U.S. domestic airline industry. The data are an extension of Caves et al. (1980) and Trethaway and Windle (1983). The original raw data set is a balanced panel of 25 firms observed over 15 years (1970–1984). After removing observations because of strikes, mergers, and missing values, the panel becomes an unbalanced one with a total of 256 observations on 25 firms. In a few cases, the time series contain gaps. Some of the models discussed above, notably Battese and Coelli (1992, 1995) and Cornwell et al. (1990), involve functions of time, $t$, which would have to be computed carefully to ensure the correct treatment of "time"; the gaps must be accounted for in the computations. Also, for firms that are not observed in the first year of the overall data set, when we consider functions of "time" with respect to a baseline, in keeping with the spirit of the stochastic frontier model, this baseline will be for the specific firm, not for the overall sample window. The unbalanced panel has 256 observations with $T_i = 4$, 7, 11, and 13 (one firm each), 12 (two firms) 9, 10, and 14 (three firms), 2 (four firms), and 15 (six firms). We will use these data to estimate frontier models with panel data and time-varying and time-invariant inefficiency.

Production and cost frontiers are fit for a five-input Cobb-Douglas production function: The inputs are labor, fuel, flight equipment, materials, and ground property. Labor is an index of 15 types of employees. Fuel is an index based on total consumption. The remaining variables are types of capital. It might be preferable to aggregate these into a single index, but for present purposes, little would be gained. Output aggregates four types of service: regular passenger service, charter service, mail, and other freight. Costs are also conditioned on two control variables: (log) average stage length, which may capture an economy of scale not reflected directly in the output variable, and load factor, which partly reflects the capital utilization rate. We also condition on the number of points served so as to attempt to capture network effects on costs. The data are described in table 2.10.

#### 2.10.3.1 Cobb-Douglas production frontiers

We first fit a Cobb-Douglas production function. This estimation illustrates a common problem that arises in fitting stochastic frontier models. The least squares residuals are positively skewed—the theory predicts they will be negatively skewed. We are thus unable to compute the usual first-round, method-of-moments estimators of $\lambda$ and $\sigma$ to begin the iterations. This finding does not prevent computation of the stochastic frontier model. However, it does necessitate some other strategy for starting the iterations. To force the issue, we simply reverse the sign of the third moment of the OLS residuals and proceed. Consistent with Waldman (1982), however, we then find that the log-likelihood function for the estimated model differs only trivially

**Table 2.10**
Airlines Data

| Variable | Mean | Standard Deviation | Description |
|---|---|---|---|
| FIRM | 11.8398438 | 7.09001883 | Firm, $i = 1, \ldots, 25$ |
| OUTPUT | 0.628784239 | 0.591862922 | Output, index |
| COST | 1172861.09 | 1197945.05 | Total cost |
| MTL | 0.751572192 | 0.642973957 | Material, quantity |
| FUEL | 0.583878603 | 0.503828645 | Fuel, quantity |
| EQPT | 0.651682905 | 0.567659248 | Equipment, quantity |
| LABOR | 0.595048662 | 0.508245612 | Labor, quantity |
| PROP | 0.656212972 | 0.692635345 | Property, quantity |
| PM | 491733.758 | 165628.591 | Materials price |
| PF | 427637.977 | 316179.137 | Fuel price |
| PE | 266391.048 | 110114.994 | Equipment price |
| PL | 669768.628 | 269367.140 | Labor price |
| PP | 40699.8592 | 19405.2501 | Property price |
| LOADFCTR | 0.526460328 | 0.120249828 | Load factor |
| STAGE | 492.642179 | 308.399978 | Average stage length |
| POINTS | 70.1328125 | 29.6541823 | Number of points served |

from the log-likelihood for a linear regression model with no one-sided error term. However, the estimates of $\sigma_u$, $\sigma_v$, $\lambda$, and $\sigma$ are quite reasonable, as are the remaining parameters and the estimated inefficiencies; indeed, the estimate of $\lambda$ is statistically significant, suggesting that there is, indeed, evidence of technical inefficiency in the data.[84] The conclusion to be drawn is that, for this data set, and more generally, when the OLS residuals are positively skewed (negatively for a cost frontier), then there is a second maximizer of the log-likelihood, OLS, that may be superior to the stochastic frontier. For our data, the two modes produce roughly equal log-likelihood values. For purposes of the analysis, the finding does suggest that one might want to take a critical look at the model specification and its consistency with the data before proceeding.

The least squares and MLEs of the parameters are given in table 2.11. The Pitt and Lee (1981) random-effects model is also fitted, which assumes that technical inefficiency is fixed through time and still half-normally distributed. The parameter estimates appear in table 2.11. Figure 2.16 shows the relationship between the two sets of estimates of $E[u_i|\varepsilon_i]$. Unfortunately, they are far from consistent. Note the widely different estimates of $\sigma_u$: 0.07 in the pooled model and 0.27 in the Pitt and Lee (1981) model. The time-invariant estimates vary widely across firms and are, in general, far larger. The time-varying values actually display relatively little within firm variation—there does not appear to be very much time variation in inefficiency suggested by these results. We might surmise that the time-invariant estimates are actually

**Table 2.11**
Estimated Cobb-Douglas Production Frontiers (Standard Errors in Parentheses)

| Variable | Least Squares | Pooled Frontier | Random Effects |
|---|---|---|---|
| Constant | −1.1124 (0.0102) | −1.0584 (0.0233) | −0.8801 (0.0302) |
| Ln fuel | 0.3828 (0.0712) | 0.3835 (0.0704) | 0.2110 (0.0951) |
| Ln materials | 0.7192 (0.0773) | 0.7167 (0.0765) | 0.8170 (0.0666) |
| Ln equipment | 0.2192 (0.0739) | 0.2196 (0.0730) | 0.3602 (0.120) |
| Ln labor | −0.4101 (0.0645) | −0.4114 (0.0638) | −0.3166 (0.0770) |
| Ln property | 0.1880 (0.0298) | 0.1897 (0.0296) | 0.1131 (0.0224) |
| $\lambda$ | 0.0 | 0.43515 | 2.2975 |
| $\sigma$ | 0.1624 | 0.16933 | 0.29003 |
| $\sigma_u$ | 0.0 | 0.06757 | 0.26593 |
| $\sigma_v$ | 0.1624 | 0.15527 | 0.11575 |
| Log-likelihood | 105.0588 | 105.0617 | 155.3240 |

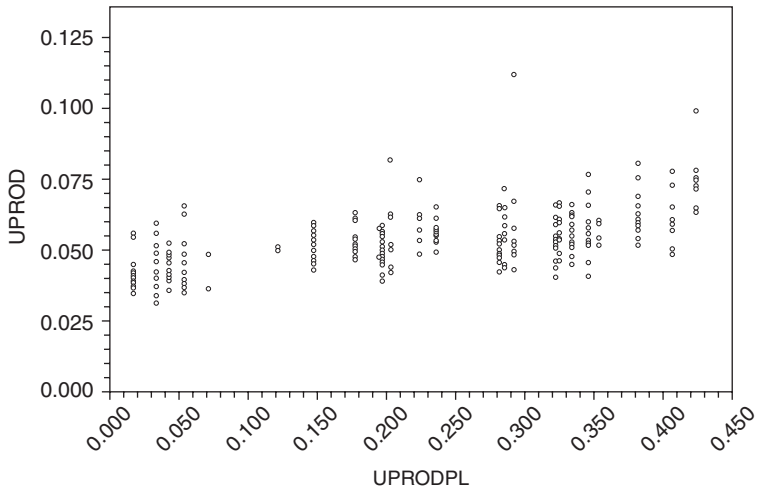


**Figure 2.16.** Pooled Time-Varying Versus Time-Invariant Inefficiencies

dominated by heterogeneity not related to inefficiency. In sum, these results are so inconsistent that, if anything, they suggest a serious specification problem with at least one of the two models. Let us turn to the cost specification to investigate.

#### 2.10.3.2 Stochastic cost frontiers

Estimates of the Cobb-Douglas stochastic frontier cost function are given in table 2.12, with the least squares results for comparison. Cost and the

**Table 2.12**
Estimated Stochastic Cost Frontier Models (Standard Errors in Parentheses)

| Variable | Least Squares | Half-Normal | Truncated Normal |
|---|---|---|---|
| Constant | −13.610 (0.0865) | −13.670 (0.0848) | −13.782 (0.145) |
| $\ln(P_M/P_P)$ | 1.953 (0.0754) | 1.9598 (0.0726) | 1.9556 (0.0666) |
| $\ln(P_F/P_P)$ | −0.6562 (0.0141) | −0.6601 (0.0139) | −0.6590 (0.01516) |
| $\ln(P_L/P_P)$ | −0.06088 (0.0533) | −0.07540 (0.0532) | −0.08667 (0.0577) |
| $\ln(P_E/P_P)$ | −0.1935 (0.0690) | −0.1840 (0.0663) | −0.1652 (0.0546) |
| Ln $y$ | 0.01054 (0.0133) | 0.01063 (0.0129) | 0.007384 (0.0145) |
| $1/2\ln^2$ y | 0.009166 (0.00435) | 0.008714 (0.00427) | 0.007919 (0.00444) |
| Constant | NA | NA | −0.1372 (0.777) |
| Load factor | −0.4712 (0.103) | −0.4265 (0.0992) | 0.5603 (0.318) |
| Ln stage length | 0.03828 (0.00889) | 0.03495 (0.00858) | −0.04397 (0.0437) |
| Points | 0.00007144 (0.000252) | 0.00001464 (0.000250) | −0.0002034 (0.000285) |
| $\lambda$ | 0.0 | 0.88157 | 1.05196 |
| $\sigma$ | 0.08915 | 0.10285 | 0.09214 |
| $\sigma_u$ | 0.0 | 0.06801 | 0.06678 |
| $\sigma_v$ | 0.08915 | 0.07715 | 0.06348 |
| Log-likelihood | 260.7117 | 261.1061 | 261.3801 |

remaining prices are normalized on the property price. Additional "shift factors" that appear in the cost equation are load factor, the log of stage length, and the number of points served. These three variables affect costs the way we might expect. Note at the outset that three of the price coefficients have the wrong sign, so the model is suspect from this point on. But let us continue for the sake of the example. We compute the JLMS estimates of $E[u_i|\varepsilon_i]$ from the MLEs of the estimated cost frontier. They are essentially uncorrelated ($r = 0.04$) with their counterparts from the production frontier. As noted above, this adds to the impression that there is something amiss with the specification of the model—we suspect the production model. The kernel density estimator for $\exp(-u_i)$ based on the JLMS estimates in figure 2.17 appears reasonable, and at least numerically consistent with the production model. However, like other descriptive statistics, it does mask the very large differences between the individual production and cost estimates. Table 2.12 also presents results for the normal–truncated-normal model in which

$$u_i = |U_i|, E[U_i] = \mu_0 + \mu_1(\text{load factor})_i + \mu_2 \ln(\text{stage length})_i + \mu_3 \text{points}_i$$

That is, these three exogenous influences are now assumed to shift the distribution of inefficiency rather than the cost function itself. Based on the estimates and statistical significance, this model change does not appear to improve it. Surprisingly, the estimated inefficiencies are almost the same.

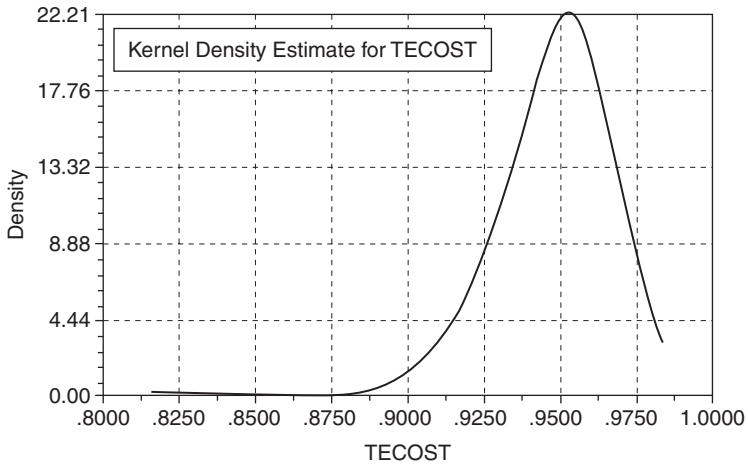


**Figure 2.17.** Kernel Estimator for $E[\exp(-u_i)]$

**Table 2.13**
Estimated Stochastic Cost Frontier Models (Standard Errors in Parentheses)

| | Time-Invariant Inefficiency | | Time-Varying Inefficiency | |
|---|---|---|---|---|
| Variable | Fixed Effect | Random Effect | Fixed Effect | Random Effect[a] |
| Constant | NA | −13.548 (0.373) | NA | −13.540 (0.0552) |
| $\ln(P_M/P_P)$ | 1.7741 (0.0869) | 2.0037 (0.0763) | 1.8970 (0.101) | 2.0092 (0.0457) |
| $\ln(P_F/P_P)$ | −0.5347 (0.0180) | −0.6440 (0.0260) | −0.7115 (0.020) | −0.6417 (0.00962) |
| $\ln(P_L/P_P)$ | −0.01503 (0.0525) | −0.07291 (0.0952) | −0.04252 (0.0625) | −0.07231 (0.0377) |
| $\ln(P_E/P_P)$ | −0.2225 (0.0753) | −0.2643 (0.0632) | −0.05125 (0.0898) | −0.2711 (0.0383) |
| $\ln y$ | −0.1990 (0.0473) | 0.01781 (0.0360) | 0.03840 (0.0404) | 0.01580 (0.00932) |
| $1/2\ln^2 y$ | −0.009713 (0.00824) | 0.0119 (0.00833) | 0.008306 (00872) | 0.01221 (0.00307) |
| Load factor | −0.4918 (0.183) | −0.4482 (0.172) | −0.4148 (0.180) | −0.4576 (0.0500) |
| Ln Stage length | −0.001397 (0.0114) | 0.03326 (0.0378) | 0.05870 (0.0133) | 0.032823 (0.00443) |
| Points | −0.0006279 (0.0005) | −0.000134 (0.000743) | 0.000631 (0.0006) | −0.000119 (0.0002) |
| $\lambda$ | 0.0 | 0.58809 | 0.5243 | 0.50148 |
| $\sigma$ | 0.07526 | 0.09668 | 0.10475 | 0.08900 |
| $\sigma_u$ | 0.0 | 0.04901 | 0.04865 | 0.03990 |
| $\sigma_v$ | 0.07526 | 0.08334 | 0.09278 | 0.07956 |
| Log-likelihood | 317.2061 | 263.2849 | 247.2508 | 262.4393 |

[a] Estimated standard deviation of w is 0.03306.

### 2.10.3.3 Panel-data models for costs

Table 2.13 presents estimates of the fixed-effects linear regression and Pitt and Lee random-effects models. The behavior of the latter was discussed above. Figure 2.18 shows the results for the Schmidt and Sickles (1984) calculations

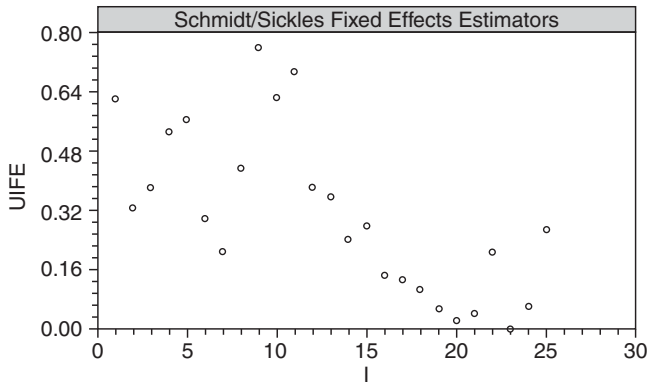


**Figure 2.18.** Estimated $E[u_i|E_i]$ from Fixed-Effects Model

based on the fixed effects. Note, again, that the estimates of $u_i$ are vastly larger for this estimator than for the pooled stochastic frontier cost or production model. We also fit a "true" fixed-effects model with these data, with some surprising results. The model is

$$\ln(C/P_P)_{it} = \sum_k \beta_k \ln(P_k/P_P) + \beta_y \ln y_{it} + \beta_{yy}\left(\frac{1}{2}\ln^2 y_{it}\right) + \gamma_1(\text{load factor})_{it} + \gamma_2 \ln(\text{stage})_{it} + \gamma_3 \text{points}_{it} + \sum_i \alpha_i d_{it} + v_{it} + u_{it},$$

that is, a stochastic cost frontier model with half-normal inefficiency and with the firm dummy variables. The log-likelihood function has two distinct modes. At one, the values of the parameters are quite reasonable, and the value of the log-likelihood is 247.2508, *compared to 261.1061 for the linear model without the firm dummy variables.* A second maximum of the log-likelihood occurs at the least squares dummy variable estimator—the estimated value of $\lambda$ is 0.00004226—where the log-likelihood value is 317.2061. We conclude that this model is saturated. While the model that assumes that there is no unobserved heterogeneity and that inefficiency is time invariant (the Pitt and Lee model) creates extreme and apparently distorted values for the inefficiency, this model that assumes that all time-invariant effects are heterogeneity and that inefficiency varies haphazardly over time appears to be overspecified. Finally, to continue this line of inquiry, we fit the "true random-effects model,"

$$\ln(C/P_P)_{it} = (\alpha + w_i) + \sum_k \beta_k \ln(P_k/P_P) + \beta_y \ln y_{it} + \beta_{yy}\left(\frac{1}{2}\ln^2 y_{it}\right) + \gamma_1(\text{load factor})_{it} + \gamma_2 \ln(\text{stage})_{it} + \gamma_3 \text{points}_{it} + v_{it} + u_{it},$$

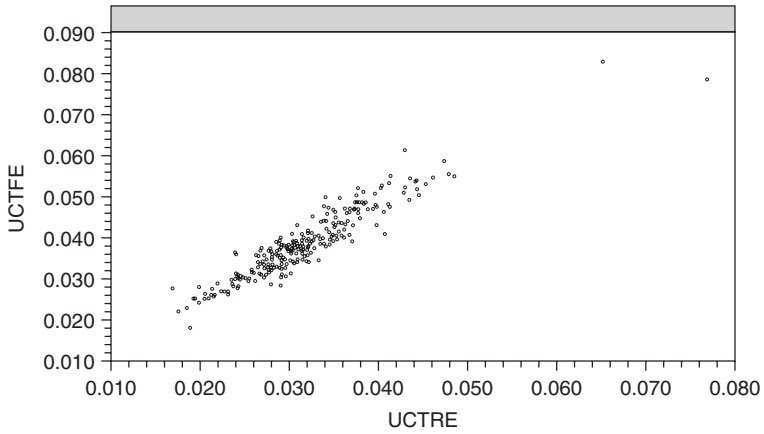


**Figure 2.19.** True Random-Effects and True Fixed-Effects Estimators

where $w_i$ picks up time-invariant heterogeneity assumed to be uncorrelated with everything else in the model, and $v_{it} + u_{it}$ are the familiar stochastic frontier specification. This model is fit by MSL, using 100 Halton draws for the simulations. Note that this model is an extension of the pooled stochastic frontier model, not the Pitt and Lee model. Figure 2.19 plots the estimated inefficiencies from the two true effects models. The striking agreement is consistent with results found in other studies. In general (see Kim and Schmidt, 2000, for commentary), the differences from one specification to another do not usually hang so much on whether one uses a fixed- or random-effects approach as they do on other aspects of the specification. On the other hand, note also the above findings that distributional assumptions do not appear to be a crucial determinant, either. Nor, it turns out, does the difference between Bayesian and classical treatments often amount to very much. One conclusion that does appear to stand out from the results here, and in Greene (2004a, 2004b, 2005), is that the assumption of time invariance in inefficiency does bring very large effects compared to a model in which inefficiency varies through time.

A final note, the log-likelihood for the true random-effects model is 262.4393, compared to 261.1061 for the pooled model. Chi squared is only 2.666, so we would not reject the hypothesis of the pooled model. The evidence for a panel-data treatment with these data is something less than compelling. As a final indication, we use the Breusch and Pagan (1980) Lagrange multiplier statistic from the simple linear model. The value is only 1.48. As a chi squared with one degree of freedom, this reinforces the above conclusion: For these data, a pooled model is preferable to any panel-data treatment.

### 2.10.4 Random- and fixed-effects models: data on U.S. banks

Data for this study are taken from the Commercial Bank Holding Company Database maintained by the Chicago Federal Reserve Bank.[85] Data are derived

from the Report of Condition and Income (Call Report) for all U.S. commercial banks that report to the Federal Reserve Banks and the Federal Deposit Insurance Corporation. A random sample of 500 banks from a total of more than 5,000 was used. This is a balanced panel of five observations (1996–2000) on each of 500 banks. Observations consist of total costs, $C_{it}$, five outputs, $y_{mit}$, and the unit prices, denoted $w_{kit}$, of five inputs, $x_{kit}$. The measured variables used to construct the data set used in the analysis are described in table 2.14 (descriptive statistics are omitted for lack of meaning and interest).

The transformed variables contained in the maintained data set and used in the study to follow (the names in the data file) are given in table 2.15. The banking data are used typically to fit multiple-output translog models (see, e.g., Kumbhakar and Tsionas, 2004, 2005a, 2005b). In the interest of maintaining

**Table 2.14**
Data Used in Cost Frontier Analysis of Banking

| Variable | Description |
|---|---|
| $C_{it}$ | Total cost of transformation of financial and physical resources into loans and investments = the sum of the five cost items described below |
| $y_{1it}$ | Installment loans to individuals for personal and household expenses |
| $y_{2it}$ | Real estate loans |
| $y_{3it}$ | Business loans |
| $y_{4it}$ | Federal funds sold and securities purchased under agreements to resell |
| $y_{5it}$ | Other assets |
| $w_{1it}$ | Price of labor, average wage per employee |
| $w_{2it}$ | Price of capital = expenses on premises and fixed assets divided by the dollar value of premises and fixed assets |
| $w_{3it}$ | Price of purchased funds = interest expense on money market deposits plus expense of federal funds purchased and securities sold under agreements to repurchase plus interest expense on demand notes issued by the U.S. Treasury divided by the dollar value of purchased funds |
| $w_{4it}$ | Price of interest-bearing deposits in total transaction accounts = interest expense on interest-bearing categories of total transaction accounts |
| $w_{5it}$ | Price of interest-bearing deposits in total nontransaction accounts = interest expense on total deposits minus interest expense on money market deposit accounts divided by the dollar value of interest-bearing deposits in total nontransaction accounts |
| $t$ | Trend variable; $t = 1, 2, 3, 4, 5$ for years 1996, 1997, 1998, 1999, 2000 |

**Table 2.15**
Variables Used in Cost Frontier Analysis of Banking

| Variable | Description |
|---|---|
| $C$ | Log(Cost/$w_5$) |
| $y_1, \ldots, y_5$ | Logs of outputs |
| $y$ | Log of sum of all five outputs |
| $w_1, \ldots, w_4$ | Log $(w_1/w_5), \ldots, \log(w_4/w_5)$ |
| $y_{11}, y_{12}, \ldots$ | 1/2 Squares and cross-products of log output variables |
| $w_{11}, w_{12}, \ldots$ | 1/2 Squares and cross-products of log price ratio variables |
| $w_1y_1, \ldots, w_4y_5$ | Cross products of log price ratios times log outputs |
| $t_2$ | $1/2t^2$ |
| $tw_1, \ldots, tw_4$ | Cross products of $t$ with log price ratios |
| $ty_1, \ldots, ty_5$ | Cross products of $t$ with log outputs |

a simple example for our application, I have combined the outputs into a single scale variable, which is the simple sum of the five outputs. Whether the aggregation would be appropriate given the technology is debatable—it would be testable—but in the interest of keeping the presentation simple, we will maintain the hypothesis. In addition, though these data are detailed and "clean enough" to enable estimation of translog models, we will, again in the interest of simplicity, restrict attention to variants of the (more or less standard) Cobb-Douglas cost function with the additional squared term in log output.

In this application, we will examine some additional panel-data estimators, including fixed-effects, random-effects, random-parameters, and the Battese and Coelli (1992, 1995) model of systematic time variation in inefficiency.

#### 2.10.4.1 Estimating inefficiency and unobserved heterogeneity

The observations in this data set are relatively homogeneous. They do differ substantially with respect to scale. However, the technology of banking is well known and smoothly dispersed, and there is little reason to expect latent heterogeneity to be a major factor. In this application, we will examine the impact of the different specifications for unobserved heterogeneity on estimates of cost inefficiency. Table 2.16 presents estimated parameters for simplest forms of the five common specifications:

$$\ln\left(\frac{C}{w_5}\right) = \alpha + \gamma_y \ln y + \gamma_{yy}\left(\frac{1}{2}\ln^2 y\right) + \sum_{k=1}^{4}\beta_k \ln\left(\frac{w_k}{w_5}\right) + v + u$$

**Table 2.16**
Estimated Stochastic Cost Frontier Models

| | Pooled | | Time-Varying Effects | | Time-Invariant Effects | |
|---|---|---|---|---|---|---|
| Variable | Half-Normal | Truncated[a] | Random[b] | Fixed | Random | Fixed |
| Constant | −0.066983 | −0.16838 | −0.065942 | Varies | 0.51228 | Varies |
| Ln$y$ | 0.66914 | 0.69865 | 0.66959 | 0.65829 | 0.58515 | 0.58556 |
| $1/2\ln^2 y$ | 0.023879 | 0.021374 | 0.023835 | 0.024922 | 0.030907 | 0.030743 |
| Ln$w_1/w_5$ | 0.38815 | 0.38733 | 0.38764 | 0.39766 | 0.39721 | 0.38387 |
| Ln$w_2/w_5$ | 0.020565 | 0.02010 | 0.020758 | 0.016966 | 0.032037 | 0.036016 |
| Ln$w_3/w_5$ | 0.17959 | 0.17730 | 0.17995 | 0.17259 | 0.17780 | 0.18758 |
| Ln$w_4/w_5$ | 0.13479 | 0.13442 | 0.13483 | 0.133419 | 0.13784 | 0.13823 |
| $\lambda$ | 1.81064 | 18.33032 | 1.82158 | 1.88219 | 0.30418 | 0.0 |
| $\sigma$ | 0.31866 | 3.07476 | 0.31796 | 0.40601 | 0.23572 | 0.22750 |
| $\sigma_u$ | 0.27894 | 3.07019 | 0.27872 | 0.35854 | 0.06860 | 0.0 |
| $\sigma_v$ | 0.15406 | 0.16749 | 0.15301 | 0.19049 | 0.22552 | 0.22750 |
| Log-likelihood | 183.9359 | 207.0714 | 184.0844 | 234.4165 | 136.6902 | 436.8185 |

[a] MLE of $\mu$ is 60.03185.
[b] MLE of $\sigma_w$ is 0.01891958.

$$\text{Pooled: } \ln(C/w_5)_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} + u_{it}$$

- This model is estimated by ML as discussed above. The JLMS estimator is used to estimate $u_{it}$.

$$\text{Random Effects: } \ln(C/w_5)_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} + u_i$$

- This is the Pitt and Lee (1981) model, also fit by ML. The form of the log-likelihood appears in Pitt and Lee (1981) and Greene (2000). The JLMS estimator is used by replacing $\varepsilon_{it}$ with $\bar{\varepsilon}_i$ and $\sigma^2$ with $\sigma^2/T$ (see Kim and Schmidt, 2000).

$$\text{Fixed Effects: } \ln(C/w_5)_{it} = \alpha_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} + (\alpha_i - \alpha_0)$$

- This is the Schmidt and Sickles (1984) approach, fit be ordinary (within-groups) OLS, followed by translation of the constants: $u_i = a_i - \min(a_i)$.

$$\text{True Random Effects: } \ln(C/w_5)_{it} = (\alpha + w_i) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} + u_{it}$$

- The model is developed in Greene (2005). The parameters are estimated by MSL. The JLMS estimator is employed by integrating $w_i$ out of $E[u_{it}|\varepsilon_{it}(w_i)]$. That is, $\varepsilon_{it}$ is a function of $w_i$, and then $w_i$ is integrated out of $u_{it}$.

$$\text{True Fixed Effects: } \ln(C/w_5)_{it} = \alpha_i + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} + u_{it}$$

- The model is estimated by brute force ML using the methods described in Greene (2004a). The JLMS estimator is used directly for $u_{it}$.

Parameter estimates are given table 2.16 to enable comparison of the models. Standard errors are omitted in the interest of brevity.

The estimated parameters are consistent across frameworks but differ surprisingly with respect to the assumption of whether the inefficiency is assumed to be time invariant or not. This finding is consistent with what we have observed elsewhere. In terms of its impact on model estimates, the assumption of time invariance seems to have much greater influence than the difference between fixed and random effects. Note, within the assumption of time-varying inefficiency, that neither the fixed- nor the random-effects model is preferred to the pooled model based on the likelihood ratio statistic. (The likelihood function rises substantially for the fixed-effects model, but with 499 degrees of freedom, the value of 100.96 is far from significant.) The truncation model displays the characteristic erratic behavior. The technology parameters are quite stable, but the truncation model substantially alters the estimated distribution of $u_{it}$. Superficially, the truncation model appears more reasonable. Figure 2.20 compares the estimated distributions—the upper figure is for $E[u_{it}|\varepsilon_{it}]$ for the half-normal model.

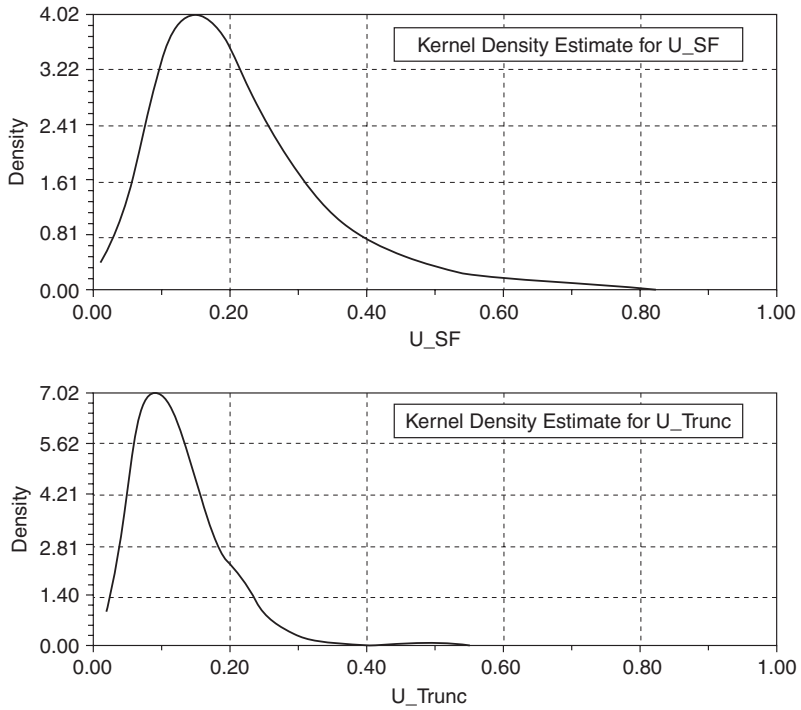


**Figure 2.20.** Estimated Mean Inefficiencies for Half-Normal (top) and Truncated-Normal (bottom) Models

**Table 2.17**
Descriptive Statistics for Estimated Inefficiencies

| Model | Mean | SD | Skewness | Minimum | Maximum |
|---|---|---|---|---|---|
| Pooled | 0.220143 | 0.127907 | 1.59129 | 0.0371616 | 0.795649 |
| True fixed effects | 0.255033 | 0.118152 | 1.61515 | 0.0658233 | 1.02899 |
| True random effects | 0.220369 | 0.130749 | 1.84823 | 0.0372414 | 1.18654 |
| Random effects | 0.0546 | 0.0168001 | 2.07666 | 0.0266957 | 0.165469 |
| Fixed effects | 0.291174 | 0.106474 | 0.472136 | 0 | 0.764483 |
| Truncated normal | 0.128167 | 0.0684533 | 1.96499 | 0.0341525 | 0.54011 |
| Latent class | 0.110435 | 0.082082 | 2.13809 | 0.0157056 | 0.703589 |
| Random parameters | 0.199054 | 0.1217 | 1.89409 | 0.0340895 | 1.08773 |

**Table 2.18**
Correlations among Inefficiency Estimates

| | Pooled | Truncated | True RE | True FE | Random | Fixed |
|---|---|---|---|---|---|---|
| Pooled | 1.00000 | | | | | |
| Truncated | 0.44376 | 1.00000 | | | | |
| True FE | 0.99567 | 0.44473 | 1.00000 | | | |
| True RE | 0.90975 | 0.10552 | 0.91713 | 1.00000 | | |
| Random | 0.44354 | 0.99716 | 0.44570 | 0.10565 | 1.00000 | |
| Fixed | 0.44675 | 0.95960 | 0.44159 | 0.08743 | 0.96629 | 1.00000 |

FE, fixed effects; RE, random effects.

Descriptive statistics for the estimates of $E[u_{it}|\varepsilon_{it}]$ (or $E[u_i|\varepsilon_{it}, t = 1, \ldots, T]$ in the case of the time-invariant models) are given in tables 2.17 and 2.18. For the time-invariant cases, consistent with the model, the fixed value of $u_i$ is repeated for the five observations for bank $i$. Among the notable features of the results are the high correlation between random- and fixed-effects estimates, but the far lower correlations across the two modeling platforms, time-varying and time-invariant effects. This is once again consistent with results observed elsewhere. Finally, scatter plots of the sets of estimates are consistent with what is suggested in tables 2.17 and 2.18. When estimates from one model that assumes $u_{it}$ varies across time are plotted against another, the estimates are essentially the same. However, as observed in Greene (2004a, 2004b), when, for example, the estimates of $u_{it}$ (or the group means) from either true effects model are plotted against (repeated) $u_i$ from the model with time-invariant inefficiency, the plot confirms that the estimates are almost uncorrelated.

#### 2.10.4.2 Parameter heterogeneity

Finally, we consider the two classical methods of allowing for parameter heterogeneity in the model, the random-parameters model and the latent class

model, which allows for discrete parameter variation. Both depart from the normal–half-normal stochastic frontier model.

The random-parameters model is

$$y_{it} = \alpha_i + \boldsymbol{\beta}_i^T \mathbf{x}_{it} + v_{it} + u_{it}$$
$$(\alpha_i, \boldsymbol{\beta}_i^{\mathrm{T}})^{\mathrm{T}} \sim N[(\alpha_0, \boldsymbol{\beta}_0^{\mathrm{T}})^{\mathrm{T}}, \boldsymbol{\Sigma}].$$

The technology parameters are allowed to vary randomly (normally) across firms. The marginal distributions of the random components, $u_{it}$ and $v_{it}$, are assumed to be common. The model is estimated by MSL as described in Greene (2004a).[86] It is useful to digress briefly to document the computation of the estimators of $E[u_{it}|\varepsilon_{it}]$. For convenience, let $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ denote the full vector of parameters, so $\boldsymbol{\theta}_i$ is what appears in the model. We can write the random-parameters component of the model as

$$\boldsymbol{\theta}_i = \boldsymbol{\theta} + \mathbf{w_i},$$

where $\mathbf{w_i} \sim N[\mathbf{0}, \boldsymbol{\Sigma}]$.

During estimation, we go a step further, and write $\mathbf{w_i} = \boldsymbol{\Gamma}\mathbf{h_i}$, where $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} = \boldsymbol{\Sigma}$ and $\mathbf{h_i} \sim N[\mathbf{0}, \mathbf{I}]$. Then, the JLMS estimator conditioned on $\mathbf{w_i}$ is

$$\hat{E}[u_{it}|\varepsilon_{it}(\mathbf{w}_i)] = \frac{\lambda\sigma}{1+\lambda^2}\left[\frac{-\varepsilon_{it}(\mathbf{w}_i)\lambda}{\sigma} + \frac{\phi[-\varepsilon_{it}(\mathbf{w}_i)\lambda/\sigma]}{\Phi[-\varepsilon_{it}(\mathbf{w}_i)\lambda/\sigma]}\right],$$

where $\varepsilon_{it}(\mathbf{w_{ii}}) = \mathbf{y_{it}} - (\boldsymbol{\theta} + \mathbf{w_i})^{\mathrm{T}}(\mathbf{1}, \mathbf{x_{it}})$.

We now must integrate $\mathbf{w_i}$ out of the expression; the unconditional estimator will be

$$\hat{\hat{E}}[u_{it}|\text{data}] = E_{\mathbf{w}_i}\hat{E}[u_{it}|\varepsilon_{it}(\mathbf{w}_i)]$$
$$= \int_{\mathbf{w}_i} \frac{\lambda\sigma}{1+\lambda^2}\left[\frac{-\varepsilon_{it}(\mathbf{w}_i)\lambda}{\sigma} + \frac{\phi[-\varepsilon_{it}(\mathbf{w}_i)\lambda/\sigma]}{\Phi[-\varepsilon_{it}(\mathbf{w}_i)\lambda/\sigma]}\right] f(\mathbf{w}_i)\, d\mathbf{w}_i.$$

(This is essentially the same as the Bayesian posterior mean estimator of the same quantity.) We are conditioning on all the data for this observation, including the dependent variable. Thus, what we have denoted $f(\mathbf{w_i})$ is actually $f(\mathbf{w_i}|\text{data}_\mathbf{i})$. The simulation-based estimator will condition out the dependent variable (for discussion, see Train, 2003, chapter 10; Greene, 2003a). The integral cannot be computed in closed form, so it is approximated by simulation. The estimator is

$$\hat{\hat{E}}^S[u_{it}|\text{data}] = \frac{1}{R}\sum_{r=1}^{R} \hat{f}_{ir} \frac{\lambda\sigma}{1+\lambda^2}\left[\frac{-\varepsilon_{it}(\mathbf{w}_{ir})\lambda}{\sigma} + \frac{\phi[-\varepsilon_{it}(\mathbf{w}_{ir})\lambda/\sigma]}{\Phi[-\varepsilon_{it}(\mathbf{w}_{ir})\lambda/\sigma]}\right],$$

where draws from the distribution of $\mathbf{w_i}$ are obtained as $\boldsymbol{\Gamma}\mathbf{h}_i$, where $\mathbf{h_i}$ is a vector of primitive draws from the standard normal distribution and recall

$\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} = \boldsymbol{\Sigma}$.[87] The weights in the summation are

$$\hat{f}_{ir} = \frac{L_{ir}}{\frac{1}{R}\sum_{r=1}^{R} L_{ir}},$$

where $L_{ir}$ is the joint likelihood (not the log) for the $T$ observations for individual (bank) $i$ computed at $\boldsymbol{\theta}_{ir}, \lambda, \sigma$. Note that the full set of computations is done ex post based on the MLEs of $\lambda, \sigma, (\alpha_0, \boldsymbol{\beta}_0)$, and $\boldsymbol{\Gamma}$. (In implementation, it is convenient to compute this quantity at the same time the simulated log-likelihood is computed, so it does not actually require very much additional computation—at the final iteration, these conditional estimates are present as a byproduct of the computation of the likelihood.)

Estimation of the latent class model is described in section 2.4.9.2. For estimates of $E[u_{it}|\varepsilon_{it}]$, we use the following strategy. [Again, I sketch only the overview. For further details, Greene (2003a, chapter 16) has additional material.] The latent class stochastic frontier model estimates consist of $(\alpha_j, \boldsymbol{\beta}_j, \lambda_j, \sigma_j, \pi_j)$, where $j$ indicates the $j$th of $J$ classes and $\pi_j$ is the unconditional (counterpart to "prior") probability of membership in the $j$th class. The conditional probabilities of class membership for bank $i$ are obtained via Bayes theorem; these equal

$$\pi(j|i) = \frac{\pi_j L(i|j)}{\sum_{j=1}^{J} \pi_j L(i|j)},$$

where $L(i|j)$ is the likelihood (not its log) for bank $i$, computed at the parameters specific to class $j$. (These are the counterparts to posterior probabilities in a Bayesian treatment.) Let $E[u_{it}|\varepsilon_{it}, j]$ denote the JLMS estimator of $E[u_{it}|\varepsilon_{it}]$ in specific class $j$—that is, computed using the parameters of class $j$. Then, our estimate of $E[u_{it}|\varepsilon_{it}]$ is

$$\hat{E}[u_{it}|\varepsilon_{it}] = \sum_{j=1}^{J} \pi(j|i)\hat{E}[u_{it}|\varepsilon_{it}, j].$$

The computations are done ex post, based on the MLEs of $(\alpha_j, \beta_j, \lambda_j, \sigma_j, \pi_j)$, $j = 1, \ldots, J$. (As in the random-parameters model, this computation is actually done at the same time the log-likelihood is computed, each time it is computed, so that at convergence, the estimated inefficiencies are already computed as a byproduct of the optimization process.)

Table 2.19 contains estimates of the parameters for the pooled stochastic frontier model, a full random-parameters model, and a three-class latent class model. (Three was chosen more or less arbitrarily for this pedagogy. In practice, one would do a more systematic search for the right number of classes.) The full covariance matrix for the random parameters (not shown) is computed using $\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{\mathrm{T}}$, where $\boldsymbol{\Gamma}$ is the lower triangular, Cholesky decomposition of the correlation matrix and $\boldsymbol{\Lambda}$ is the diagonal matrix of standard deviations

**Table 2.19**
Estimates of Stochastic Frontier Models with Parameter Heterogeneity

| | | Random Parameters | | Latent Class | | |
|---|---|---|---|---|---|---|
| Variable | Pooled | Means | SD | 1: $\pi_1$=0.2765 | 2: $\pi_2$=0.3656 | 3: $\pi_3$=0.3579 |
| Constant | −0.066983 | 0.60582 | 0.94844 | 0.97366 | −1.76168 | 2.86413 |
| Ln $y$ | 0.66914 | 0.62883 | 0.08092 | 0.48163 | 0.92320 | 0.49111 |
| $1/2 \ln^2 y$ | 0.023879 | 0.027914 | 0.00763 | 0.039745 | 0.0025294 | 0.040041 |
| Ln$w_1/w$P$_5$ | 0.38815 | 0.31048 | 0.06313 | 0.38237 | 0.444271 | 0.067207 |
| Ln$w_2/w_5$ | 0.020565 | 0.025300 | 0.05939 | 0.064287 | −0.036128 | 0.026086 |
| Ln$w_3/w_5$ | 0.17959 | 0.14430 | 0.15692 | 0.15152 | 0.22077 | −0.00040723 |
| Ln$w/w_5$ | 0.13479 | 0.10129 | 0.06767 | 0.143330 | 0.15303 | −0.018279 |
| $\lambda$ | 1.81064 | 2.27161 | 0.0 | 2.23409 | 1.20080 | 2.62612 |
| $\sigma$ | 0.31866 | 0.29715 | 0.0 | 0.39960 | 0.23755 | 0.25030 |
| $\sigma_u$ | 0.27894 | 0.27196 | 0.0 | 0.36473 | 0.18255 | 0.23392 |
| $\sigma_v$ | 0.15406 | 0.11972 | 0.0 | 0.16325 | 0.15202 | 0.089073 |
| Log -likelihood | 183.9359 | 249.0411 | | | 310.7142 | |

that are shown in table 2.19. I emphasize that the estimated "standard deviations" (SD) in table 2.19 are not standard errors (one would not divide the means by the standard deviations to compute $t$ ratios). These are the estimates of the standard deviations of the marginal distributions of the parameters distributed across the banks in the sample. The sampling "standard errors" are not shown below. As the results suggest, particularly in the latent class model, there is a fair amount of variation across firms in the frontier model parameters. For the present purposes, the more important question is the impact on the estimated inefficiencies. This is shown in the composite scatter plots in figure 2.21. The two upper figures plot the heterogeneous models against the pooled, base-case stochastic frontier model. The lower panel plots the two random-parameters models. There is, as one might expect, strong similarity across the three sets of estimates. Nonetheless, it is evident that the effects are not negligible. To assess whether the greater generality of the random-parameters approaches are indicated as necessary by the data, we can revert back to a likelihood ratio test. For the random-parameters model, the chi-squared statistic is 130.21 with 28 degrees of freedom (the number of free elements in $\mathbf{\Sigma}$). The critical value is 41.33, so the hypothesis of homogeneity would be rejected. For the latent class model, the chi-squared statistic is 253.56. The number of degrees of freedom is unclear, since if the parameters are constrained across the three classes, the same model results regardless of the unconditional values of $\pi_j$. This suggests that 18 is the appropriate count. If, instead, one must also assume that the three values of $\pi_j$ equal one-third, then 20 is the appropriate count. In either case, the critical value would be far below the sample statistic. Note, finally, that the framework does not provide an obvious way to choose between continuous and discrete parameter variation.

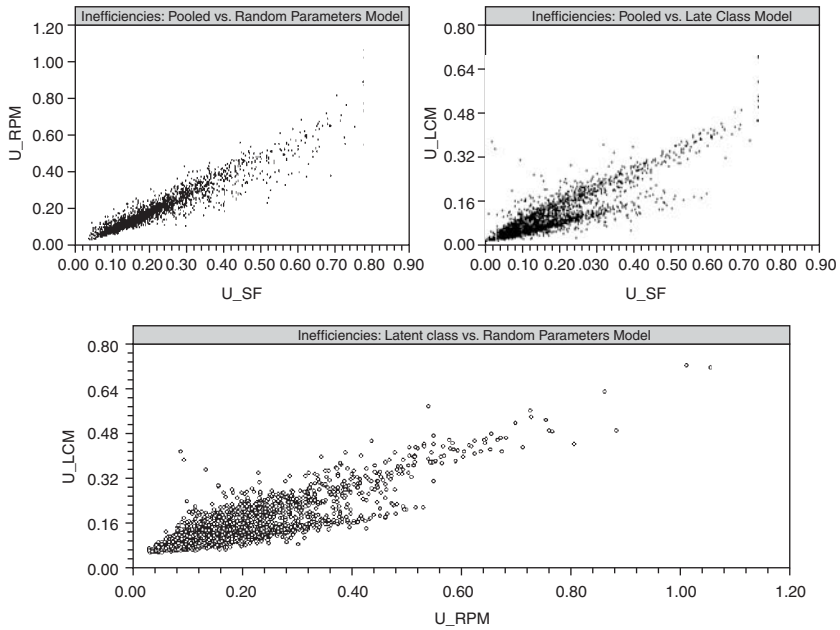


**Figure 2.21.** Estimated Inefficiencies from Pooled (top) and Random-Parameter (bottom) Models

### 2.10.5 Heterogeneity in production: WHO data

These data are a country-level panel on health care attainment. The two main variables of interest are "disability-adjusted life expectancy" (DALE) and "composite health attainment" (COMP). The former is a standard variable used to measure health care attainment. The latter is an innovative survey-based measure created by the researchers at WHO. The health attainments are viewed as the outputs of a production (function) process and were modeled in this fashion by WHO (2000) and Greene (2004b). Two input variables are health expenditure (HEXP) and education levels (EDUC). There are a number of other covariates in the data set that I view as shifters of the production function or as influences on the level of inefficiency, but not direct inputs into the production process. The data are measured for five years, 1993–1997. However, only COMP, DALE, HEXP, and EDUC actually vary across the years; the other variables are time invariant, dated 1997. In addition, as discussed by Gravelle et al. (2002a, 2002b), among others, there is relatively little actual time (within country) variation in these data; the within-groups variation for the time-varying variables accounts for less than 2% of the total. This rather limits what can be done in terms of panel-data analysis. However, in spite of this limitation, this data set provides an interesting platform for placing heterogeneity in a stochastic frontier model. [The examples to follow will build on Greene (2004b).] The WHO data are described in table 2.20.

**Table 2.20**
World Health Organization Data on Health Care Attainment

| Variable | Mean | SD | Description |
|---|---|---|---|
| COMP | 75.0062726 | 12.2051123 | Composite health care attainment |
| DALE | 58.3082712 | 12.1442590 | Disability-adjusted life expectancy |
| HEXP | 548.214857 | 694.216237 | Health expenditure per capita, PPP units |
| EDUC | 6.31753664 | 2.73370613 | Education, years |
| WBNUMBER | 138.989286 | 79.8358634 | World Bank country number |
| COUNTRY | 97.3421751 | 54.0810680 | Country number omitting internal units |
| OECD | 0.279761905 | 0.449149577 | OECD member country, dummy variable |
| SMALL | 0.373809524 | 1.20221479 | Zero or number if internal state or province |
| YEAR | 1995.21310 | 1.42464932 | Year (1993–1997) (T = year — 1992; Tyy = year dummy variable) |
| GDPC | 8135.10785 | 7891.20036 | Per capita GDP in PPP units |
| POPDEN | 953.119353 | 2871.84294 | Population density per square Kilometer |
| GINI | 0.379477914 | 0.090206941 | Gini coefficient for income distribution |
| TROPICS | 0.463095238 | 0.498933251 | Dummy variable for tropical location |
| PUBTHE | 58.1553571 | 20.2340835 | Proportion of health spending paid by government |
| GEFF | 0.113293978 | 0.915983955 | World bank government effectiveness measure |
| VOICE | 0.192624849 | 0.952225978 | World bank measure of democratization |

I have placed these data on my home page (http://http://www.stern.nyu.edu/∼wgreene (Publications))for the interested reader who wishes to replicate or extend our results. Some of the variables listed in table 2.20 (e.g., PUBTHE, SMALL) are not used here but are listed as a guide for the reader. These data and the issue they address have been analyzed and discussed widely by researchers at many sites. Greene (2004b) is part of that discussion. I do not replicate any of these studies here. Rather, we will use a subset of the data set (actually, most of it) to examine a few additional models that were not estimated above. Note some features of the data set and analysis: First, the WHO data consist of an unbalanced panel on 191 countries plus a large number of smaller political units (e.g., states of Mexico, Canadian provinces); 140 of the countries were observed in all five years (1993–1997), one (Algeria) was observed in four years, and the remaining units were all observed once, in 1997. Purely

for convenience and for purposes of our pedagogy here, we will limit our attention to the balanced panel of the 140 countries observed in all five years. Second, given that the outcome variables in the model (life expectancy and composite health care attainment) are not obviously quantitative measures such as cost or physical output units, the numerical values of efficiency measures ($u_{it}$) have ambiguous meaning. To accommodate this, the researchers at WHO focused their attention on rankings of efficiency measures, rather than on values. Third, the WHO study was innovative in several respects, notably in its attempt to include many (all) countries, in contrast to above studies that nearly always focused on the 30 member countries of the Organisation for Economic Co-operation (OECD). However, little attention was paid in the WHO studies (Evans et al., 2000a, 2000b) to the distinction between OECD and non-OECD countries in the results, perhaps by design. Greene (2004b) found a striking distinction in the results between the two groups. In short, nearly all of the "action" in the inefficiency distributions pertains to the non-OECD observations. The OECD countries area always clustered near the origin. This is an important angle that might be pursued in further analysis.

The WHO study treated the process of health care provision at the national level as a production process,

$$\text{health}_{it} = f(\text{education}_{it}, \text{expenditure}_{it}).$$

Whether it is reasonable to view the outcome here as an optimization process in which agents maximized the production of "health" while using these two inputs is, of course, debatable. For better or worse, the model employed is

$$\ln \text{health}_{it} = \alpha + \beta_1 \ln \text{HEXP}_{it} + \beta_2 \ln \text{EDUC}_{it} + \beta_3 \ln^2 \text{EDUC}_{it} + v_{it} - u_{it}.$$

Differences among subsequent researchers concerned the functional form, the stochastic specification, and the method of handling the cross heterogeneity. We will explore a few of those issues here, though not with an eye toward commenting on other received analyses. We are interested in two modeling aspects in this section. As noted above, in some applications, notably this one, there are covariates that arguably affect production and/or efficiency. The modeling question raised above is, "where do we put the $z$'s?" That is, how should measured heterogeneity enter the model? Unfortunately, there are numerous choices, and no clearly right answer, as will become obvious presently. The number of possibilities is yet doubled here, as we have two outcome variables to study. Without attempting to resolve the question, I present a handful of model estimates under different formulations to illustrate the techniques. We have no basis on which to prefer any particular one at this juncture. The interested reader may wish to continue the analysis. The second feature we examine, briefly further below, is the extent to which accommodating measured (and unmeasured) heterogeneity affects estimates of inefficiency. It is straightforward to make a case that,

under most reasonable specifications, inappropriate treatment of heterogeneity will distort estimates of inefficiency. Whether it will affect rankings of inefficiencies, which were the focus of attention in the WHO study, is, however, unresolved.

#### 2.10.5.1 Accommodating measured heterogeneity

We will define the vectors

$$\begin{aligned}\mathbf{x}_{it} &= \ln \text{HEXP}_{it}, \ln \text{EDUC}_{it}, \ln^2 \text{EDUC}_{it},\\ \mathbf{z}_{i,p} &= \text{TROPICS}_i, \ln \text{POPDEN}_i,\\ \mathbf{z}_{i,e} &= \text{GINI}_i, \ln \text{GDPC}_i, \text{GEFF}_i, \text{VOICE}_i, \text{OECD}_i.\end{aligned}$$

Note that the latter two are time invariant; only the outputs and inputs are measured in all years. We will condition the production directly on $\mathbf{z}_{i,p}$. The other vector of covariates will enter the efficiency models at various points as shown below. Obviously, other arrangements of the variables are possible. It seems natural that location and population density are less policy related than the other variables and so appear more naturally as shift parameters in the production function. Other assignments might also seem appropriate; the interested reader may wish to experiment—for example, Greene (2004b) also included a time trend in $\mathbf{z}_{i,e}$. Tables 2.21–2.23 present estimates for the following models:

*Stochastic Frontier: Normal–Half-Normal (Aigner et al., 1977)*

$$\begin{aligned}&\ln \text{health}_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + \boldsymbol{\theta}_p^{\mathrm{T}}\mathbf{z}_{i,p} + \boldsymbol{\theta}_e^{\mathrm{T}}\mathbf{z}_{i,e} + v_{it} - u_{it}\\ &v_{it} \sim N[0, \sigma_v^2]\\ &u_{it} = |U_{it}|,\ U_{it} \sim N[0, \sigma_u^2]\end{aligned}$$

*Normal–Truncated Normal (Stevenson, 1980)*

$$\begin{aligned}&\ln \text{health}_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + \boldsymbol{\theta}_p^{\mathrm{T}}\mathbf{z}_{i,p} + v_{it} - u_{it}\\ &v_{it} \sim N[0, \sigma_v^2]\\ &u_{it} = |U_{it}|,\ U_{it} \sim N[\mu + \boldsymbol{\theta}_e^{\mathrm{T}}\mathbf{z}_{i,e}, \sigma_u^2]\end{aligned}$$

*Heteroskedastic Normal (singly or doubly; Hadri, 1999, and Hadri et al., 2003a,b)*

$$\begin{aligned}&\ln \text{health}_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + \boldsymbol{\theta}_p^{\mathrm{T}}\mathbf{z}_{i,p} + v_{it} - u_{it}\\ &v_{it} \sim N[0, \sigma_{vi}^2];\ \sigma_{vi} = \sigma_v \times \exp(\boldsymbol{\gamma}_{pv}^{\mathrm{T}}\mathbf{z}_{i,e})\\ &u_{it} = |U_{it}|,\ U_{it} \sim N[0, \sigma_{ui}^2];\ \sigma_{ui} = \sigma_u \times \exp(\boldsymbol{\gamma}_{pu}^{\mathrm{T}}\mathbf{z}_{i,e})\end{aligned}$$

**Table 2.21**
Estimated Heterogeneous Stochastic Frontier Models for lnDALE

| Variable | Half-Normal Model | | | Truncated-Normal Model |
|---|---|---|---|---|
| Constant | 3.50708 | 3.50885 | 3.28004 | 3.90626* |
| EXP. | 0.066364 | 0.065318 | 0.019171 | 0.03532* |
| EDUC. | 0.288112 | 0.307518 | 0.277322 | 0.22911* |
| EDUC.$^2$ | −0.110175 | −0.12711 | −0.11729 | −0.12480* |
| TROPICS | | −0.025347 | −0.016577 | −0.12480 |
| LnPOPDEN | | 0.0013475 | −0.00028902 | 0.0014070 |
| | Shift Production Function | | Production Function | Mean of $U_i$ |
| Constant | | | | 2.33052* |
| GINI | | | −0.21864 | −0.090319 | 1.90648* |
| LnGDPC | | | 0.072409 | −0.0096963 | −0.40107* |
| GEFF | | | −0.0088535 | 0.010164 | — |
| | | | | 0.0047021 | |
| VOICE | | | 0.012679 | 0.016304* | — |
| | | | | 0.00092454 | |
| OECD | | | −0.045681 | −0.018195 | −2.82321 |
| | Noise and Inefficiency Distributions | | | |
| $\lambda$ | 5.72629 | 5.19739 | 6.31057 | 9.92754 |
| $\sigma$ | 0.21063 | 0.20669 | 0.20223 | 0.20818 |
| $\sigma_u$ | 0.21063 | 0.20297 | 0.19974 | 0.20713 |
| $\sigma_v$ | 0.03623 | 0.03905 | 0.03165 | 0.02086 |
| Log-likelihood | 501.4585 | 506.1130 | 536.9086 | 859.4868 |

* Statistically significant at the 95% level.

The truncation and heteroskedasticity models can be combined and permuted. The formulation of the Alvarez et al (2006). scaling model shows one possibility:

*Scaling (Alvarez, Amsler, Orea and Schmidt, 2006)*

$$\ln \text{health}_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + \boldsymbol{\theta}_p^{\mathrm{T}}\mathbf{z}_{i,p} + v_{it} - u_{it}$$

$$v_{it} \sim N[0, \sigma_v^2]$$

$$u_{it} = |U_{it}|,\ U_{it} \sim N[\mu_i, \sigma_{ui}^2];\ \mu_i = \mu \times \exp(\boldsymbol{\gamma}_{pu}^{\mathrm{T}}\mathbf{z}_{i,e})\sigma_{ui}$$

$$= \sigma_u \times \exp(\boldsymbol{\gamma}_{pu}^{\mathrm{T}}\mathbf{z}_{i,e})$$

Note that the scale factors on the mean and standard deviation of the distribution of $u_{it}$ are identical. This constraint can be relaxed, though if so, the model no longer obeys the scaling property suggested by the authors. Alvarez et al. suggested linear rather than loglinear scale factors. This is potentially problematic since the linear function is not constrained to be positive, and it is not possible to impose the constraint on the optimization procedure. As a final candidate for a platform for the measured heterogeneity, we consider a latent class formulation in which allows both the production and efficiency

**Table 2.22**
Estimated Heteroskedastic Stochastic Frontier Models

| Variable | Half-Normal | Hetero-skedasticity in $u$ | Heteroskedasticity in Both $u$ and $v$ | | Heteroskedasticity in $u$ and $v$; $u$ Time Invariant | | Scaling Model; Heterogeneity in $E[U]$, Same Scale for $\sigma_u$ | |
|---|---|---|---|---|---|---|---|---|
| Constant | 3.28004 | 3.67243 | 3.69419 | | 3.91430 | | 3.64980 | |
| EXP. | 0.019171 | 0.037812 | 0.04056 | | 0.016895 | | 0.041866 | |
| EDUC. | 0.277322 | 0.34194 | 0.31867 | | 0.10816 | | 0.32555 | |
| EDUC.$^2$ | −0.11729 | −0.17715 | −0.17415 | | 0.011575 | | −0.16676 | |
| TROPICS | −0.016577 | −0.011027 | −0.010097 | | 0.025598 | | −0.008782 | |
| lnPOPDEN | −0.000289 | 0.000299 | 0.00028812 | | 0.003334 | | −0.000493 | |
| | | | $\sigma_v$ | $\sigma_u$ | $\sigma_v$ | $\sigma_u$ | $\mu$ | $\sigma_u$ |
| Constant | | 1.78679 | −1.23320 | 1.29694 | 0.75772 | 0.68495 | −23.1551 | 2.08169 |
| GINI | | 9.64724 | −3.19744 | 10.34564 | 4.28185 | 7.20613 | 5.77792 | |
| LnGDPC | | −1.13106 | −0.64829 | −1.10138 | −1.0714 | −0.63463 | −0.65734 | |
| GEFF. | | −0.24609 | −0.93913 | −0.169847 | −0.27222 | −0.41316 | −0.050936 | |
| VOICE | | 0.14326 | 0.039271 | 0.055723 | 0.58151 | 0.082344 | −0.046971 | |
| OECD | | −1.65801 | 1.562677 | −2.09808 | −0.27988 | 0.020814 | −1.17259 | |
| | | | Variance Parameters | | Variance Parameters | | Variance Parameters | |
| $\lambda$ | 6.31057 | 4.08669[a] | 4.80460[a] | | 2.11404[a] | | 15.57378[a] | |
| $\sigma$ | 0.20223 | 0.24556[a] | 0.25104[a] | | 0.75772[a] | | 0.45450[a] | |
| $\sigma_u$ | 0.19974 | 0.24467[a] | 0.28975[a] | | 0.68495[b] | | 0.45382[a] | |
| $\sigma_v$ | 0.03165 | 0.05987 | 0.06218[a] | | 0.32400[b] | | 0.024914[a] | |
| Log-likelihood | 536.9086 | 812.9505 | 829.5840 | | 1910.944 | | 817.2499 | |

[a] Computed by averaging the sample estimates of country-specific variances

**Table 2.23**
Estimated Latent Class Stochastic Frontier Model

| Variable | Half-Normal | Class 1 | Class 2 |
|---|---|---|---|
| Constant | 3.28004 | 3.53884 | 2.91203 |
| EXP. | 0.019171 | 0.044493 | 0.025945 |
| EDUC. | 0.277322 | 0.33199 | −0.072499 |
| EDUC. | −0.11729 | −0.15674 | 0.12832 |
| TROPICS | −0.016577 | −0.001768 | −0.0079229 |
| LnPOPDEN | −0.00028902 | −0.0033528 | 0.0058591 |
| GINI | −0.21864 | −0.185551 | −0.48646 |
| LnGDPC | 0.072409 | 0.016297 | 0.12076 |
| GEFF. | −0.0088535 | 0.00056079 | 0.13722 |
| VOICE | 0.012679 | 0.013583 | −0.17573 |
| OECD | −0.045681 | −0.022626 | 0.10688 |
| Class probability | 1.00000 | 0.83916 | 0.16084 |
| $\lambda$ | 6.31057 | 1.86032 | 8.50170 |
| $\sigma$ | 0.20223 | 0.071261 | 0.11716 |
| $\sigma_u$ | 0.19974 | 0.062768 | 0.116365 |
| $\sigma_v$ | 0.03165 | 0.033740 | 0.013687 |
| Log-likelihood | 536.9086 | 1011.858 | |

heterogeneity to enter the production function, and the efficiency heterogeneity also to enter the class probabilities. The model is

*Latent class (Greene, 2004a; Orea and Kumbhakar, 2004)*

$$\ln \text{health}_{it}|j = \alpha_j + \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_{it} + \boldsymbol{\theta}_{p,j}^{\mathrm{T}}\mathbf{z}_{i,p} + \boldsymbol{\theta}_{e,j}^{T}\mathbf{z}_{i,e} + v_{it} - u_{it}$$
$$v_{it}|j \sim N[0, \sigma_{v,j}^2]$$
$$u_{it}|j = |U_{it}|j|,\ U_{it}|j \sim N[0, \sigma_{uj}^2]$$
$$\text{Class probability: } \pi_{i,j} = \exp(\tau_{0j} + \boldsymbol{\tau}_j^{\mathrm{T}}\mathbf{z}_{i,e}) / \textstyle\sum_j \exp(\tau_{0j} + \boldsymbol{\tau}_j^{\mathrm{T}}\mathbf{z}_{i,e})$$

The latent class model can be applied in a cross-section or pooled model. Since we have a panel model, we will fit this as such—the force of the treatment is that the class probabilities apply unchanged to all five observations for each country.

My purpose here is to illustrate computation of the models. The JLMS estimator of $E[u|\varepsilon]$ is computed using all the above results. Since there are so many combinations of the models available, each with its own implied set of estimates, I forgo a secondary analysis of the implied inefficiency estimates, with one exception. An important specification issue for analysts—the subject of this exercise—is the effect of the measured covariates, the "*z*s," on estimates of $E[u|\varepsilon]$ or $E[\exp(-u)|\varepsilon]$. To pursue this issue, researchers often estimate the generic frontier model without the covariates and then,

**Table 2.24**
Second-Step Regression of Estimates of $E[u|\varepsilon]$ on Covariates

| | Stochastic Frontier Production | | Heterogeneous Truncated Normal | |
|---|---|---|---|---|
| Variable | Estimate | *t*-Ratio | Estimate | *t*-Ratio |
| Constant | 0.27632 | 8.802 | 0.23059 | 14.790 |
| Tropics | 0.018463 | 3.431 | 0.0021116 | 0.790 |
| LnPOPDEN | −0.0010252 | −0.905 | −0.00024979 | −0.444 |
| GINI | 0.15700 | 5.537 | 0.056483 | 4.011 |
| lnGDPC | −0.027559 | −7.842 | −0.019621 | −11.243 |
| GEFF. | 0.010052 | 2.165 | 0.0039423 | 1.710 |
| VOICE | −0.0031805 | −0.888 | −0.0025433 | −1.431 |
| OECD | 0.035059 | 4.661 | −0.017854 | −4.780 |
| $R^2$ | 0.2576678 | | 0.5346847 | |
| SE | 0.0533936 | | 0.026517 | |

in a second step, regress the estimated (in)efficiencies on the covariates. Wang and Schmidt (2002) have cautioned against this, arguing that the omission of the covariates at the "first step" is tantamount to the omitted variable problem in ordinary regression. Nonetheless, this procedure is fairly common and, indeed, is routine in the DEA literature. (In fact, the first step of DEA provides no mechanism for including the $z$ values in the model, so this is to be expected.) Table 2.24 shows a second-step analysis of the estimates from the generic model and from the truncated regression model.

Table 2.21 shows the impact of introducing the observed indicators of heterogeneity directly into the production model and into the mean of $U_i$. The first three columns show the estimation of the half-normal model with progressively more extensive lists of covariates. The base production parameters change relatively little. However, the estimate of $\sigma_u$ gets progressively smaller, though less than we might have expected. The last set of results shows the normal–truncated-normal model, with the full set of effects both in the production function and in the inefficiency distribution. Intuition might suggest, however incorrectly, that estimation of the model with the variables in both the production function and in $E[U_i]$ would be difficult because of weak identification—a multicollinearity problem, if nothing else. In fact, estimation of the model was routine. For the example, coefficients that were "significant" in this model are indicated by asterisks. Of the 19 parameters estimated in the full model, 12 "$t$-ratios" were larger than 1.0, and only three were less than 0.5. Figure 2.22 shows the kernel density estimators for the sample of estimates of $E[u_i|\varepsilon_i]$ for the least specified model, at the left, and the most general model, at the right. The x-axes of the two figures are the same. The much tighter distribution of the latter is consistent with expectations about introducing heterogeneity into

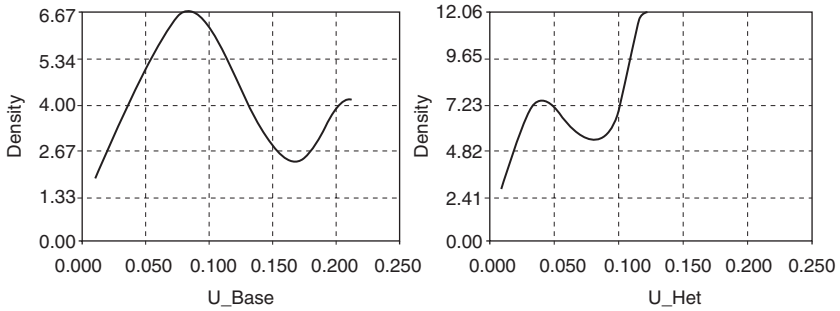


**Figure 2.22.** Kernel Density Estimates for Inefficiency Distributions

the model. (I have no theory at this point for the bimodality of the two estimates.)

Table 2.22 displays the estimated parameters for models in which the heterogeneity appears in the variances rather than the means. The results do illustrate the computations. It is difficult to frame a prior for whether heterogeneity in the mean of the variance would be the preferred model. That would depend on the application. One interesting outcome is shown in figure 2.23, which plots the estimates of $E[u|\varepsilon]$ for the doubly heteroskedastic model. Though the shape of the distribution is more in line with priors, its range is much larger than that for the preceding model, in which the heterogeneity is in the mean. This may suggest a basis on which to formulate the preferred model. The third set of results displays the Alvarez Amsler, Orea and Schmidt (2006). "scaling model." Again, it is difficult to form priors, but note here that the assumption of the scaling model has the result that nearly all of the variation in $\varepsilon$ (and some not found before) is shifted from $v$ to $u$, compared to the truncation model in table 2.21 and the doubly heteroskedastic model in table 2.22.

The final set of results, in table 2.23, show a two-class latent class model. In the model estimated, the efficiency covariates, $z_{i,e}$, are also determinants of the class probabilities (in the form of a binomial logit model with these as the independent variables).

Table 2.24 displays regression results when the JLMS estimates of $E[u|\varepsilon]$ are regressed on the observed indicators of heterogeneity. The estimates computed from the half-normal stochastic frontier model contain only expenditure, education, and its square. In those computed from the normal–truncated-normal model, all variables listed appear in the production function, and the GINI coefficient, lnGDP, and so on, also appear in the mean of the inefficiency distribution. Table 2.24 reveals that the heterogeneity significantly improves the prediction of $E[u|\varepsilon]$. The stochastic frontier results confirm our expectation, that omitted heterogeneity is an important element of the

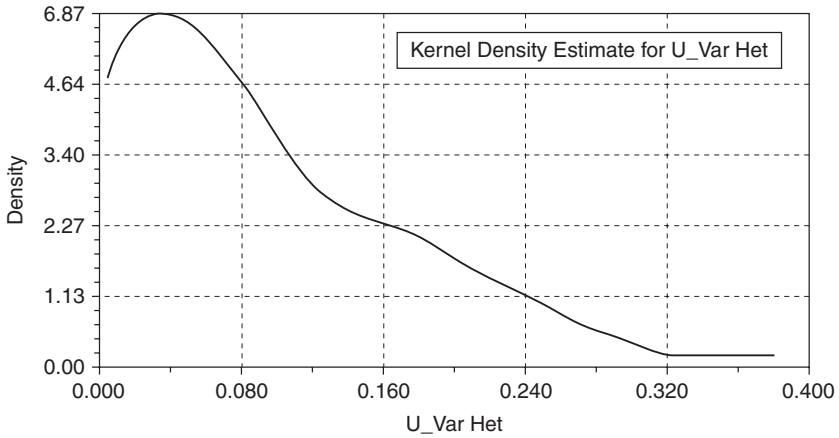


**Figure 2.23.** Kernel Density for Inefficiencies on Doubly Heteroskedastic Model

measurement of inefficiency in these data. Intuition might suggest something amiss in the normal–truncated-normal results. Since the heterogeneity is already in the equation, shouldn't it be absent from the residuals? Yes, but no, because the JLMS estimator of $E[u|e]$ is not the residual; it is explicitly a function of the data. Thus, there is no surprise in the normal–truncated-normal results in table 2.24. Note also that the fit of the "regression" is considerably in the truncated-normal model. The much lower value of $s$ (0.0265 vs. 0.0534) reflects the purging of these heterogeneity effects from the estimates of inefficiency. Table 2.24 casts no light on whether the omission of heterogeneity has significantly impacted the estimates of $E[u|\varepsilon]$, save for the apparently lower values. Figure 2.24 plots the two sets of estimates against each other.[88] What the figure reveals is that there is much less correlation between the two than one might hope for—the simple correlation is about 0.7. If we correlate the ranks, instead, the rank correlation is about 0.65. As a final exercise, we compute the country means of the estimates and then compute the ranks. The scatter plot of the two sets of ranks is shown in figure 2.25. The plot is illuminating. It shows, first, that, in fact, the rankings are crucially dependent on the treatment of heterogeneity. This was the original premise in Greene (2004b). Second, the nicely arranged line of points at the upper left of the figure consists of the 30 OECD countries whose high rankings (low estimates) are clearly evident.

#### 2.10.5.2 The effect of mishandled heterogeneity on inefficiency measurement

The possibility that unmeasured heterogeneity could masquerade as technical inefficiency has at least an intuitive appeal. To examine the issue, let us compare

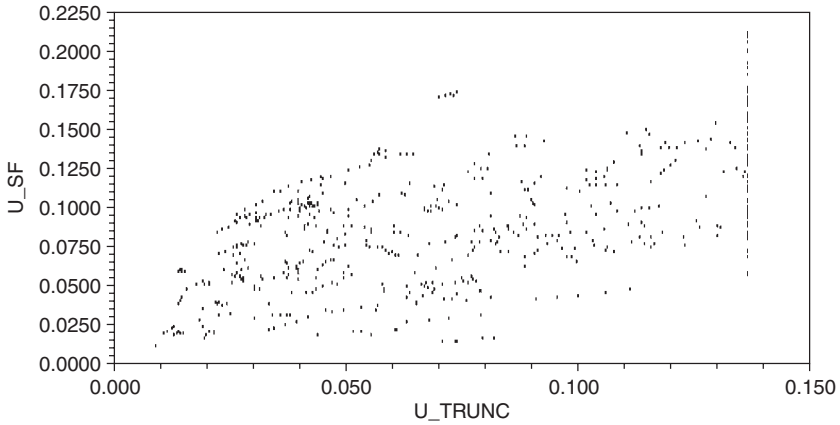


**Figure 2.24.** Scatter Plot of Estimated Inefficiencies

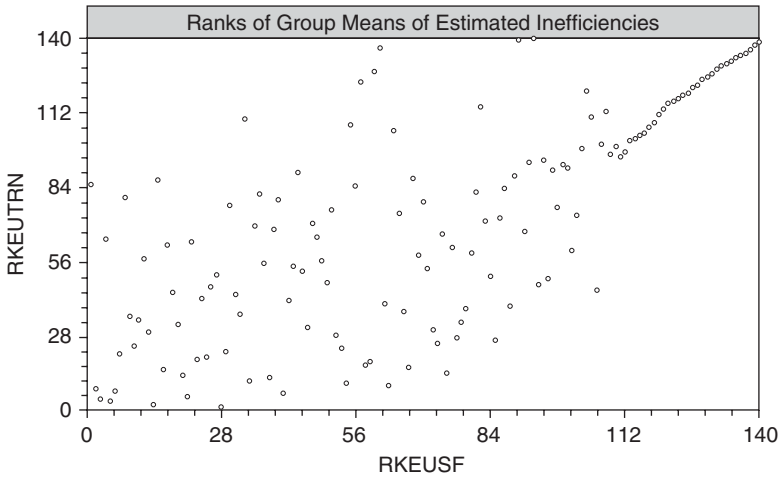


**Figure 2.25.** Plot of Ranks of Group Means of Inefficiencies

the estimates of $E[u|\varepsilon]$ from a generic, surely underspecified model,

$$\ln \text{health}_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it},$$

to several alternatives:

- True random effects: $\ln \text{health}_{it} = (\alpha + w_i) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it}$
- True fixed effects: $\ln \text{health}_{it} = \alpha_i + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + v_{it} - u_{it}$

- Heterogeneous truncated-normal model

$$\ln \text{health}_{it} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it} + \boldsymbol{\theta}_p^{\mathrm{T}}\mathbf{z}_{i,p} + v_{it} - u_{it}$$
$$v_{it} \sim N[0, \sigma_v^2]$$
$$u_{it} = |U_{it}|, U_{it} \sim N[\mu + \boldsymbol{\theta}_e^{\mathrm{T}}\mathbf{z}_{i,e}, \sigma_u^2]$$

In each case, our expectation is that the explicit treatment of heterogeneity (unobserved or measured) should purge the disturbance of this effect. The interesting question is whether the effect being purged would have initially been placed in $u_{it}$ (our conditional mean estimate of it) or in $v_{it}$. [Note that there is no ambiguity about the outcome in the deterministic frontier methodology analyzed, e.g., in Cornwell et al. (1990) and in Schmidt and Sickles (1984) or in the Pitt and Lee (1981) random-effects model. A demonstration of the effect in these data appears in Greene (2004a, 2004b).]

Table 2.25 gives descriptive statistics for the four sets of estimates of $E[u_{it}|\varepsilon_{it}]$ for both health outcomes. Note, first, that the assumption of the true random-effects model, that the unobserved heterogeneity is uncorrelated with the observed variables, seems extremely unlikely to be correct. The results in table 2.25 seem consistent with this: The estimated inefficiencies are an order of magnitude smaller than the other three. For the others, we see the anticipated effect. The average values are significantly smaller for the models that accommodate heterogeneity (truncation and true fixed effects). The kernel density estimators in figure 2.26 show that the latter distributions are also much tighter. The left pair is for *DALE*; the right is for *COMP*. The upper figure of each pair is the density estimator for the results based on the true fixed-effects estimator. The lower one is the estimator for the base model with no terms for heterogeneity.

**Table 2.25**
Descriptive statistics for Estimates of $E[u|\varepsilon]$

| Model | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|
| | | Ln DALE | | |
| Base | 0.11580 | 0.061660 | 0.12211 | 0.21060 |
| True fixed effect | 0.077081 | 0.012237 | 0.042582 | 0.17549 |
| True random effect | 0.011091 | 0.0059746 | 0.0013537 | 0.074813 |
| Truncation | 0.088570 | 0.043287 | 0.0094572 | 0.13648 |
| | | Ln COMP | | |
| Base | 0.069964 | 0.034603 | 0.0075750 | 0.11065 |
| True fixed effect | 0.042728 | 0.010689 | 0.018934 | 0.13264 |
| True random effect | 0.0 | 0.0 | 0.0 | 0.0 |
| Truncation | 0.038745 | 0.014894 | 0.00308415 | 0.048302 |

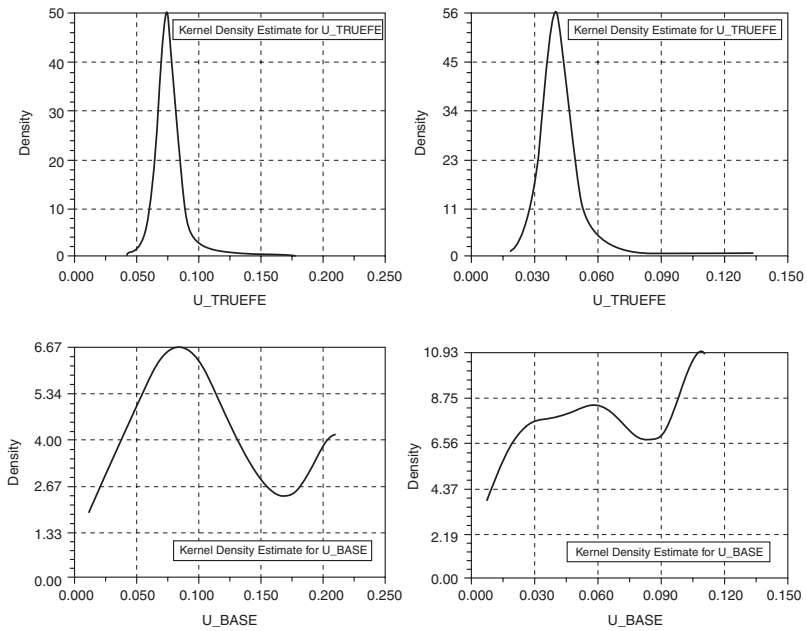


**Figure 2.26.** Kernel Density Estimators for Estimates of $E[uE]$

## 2.11 Conclusions

Current practice includes two approaches to efficiency measurement: the programming approach and the econometric approach. The deterministic frontier models presented in section 2.3 represent a hybrid of these two approaches. Although it is difficult to draw general conclusions from a single study, the results of this one concur with the common perception that the main advantage of the econometric approach lies in its ability to shift the deleterious effect of measurement error away from estimates of efficiency. The values produced by the deterministic estimators section 2.10.3 seem not only to be implausibly large, but also to distort the expected relationship between cost and production frontiers.

The stochastic frontier approach has a number of virtues, notably its internal consistency and its ease of implementation. For single-equation, cross-section analysis, with modern computer software, the stochastic frontier model is not appreciably more complex than a linear regression model. The possibility of adding a shift parameter to it, and the numerous interesting ancillary calculations derived by Jondrow et al. (1982) and Battese and Coelli (1992, 1995) suggest that the half-normal model is the most useful formulation. Other variants such as the truncated-normal model with heterogeneity in the mean allow for great flexibility in the modeling tools.

Panel data open up numerous interesting possibilities. Approaches based on regression analysis of the fixed- and random-effects models have the appeal of robustness and the potential for a consistent estimator of inefficiency. The fixed-effects model does carry with it the necessity that the analyst revert back, essentially, to the deterministic frontier model. The random-effects model, on the other hand, has the appeal of the single-equation stochastic frontier. However, as in other settings, the drawback to this approach is that the effects must be assumed to be uncorrelated with the regressors (factors). This is likely to be minor in this context. It is routinely assumed in any event. The impact of the assumption of time invariance of inefficiency seems to be the one large effect of model specification. Consistent with other researchers, we have found in this study that estimates of technical and cost inefficiency are quite robust to distributional assumptions, to the choice of fixed or random effects and to methodology, Bayesian versus classical, but they are quite sensitive to the crucial assumption of time invariance (or the lack thereof).

## Notes

1. Some econometric issues that arise in the analysis of primal productions and dual cost functions are discussed in Paris and Caputo (2004).
2. Some current research is directed at blurring this distinction by suggesting a statistical underpinning for DEA. Because DEA is the subject of subsequent chapters in this book, I do not visit the topic here.
3. A more complete listing appears in chapter 1.
4. A second edition of the latter is forthcoming as of this writing.
5. This does not fully solve the problem of zero values in the data, because the appropriate standard errors for the Box-Cox model still require the logarithms of the variables. See Greene (2003a, p. 174).
6. A few other applications do note the idea, including Koop et al. (1994, 1997), Tsionas (2002), and Kumbhakar and Tsionas (2005a). Mention of the "regularity conditions" (to be kept distinct from the regularity conditions for maximum likelihood estimators) is common in the frontier applications, though relatively few actually impose them. It is more common to "check" the conditions after estimation. For example, Farsi and Filippini (2003) estimated a translog cost frontier for Swiss nursing homes and observed ex post that the estimated parameters did not satisfy the concavity conditions in the input prices. This result was attributed to the price-setting mechanism in this market.
7. Førsund et al. (1980, pp. 21–23) argue that economic dogma has essentially painted its proponents into a corner. Every action taken by an economic agent must be efficient, or else it would not have been taken. This takes a bit of artful footwork in some cases.
8. See chapter 1 for a more complete exposition.
9. There is a tendency on the part of many authors in economics to equate an *estimation technique* with a *model*. In few places is this more evident than in the literature on DEA.

10. A crucial assumption that was discussed early in this literature, but is now implicit, is that there is no correlation between $\mathbf{x_i}$ and $\varepsilon_i$ in the model. Refer to Zellner, Kmenta, and Dreze (1966) for discussion of the proposition that deviations of the observed factor demands $\mathbf{x_i}$ from the cost-minimizing or profit-maximizing values could be uncorrelated with the deviation of $y_i$ from its ideal counterpart as specified by the production function. Further discussion and a model specification appear in Sickles and Streitweiser (1992).

11. See Richmond (1974) for the application to this distribution. Afriat (1972) examined $\text{TE}_i$ similarly under the assumption of a beta distribution.

12. See Greene (1980a) for discussion.

13. See Greene (1980a) for technical details and Deprins and Simar (1985, 1989b) for some corrections to the derivations.

14. See Christensen and Greene (1976) for discussion. The outer transformation is strictly monotonic, and the inner function is linearly homogeneous.

15. The constrained linear programming solution is not the maximizer of the log-likelihood function.

16. This is an antecedent to the recent DEA literature (e.g., Bankar, 1993, 1997) that has attempted to cast the linear programming approach as the maximizer of a log-likelihood function. An application, among many, that compares econometric approaches to this linear programming methodology is Ferrier and Lovell (1990).

17. As we can see from the expression for $E[e^{-u_i}]$, when $\theta = 1$, $E[e^{-u_i}]$ is $2^{-P}$, which is Richmond's result.

18. For discussion, see Lovell (1993), Ali and Seiford (1993), and chapter 3 of this volume.

19. For extensive commentary on this issue, see Schmidt (1985). Banker and Maindiratta (1988) show how DEA gives an upper bound for efficiency. With input price data, one can also use the technique to compute a lower bound.

20. An application that appeared concurrently is Battese and Corra (1977).

21. Recent research has begun to investigate the possibility of correlation across the two components of the composed disturbance. The econometric issues are considerable; e.g., identification is a problem. The underlying economics are equally problematic. As of this writing (mid-2007), the returns on this model extension are far from complete, so I eschew further consideration of the possibility.

22. See Schmidt and Lin (1984) for this test and Coelli (1995) for a slightly different form of the test. The statistic appears in a slightly different form in Pagan and Hall (1983).

23. This need not be the case. The skewness of $\varepsilon_i$ is entirely due to $u_i$, and as long as $u_i$ is positive, in fact, the skewness could go in either direction. Nonetheless, in the most common formulations of the stochastic frontier model, involving the normal distribution, the skewness provides an important diagnostic check on the model specification.

24. See Train (2003), Greene (2003a, section 17.8; 2005), and Greene and Misra (2003).

25. The derivation appears in many other sources, e.g., Pitt and Lee (1981), Greene (1990), and Kumbhakar and Lovell (2000).

26. An alternative parameterization that is convenient for some other forms of the model is $\gamma = \sigma_u^2/\sigma^2$. See Battese and Corra (1977), Battese (1992), Coelli (1991), and Greene (2000, chapter 28).

27. The standard statistics, LM, Wald, and LR, are quite well defined, even at $\lambda = 0$, which presents something of a conundrum in this model. There is, in fact, no problem computing a test statistic, but problems of interpretation arise. For related commentary, see Breusch and Pagan (1980). The corresponding argument regarding testing for a one-sided error term would be the same. In this case, the parametric "restriction" would be $\lambda \to +\infty$ or $(1/\lambda) \to 0$, which would be difficult to test formally. More encouraging and useful results are given in Coelli (1995), who shows that the likelihood ratio statistic has a limiting distribution that is a tractable mix of chi-squared variates.

28. The log-likelihood for the normal–half-normal model has two roots, one at OLS with $\lambda = 0$ and one at the MLE with positive $\lambda$. In the event noted, the first solution is "superior" to the second.

29. It does create a bit of a dilemma for the practitioner. In itself, the result is an important diagnostic for the model specification. However, it does not carry over to other model formulations and more elaborate extensions. As such, one might choose to proceed despite the warning. Then again, some of the estimators of these elaborate models use the "plain vanilla" ALS frontier estimates as starting values for the iterations. In this case, at least the warning will be heard. I note, for the benefit of the practitioner, that the occurrence of this result is not indicative of a problem with the data or the software—it signals a mismatch between the model and the data. The appropriate conclusion to draw is that the data do not contain evidence of inefficiency. A more encouraging result, however, is that this result is specific to the half-normal model above. Other formulations of the model, or more highly developed specifications, might well reveal the presence of inefficiency. That is, this finding can emerge from several sources.

30. Of course, this assumption is no less restrictive than half-normality.

31. One apparently reliable strategy is based on OLS for the slopes, and the method of moments for the remaining parameters.

32. My literature search returned roughly 35 references in total. Most are described in the various sections to follow, so I eschew a rote listing of them here. I will wait for my next edition of this survey before applying any generic appellation to the nascent Bayesian literature, because at least 16 of those 35 studies were produced by the same four authors. Suffice to say, as of this writing, the Bayesian methodology has made a significant contribution to the larger literature.

33. Other treatments that contain useful entry-level introductory material are Osiewalski and Steel (1998), Kleit and Terrell (2001), Tsionas (2001a), and Kim and Schmidt (2000).

34. Impressions (and assertions) to the contrary notwithstanding, neither Bayesian nor classical procedures estimate $u_i$, conditionally or otherwise. They estimate the conditional mean function, $E[u_i|\text{data}]$, the mean of the conditional distribution of the population that generated $u_i$. Section 2.8 revisits this issue.

35. For the production or cost model, Koop and Steel (2001) suggest a refinement to include $p(\beta) \propto$ an indicator function that includes the regularity conditions. [This device is used by Kleit and Terrell (2001).] For example, in a Cobb-Douglas model, we require the elasticities to be positive. As stated, their "indicator function" cannot actually be a "prior" in a flexible functional form, since the regularity conditions are only local and functions of the present data. Given the ambiguity, we will maintain the simpler prior over the technology parameters and leave the question to be resolved elsewhere.

36. Note a point here that appears to have been lost in the received Bayesian applications that frequently cite the shortcomings of the JLMS (Jondrow, Lovell, Materov, and Schmidt, 1982) "estimator" of $u_i$. The conditional mean being estimated at the data augmentation step of the Gibbs sampler is precisely the same conditional mean function that is computed by the classical estimator using the JLMS results. This is not surprising, of course, since, in fact, conditional means are all that can be estimated in this context under either methodology. I return to this point below.

37. The closed form for a few integer values may be found in Amemiya (1973).

38. However, there is some evidence given by van den Broeck et al. (1994) that Greene's results may have been influenced by some complications in the numerical procedures. Coincidentally, these authors (p. 17) likewise experience considerable difficulty in estimating a nonzero $\mu$ in the truncated-normal model.

39. The difficulty lies in accurately computing the moment of the truncated normal distribution [the $q(r, \varepsilon_i)$ function]. An equally complex model that has also not been used empirically is Lee's (1983) four-parameter Pearson family of distributions.

40. The authors obtain some results in this study which suggest that Greene's results were heavily influenced by the method of approximating the integrals $q(r, \varepsilon_i)$. Their results are reasonably strong, though clearly the extension to noninteger $P$ would be useful.

41. Estimation of the posterior mean of $u_{it}$ requires sampling from the truncated normal distribution. Tsionas (2002) suggests acceptance/rejection and a modification thereof for the troublesome tail areas. The inverse probability transformation discussed above would be an improvement.

42. In my literature search, I found, up to mid-2005, roughly 35 applications of Bayesian methods to frontier modeling; five of these, those mentioned above, use the Christensen and Greene (1976; CG) data, and one (Kleit and Terrell, 2001) builds on the principles in CG but uses an updated data set. [The widely circulated "classical" study by Bera and Sharma (1999) also uses these data.] Curiously, Kleit and Terrell (2001) argue that the CG data are outdated (possibly so), but also that the CG data were a "limited sample of fossil fuel electric generators" (p. 524). In fact, the CG 1970 firm-level sample contained within a percent or two the entire universe of privately owned fossil-fueled generators in the United States in 1970, whereas their updated plant-level data set included 78 of the several hundred U.S. generators in 1996. This adds a useful layer to the use of the CG data as an application. While the Bayesian methods limit their inferences to the sample data, classical ("asymptotic") methods attempt to extend the reach of the results to the broader population. But, for these data, in that year, the sample *is* the population. There remains, therefore, some scope for researchers to dig a bit deeper and examine the differences between Bayesian and classical results—small though they usually are. It is also striking that, although one of the oft-touted virtues of the Bayesian methodology is that it enables the researcher to incorporate "prior information," not one of these six studies used a single result from CG or any of the other studies in formulating their "priors" for any of the model parameters or the inefficiency estimates. In the same fashion, several studies (e.g., Tsionas, 2000b; Smith, 2004) have used the airline panel data in Greene (1997), but, again, none of these found useful prior information in any of the predecessors. I note (only) three studies: Tsionas (2001b), in which prior (DEA efficiency) estimates are incorporated in the estimation priors, and Kim and Schmidt (2000) and O'Donnell

and Coelli (2005), which use the classical MLEs for the variance parameters, for the study at hand.

43. Of course, once they are added to commercial software, the issue of difficulty of implementation becomes a moot point.

44. See van den Broeck et al. (1994), Osiewalski and Steel (1998), Koop and Steel (2001), and Tsionas (2001a) for surveys and introductory analyses.

45. This is to be expected given the well-known result that, in the presence of diffuse priors, the Bayesian posterior mean will converge to the mode of the likelihood function—in the absence of prior information, the likelihood function dominates the Bayesian estimator. See Kim and Schmidt (2000) for commentary.

46. The results are extended in Lee (1993), who addresses the issue of inference in the stochastic frontier model at the boundary of the parameter space, $\lambda = 0$.

47. There is a complication in the Bayesian approach to estimation of this model that does not arise in the classical method—the "labeling problem." A Bayesian estimator must actually decide a priori which class is which and make a distinction during estimation. An approach that is sometimes used is to identify as "class 1" the class with the largest prior class probability, and the others similarly. The classical MLE just allows the classes to fall out of the maximization process.

48. The authors analyzed the truncated-normal model and considered a cost frontier model. I have simplified the formulation a bit for purposes of the description, and changed the parameterization slightly, as well.

49. There is a minor ambiguity in Kumbhakar et al. (2005). The authors define $m$ to be the number of parameters in the model, but the definition of the kernel function is only consistent with $m$ equal the number of variables in $\mathbf{x_i}$.

50. Koop (2001) also applied this approach to the output of major league baseball players were the four outputs are singles, doubles and triples, home runs, and walks and the "inputs" are time, team, and league dummy variables—illustrative of the technique, but perhaps of limited policy relevance.

51. See Christensen and Greene (1976).

52. Some additional useful related results appear in Kumbhakar (1991b) and in Kumbhakar and Lovell (2000).

53. Note what would be the utility of the Førsund and Jansen's (1977) input-oriented efficiency formulation, $y_i = F[f(\mathrm{TE}_i\mathbf{x_i})]$. Under this assumption, the cost function would always be of the form $\ln C_i = \ln F^{-1}(y) + \ln c(w_i) - v_i - \ln \mathrm{TE}_i$. See Alvarez, Arias, and Greene (2005) for an analysis along these lines that explicitly accounts for the internal relationships and Alvarez et al. (2004) for an analysis of mixtures of input and output-oriented inefficiency models.

54. The lack of invariance to the units of measurement of output also conflicts with the useful Farrell measure of economic efficiency.

55. As I noted in the preface (section 2.1.5), my apologies to the very large number of researchers whose work is not listed here. For these two industries, there are scores of applications, and in the interest of brevity, I can list only a few of them. Those chosen are only illustrative of these very large literatures.

56. In a remarkably obscure study, Greene (1983) proposed a translog cost model in which all parameters were of the form $\theta_{k,t} = \theta_{k,0} + \theta_{k,1}t$. The Kurkalova and Carriquiry (2003) study does likewise, although with only two periods, their formulation is much simpler. Cornwell et al. (1990) added a quadratic term, $\theta_{k,2}\mathrm{t}^2$.

57. This approach has become essentially standard in the DEA literature.

58. It does not follow automatically that biases in the estimates of the parameters of the production or cost function will translate into biases in estimates of inefficiency (though the evidence does suggest it does). In a linear regression model, omitted variable biases in coefficients do not always translate into biases in forecasts. Under some reasonable assumptions, one can, e.g., safely truncate a distributed lag model.

59. These models are anticipated in Kumbhakar and Hjalmarsson (1995), who proposed representing time-varying inefficiency with equivalents to these models. Their proposed estimators do not maintain the stochastic frontier specification, however. Methodological issues are discussed in Heshmati and Kumbhakar (1994) and Kumbhakar and Heshmati (1995).

60. Swamy and Tavlas (2001) label these "first-generation" and "second-generation" methods. In the current literature, one can find a vast library of treatments on "mixed" models, "hierarchical" models, "multilevel" models, and "random-parameters" models, all of which are the same. Applications can be found in every area in the social sciences.

61. Their study demonstrated, as one might guess from the functional form, that ignoring heteroskedasticity of this form would lead to persistent biases in MLEs of the production or cost function parameters.

62. A Monte Carlo study of the properties of this model is Guermat and Hadri (1999). The exponential model with nonmonotonic effects on estimated inefficiency is examined at length in Wang (2002). An application of the doubly heteroskedastic model is Hadri et al. (2003a, 2003b).

63. See Schmidt and Sickles (1984).

64. There seems to be a presumption in some writings that the fixed-effects model when fit to panel data *must* be computed by using the "within" transformation (deviations from group means). In fact, this is merely a means to another end and, with modern software, some of which is quite capable of estimating regressions with hundreds of coefficients (even with desktop computers), may be quite unnecessary. The point is that this ought not to be construed as any sort of model in itself; it is merely a computational device usable for solving a practical problem. Note this is the motivation behind Greene's (2005) "true" fixed-effects model.

65. For example, their model involved 14 basic coefficients and a $[\alpha, \gamma, \delta]_i$ for each of eight firms, a total of 38 coefficients. This is well within the reach of any modern regression package, even on a desktop computer. The point is that there are few practical obstacles to computing estimators for the various frontier models given the current state of computer software.

66. Kumbhakar (1991a) proposes a hybrid of the frontier model and a two-way random-effects model. The disturbance specification is $\varepsilon_{it} = w_i + c_t + v_{it} - u_{it}$ (my notation) in which $w_i$, $c_t$, and $v_{it}$ constitute, with normal distributions, a more or less conventional model by Balestra and Nerlove (1968), but $u_{it}$ is the truncation of a normally distributed variable with mean $\mu_{it}$ (which may depend on exogenous variables). Thus, the fact that $u_{it}$ is positive embodies the frontier aspect of the model, but the panel nature of the model is carried by $w_i$ and $c_t$. Once again, this is essentially the same as Greene's (2004a) true random-effects model.

67. In the Bayesian framework, the distinction between fixed and random effects does not have the same interpretation as in the classical framework. As will be evident momentarily, the distinction becomes simply whether the inefficiencies are treated as parameters or as missing data. Estimation is the same either way.

68. Greene and Misra (2003) discuss simulation-based estimation of this quantity for cases in which the precise functional form is not known.

69. For the gamma model, $E[u_{it}^r|\varepsilon_{it}] = q(P + r - 1, \varepsilon_{it})/q(P - 1, \varepsilon_{it})$.

70. I have also investigated experimentally the relationship between the JLMS estimator and the actual inefficiencies when the latter are "known." Not surprisingly, I found that the overall quality of the estimator degrades with smaller $\lambda$ values, that is, with larger noise components, $v$. What was a bit surprising was that the JLMS estimator tends systematically to underpredict $u$ when $u$ is small and overpredict it when $u$ is large—again, with improvement as $\lambda$ increases.

71. There is one other small point that might be considered in either a Bayesian or a classical context. The interval thus constructed is not as short as it might be. In order to encompass $100(1 - \alpha)\%$ of this asymmetric distribution, with the narrowest interval possible, we should find the equivalent of the Bayesian HPD interval. For the truncated normal distribution, we can actually deduce what this is. Suppose we wish to capture 95% of the mass of the distribution. For a density in which more than 2.5% of the untruncated distribution is to the left of zero (and this will be most of the time), the shortest interval will be from zero to the 95th percentile of the truncated normal distribution. By construction, this interval will be shorter than one that puts 2.5% of the mass from zero to L and 2.5% of the mass to the right of $U$. A simple figure makes this obvious.

72. See Bera and Sharma (1999) and Hjalmarsson, Kumbhakar, and Heshmati (1996). Bera and Sharma also provide the expressions needed for the Battese and Coelli measures, $E[\text{TE}_i|\varepsilon_i]$.

73. There have been a number of panel-data analyses of this industry, including Sickles (1987), Sickles, Good, and Johnson (1986), Schmidt and Sickles (1984), Good, Nadiri, Roller, and Sickles (1993), Good, Roller, and Sickles (1993), Good, Roller, and Sickles (1995), Good and Sickles (1995), and Alam and Sickles (1998, 2000), and the references cited therein.

74. Results extracted from other studies, notably the Bayesian estimates reported in section 2.10.2.1, were not replicated here. Researchers who wish to replicate those results should contact the original authors.

75. There are no general-purpose econometric packages that specifically contain MLEs for deterministic frontier models, though there are any number of programs with which the linear and quadratic programming "estimators" can be computed. Likewise, the gamma model with only one-sided residuals can be programmed but presents an exceedingly difficult problem for conventional estimation.

76. A separate package, downloadable at no cost (as is *Frontier 4.2*), distributed by the Center for Efficiency and Productivity Analysis at the University of Queensland in Australia, can be used for DEA (http://http://www.scripting.com/frontier/newReleases/Frontier42.html).

77. The lack of replicability in empirical econometrics has been widely documented and is, ultimately, a major challenge and shortcoming of a large amount of contemporary research (see, e.g., Anderson et al., 2005). Even a cursory reading of the Bayesian applications of stochastic frontier modeling will suggest how difficult it would be to replicate this type of work the way it is currently documented.

78. The original data set contained the 123 observations discussed above and 35 holding company aggregates of some of the firms. My data file contains all 158 observations, but we will be using only the 123 firm-level observations.

79. Using the standard deviation of $u_i$ rather than the mean as the initial estimator of $\gamma$ to form the weighting matrix led to an implausible value of $\theta$ in excess of 6.0.

80. The parameter heterogeneity feature of these two applications takes the model in a different direction. It appears that this feature has induced the peculiar findings with respect to inefficiency, but that is aside from the specification issue.

81. Their conclusion, "This paradigm thus allows direct posterior inference on firm-specific efficiencies, avoiding the much criticized two-step procedure of Jondrow et al. (1982)," overreaches a bit. Subsequent researchers continue to rely comfortably on JLMS and the extensions suggested by Horrace and Schmidt (1996), Bera and Sharma (1999), and Kim and Schmidt (2000). And, as shown above, the Bayesian posterior estimators are essentially the same as the classical JLMS estimator.

82. Kim and Schmidt (2000) found, likewise, that the Bayesian and classical estimators when applied to the same model under the same assumptions tended to produce essentially the same results.

83. It appears that there might have been some small differences in the data used—the "fixed-parameter" estimates reported seem to resemble, but not quite equal, our classical, least squares estimates or our normal–gamma estimates. Because we are interested here only in the methodology, I leave it to subsequent researchers to fine-tune the comparison.

84. If we restrict the sample to only the firms with all 15 years of data, the entire problem vanishes, and there is no problem fitting the stochastic production frontier model. As a general rule, we would not do the specification search in this fashion, so we will not pursue this angle.

85. These data were developed and provided by S. Kumbhakar and E. Tsionas. Their assistance is gratefully acknowledged here.

86. This model as given would be the classical counterpart to a hierarchical Bayes formulation, e.g., in Tsionas (2002). The fixed values of $\lambda$ and $\sigma$ would correspond to flat priors, which would lead to the MLEs.

87. In our implementation of this procedure, we do not actually use random draws from the distribution. The terms in the simulation are built up from Halton sequences, which are deterministic. Details on this method of integration appear in Train (2003) and Greene (2003a).

88. The wall of points at the right in the scatter plot is the value produced for relatively extreme observations, where the numerical limit of our ability to compute the standard normal CDF is reached: at about 8.9 standard deviations.

## References

Absoft, 2005, *IMSL Libraries, Reference Guide*, http://www.absoft.com/Products/Libraries/imsl.html.

Adams, R., A. Berger, and R. Sickles, 1999, "Semiparametric Approaches to Stochastic Panel Frontiers with Applications in the Banking Industry," *Journal of Business and Economic Statistics*, 17, pp. 349–358.

Afriat, S., 1972, "Efficiency Estimation of Production Functions," *International Economic Review*, 13, pp. 568–598.

Aguilar, R., 1988, "Efficiency in Production: Theory and an Application on Kenyan Smallholders," Ph.D. Dissertation, Department of Economics, University of Göteborg, Sweden.

Aigner, D., T. Amemiya, and D. Poirier, 1976, "On the Estimation of Production Frontiers," *International Economic Review*, 17, pp. 377–396.

Aigner, D., and S. Chu, 1968, "On Estimating the Industry Production Function," *American Economic Review*, 58, pp. 826–839.

Aigner, D., K. Lovell, and P. Schmidt, 1977, "Formulation and Estimation of Stochastic Frontier Production Function Models," *Journal of Econometrics*, 6, pp. 21–37.

Akhavein, J., P. Swamy, and S. Taubman, 1994, "A General Method of Deriving Efficiencies of Banks from a Profit Function," Working Paper No. 94-26, Wharton School, University of Pennsylvania, Philadelphia.

Alam, I., and R. Sickles, 1998, "The Relationship Between Stock Market Returns and Technical Efficiency Innovations: Evidence from the U.S. Airline Industry," *Journal of Productivity Analysis*, 9, pp. 35–51.

Alam, I., and R. Sickles, 2000, "A Time Series Analysis of Deregulatory Dynamics and Technical Efficiency: The Case of the U.S. Airline Industry," *International Economic Review*, 41, pp. 203–218.

Albert, J., and S. Chib, 1993, "Bayesian Analysis of Binary and Polytomous Response Data," *Journal of the American Statistical Association*, 88, pp. 669–679.

Albriktsen, R., and F. Førsund, 1990, "A Productivity Study of the Norwegian Building Industry," *Journal of Productivity Analysis*, 2, pp. 53–66.

Ali, A., and L. Seiford, 1993, "The Mathematical Programming Approach to Efficiency Analysis," in *The Measurement of Productive Efficiency*, H. Fried, K. Lovell, and S. Schmidt, eds. Oxford University Press, New York.

Alvarez, A., C. Amsler, L. Orea, and P. Schmidt, 2006, "Interpreting and Testing the Scaling Property in Models Where Inefficiency Depends on Firm Characteristics," *Journal of Productivity Analysis*, 25, 3, 2006, pp. 201-212.

Alvarez, A., C. Arias, and W. Greene, 2005, "Accounting for Unobservables in Production Models: Management and Inefficiency," Working Paper, Department of Economics, University of Oviedo, Spain.

Alvarez, A., C. Arias, and S. Kumbhakar, 2004, "Additive Versus Interactive Unobservable Individual Effects in the Estimation of Technical Efficiency," Working Paper, Department of Economics, University of Oviedo, Spain.

Amemiya, T., 1973, "Regression Analysis When the Dependent Variable Is Truncated Normal," *Econometrica*, 41, pp. 997–1016.

Anderson, R., W. Greene, B. McCullough, and H. Vinod, 2005, "The Role of Data and Program Archives in the Future of Economic Research," Working Paper 2005-14, Federal Reserve Bank, St. Louis, MO.

Annaert, J., J. van den Broeck, and R. Vennet, 2001, "Determinants of Mutual Fund Performance: A Bayesian Stochastic Frontier Approach," Working Paper 2001/103, University of Gent, Ghent, Belgium.

Aptech Systems, Inc., 2005, *Gauss Reference Guide*, http://http://www.aptech.com, Kent, WA.

Arickx, F., J. Broeckhove, M. Dejonghe, and J. van den Broeck, 1997, "BSFM: A Computer Program for Bayesian Stochastic Frontier Models," *Computational Statistics*, 12, pp. 403–421.

Arrow, K., H. Chenery, B. Minhas, and R. Solow, 1961, "Capital Labor Substitution and Economic Efficiency," *Review of Economics and Statistics*, 45, pp. 225–247.

Atkinson, S., and C. Cornwell, 1993, "Estimation of Technical Efficiency with Panel Data: A Dual Approach," *Journal of Econometrics*, 59, pp. 257–262.

Atkinson, S., and J. Dorfman, 2005, "Bayesian Measurement of Productivity and Efficiency in the Presence of Undesirable Outputs: Crediting Electric Utilities for Reducing Air Pollution," *Journal of Econometrics*, 126, pp. 445–468.

Atkinson, S, R. Fare, and D. Primont, 2003 "Stochastic Estimation of Firm Technology, Inefficiency and Productivity Growth Using Shadow Cost and Distance Functions," *Journal of Econometrics*, 108, pp. 203–226.

Averch, H., and L. Johnson, 1962, "Behavior of the Firm under Regulatory Constraint," *American Economic Review*, 52, pp. 1052–1069.

Balestra, P., and M. Nerlove, 1968, "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas," *Econometrica*, 34, pp. 585–612.

Bankar, R., 1993, "Maximum Likelihood, Consistency, and Data Envelopment Analysis," *Management Science*, 39, pp. 1265–1273.

Bankar, R., 1997, "Hypothesis Tests Using Data Envelopment Analysis," *Journal of Productivity Analysis*, 7, pp. 139–159.

Banker, R., and A. Maindiratta, 1988, "Nonparametric Analysis of Technical and Allocative Efficiencies in Production," *Econometrica*, 56, pp. 1315–1332.

Battese, G., 1992, "Frontier Production Functions and Technical Efficiency: A Survey of Empirical Applications in Agricultural Economics," *Agricultural Economics*, 7, pp. 185–208.

Battese, G., and T. Coelli, 1988, "Prediction of Firm-level Technical Efficiencies with a Generalized Frontier Production Function and Panel Data," *Journal of Econometrics*, 38, pp. 387–399.

Battese, G., and T. Coelli, 1992, "Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India," *Journal of Productivity Analysis*, 3, pp. 153–169.

Battese, G., and T. Coelli, 1995, "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Model for Panel Data," *Empirical Economics*, 20, pp. 325–332.

Battese, G., T. Coelli, and T. Colby, 1989, "Estimation of Frontier Production Functions and the Efficiencies of Indian Farms Using Panel Data from ICRISTAT's Village Level Studies," *Journal of Quantitative Economics*, 5, pp. 327–348.

Battese, G., and G. Corra, 1977, "Estimation of a Production Frontier Model: With Application to the Pastoral Zone of Eastern Australia," *Australian Journal of Agricultural Economics*, 21, pp. 167–179.

Battese, G., A. Rambaldi, and G. Wan, 1997, "A Stochastic Frontier Production Function with Flexible Risk Properties," *Journal of Productivity Analysis*, 8, pp. 269–280.

Bauer, P., 1990, "A Survey of Recent Econometric Developments in Frontier Estimation," *Journal of Econometrics*, 46, pp. 21–39.

Bauer, P., A. Berger, G. Ferrier, and D. Humphrey, 1998, "Consistency Conditions for Regulatory Analysis of Financial Institutions: A Comparison of Frontier Efficiency Methods," *Journal of Economics and Business*, 50, pp. 85–114.

Beckers, D., and C. Hammond, 1987, "A Tractable Likelihood Function for the Normal-Gamma Stochastic Frontier Model," *Economics Letters*, 24, pp. 33–38.

Bera, A., and S. Sharma, 1999, "Estimating Production Uncertainty in Stochastic Frontier Production Function Models," *Journal of Productivity Analysis*, 12, pp. 187–210.

Berger, A., 1993, "Distribution Free Estimates of Efficiency in the U.S. Banking Industry and Tests of the Standard Distributional Assumptions," *Journal of Productivity Analysis*, 4, pp. 261–292.

Berger, A., and D. Humphrey, 1991, "The Dominance of Inefficiencies over Scale and Product Mix Economies in Banking," *Journal of Monetary Economics*, 28, pp. 117–148.

Berger, A., and D. Humphrey, 1992, "Measurement and Efficiency Issues in Commercial Banking," in *National Bureau of Economic Research Studies in Income and Wealth*, Vol. 56, *Output Measurement in the Service Sector*, Z. Griliches, ed., Chicago, University of Chicago Press.

Berger, A., and L. Mester, 1997, "Inside the Black Box: What Explains Differences in Efficiencies of Financial Institutions?" *Journal of Banking and Finance*, 21, pp. 895–947.

Bhattacharyya, A., S. Kumbhakar, and A. Bhattacharyya, 1995, "Ownership Structure and Cost Efficiency: A Study of Publicly Owned Passenger Bus Transportation Companies in India," *Journal of Productivity Analysis*, 6, pp. 47–62.

Bjurek, H., L. Hjalmarsson, and F. Forsund, 1990, "Deterministic Parametric and Nonparametric Estimation of Efficiency in Service Production: A Comparison," *Journal of Econometrics*, 46, pp. 213–227.

Breusch, T., and A. Pagan, 1980, "The LM Test and Its Applications to Model Specification in Econometrics," *Review of Economic Studies*, 47, pp. 239–254.

Cameron, C., T. Li, P. Trivedi, and D. Zimmer, 2004, "Modeling the Differences in Counted Outcomes Using Bivariate Copula Models with Applications to Mismeasured Counts," *Econometrics Journal*, 7, pp. 566–584.

Casella, G., and E. George, 1992, "Explaining the Gibbs Sampler," *American Statistician*, 46, pp. 167–174.

Caudill, S., and Ford, J., 1993, "Biases in Frontier Estimation Due to Heteroskedasticity," *Economics Letters*, 41, pp. 17–20.

Caudill, S., J. Ford, and D. Gropper, 1995, "Frontier Estimation and Firm Specific Inefficiency Measures in the Presence of Heteroscedasticity," *Journal of Business and Economic Statistics*, 13, pp. 105–111.

Caves, D., L. Christensen, and M. Trethaway, 1980, "Flexible Cost Functions for Multiproduct Firms," *Review of Economics and Statistics*, 62, pp. 477–481.

Cazals, C., J. Florens, and L. Simar, 2002, "Nonparametric Frontier Estimation: A Robust Approach," *Journal of Econometrics*, 106, pp. 1–25.

Chakraborty, K., B. Biswas, and W. Lewis, 2001, "Measurement of Technical Efficiency in Public Education: A Stochastic and Nonstochastic Production Function Approach," *Southern Economic Journal*, 67, pp. 889–905.

Chen, T., and D. Tang, 1989, "Comparing Technical Efficiency Between Import Substitution Oriented and Export Oriented Foreign Firms in a Developing Economy," *Journal of Development of Economics*, 26, pp. 277–289.

Christensen, L., and W. Greene, 1976, "Economies of Scale in U.S. Electric Power Generation," *Journal of Political Economy*, 84, pp. 655–676.

Cobb, S., and P. Douglas, 1928, "A Theory of Production," *American Economic Review*, 18, pp. 139–165.

Coelli, T., 1991, "Maximum Likelihood Estimation of Stochastic Frontier Production Functions with Time Varying Technical Efficiency Using the Computer Program Frontier Version 2.0," Department of Economics, University of New England, Armidale, Australia.

Coelli, T., 1995, "Estimators and Hypothesis Tests for a Stochastic Frontier Function: A Monte Carlo Analysis," *Journal of Productivity Analysis*, 6, pp. 247–268.

Coelli, T., 1996, "Frontier Version 4.1: A Computer Program for Stochastic Frontier Production and Cost Function Estimation," Working Paper 96/7, Center for Efficiency and Productivity Analysis, Department of Econometrics, University of New England, Armidale, Australia.

Coelli, T., 2000, "On the Econometric Estimation of the Distance Function Representation of a Production Technology," Working Paper 2000/42, CORE, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

Coelli, T., and S. Perelman, 1996, "Efficiency Measurement, Multiple Output Technologies and Distance Functions: With Application to European Railways," *European Journal of Operational Research*, 117, pp. 326–339.

Coelli, T., and S. Perelman, 1999, "A Comparison of Parametric and Nonparametric Distance Functions: With Application to European Railways," CREPP Discussion Paper 96/25, University of Liège, Belgium.

Coelli, T., and S. Perelman, 2000, "Technical Efficiency of European Railways: A Distance Function Approach," *Applied Economics*, 32, pp. 67–76.

Coelli, T., P. Rao, and G. Battese, 1998, *An Introduction to Efficiency and Productivity Analysis*, Kluwer Academic Publishers, Boston.

Cornwell, C., and P. Schmidt, 1996, "Production Frontiers and Efficiency Measurement," in *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, 2nd rev. ed., L. Matyas and P. Sevestre, eds., Kluwer Academic Publishers, Boston.

Cornwell, C., P. Schmidt, and R. Sickles, 1990, "Production Frontiers with Cross Sectional and Time Series Variation in Efficiency Levels," *Journal of Econometrics*, 46, pp. 185–200.

Cuesta, R., 2000, "A Production Model with Firm Specific Temporal Variation in Technical Efficiency: With Application to Spanish Dairy Farms," *Journal of Productivity Analysis*, 13, pp. 139–158.

Cuesta, R., and L. Orea, 2002, "Mergers and Technical Efficiency in Spanish Savings Banks: A Stochastic Distance Function Approach," *Journal of Banking and Finance*, 26, pp. 2231–2247.

Cummins, J., and H. Zi, 1998, "Comparison of Frontier Efficiency Models: An Application to the U.S. Life Insurance Industry," *Journal of Productivity Analysis*, 10, pp. 131–152.

Dean, J., 1951, *Managerial Economics*, Prentice Hall, Englewood Cliffs, NJ.

Debreu, G., 1951, "The Coefficient of Resource Utilization," *Econometrica*, 19, pp. 273–292.

Deprins, D., and L. Simar, 1985, "A Note on the Asymptotic Relative Efficiency of M.L.E. in a Linear Model with Gamma Disturbances," *Journal of Econometrics*, 27, pp. 383–386.

Deprins, D., and L. Simar, 1989a, "Estimation of Deterministic Frontiers with Exogenous Factors of Inefficiency," *Annals of Economics and Statistics (Paris)*, 14, pp. 117–150.

Deprins, D., and L. Simar, 1989b, "Estimating Technical Inefficiencies with Corrections for Environmental Conditions with an Application to Railway Companies," *Annals of Public and Cooperative Economics*, 60, pp. 81–102.

Econometric Software, Inc., 2000, *LIMDEP, User's Manual*, http://http://www.limdep.com, Plainview, NY.

Estima, 2005, *RATS Reference Guide*, http://http://www.estima.com, Evanston, IL.

Evans, D, A. Tandon, C. Murray, and J. Lauer, 2000a, "The Comparative Efficiency of National Health Systems in Producing Health: An Analysis of 191 Countries," GPE Discussion Paper No. 29, EIP/GPE/EQC, World Health Organization, Geneva.

Evans D, A. Tandon, C. Murray, and J. Lauer, 2000b, "Measuring Overall Health System Performance for 191 Countries," GPE Discussion Paper No. 30, EIP/GPE/EQC, World Health Organization, Geneva.

Fan, Y., Q. Li, and A. Weersink, 1996, "Semiparametric Estimation of Stochastic Production Frontiers," *Journal of Business and Economic Statistics*, 64, pp. 865–890.

Farrell, M., 1957, "The Measurement of Productive Efficiency," *Journal of the Royal Statistical Society A, General*, 120, pp. 253–281.

Farsi, M., and M. Filippini, 2003, "An Empirical Analysis of Cost Efficiency in Non-profit and Public Nursing Homes," Working Paper, Department of Economics, University of Lugano, Switzerland.

Farsi, M., M. Filippini, and M. Kuenzle, 2003, "Unobserved Heterogeneity in Stochastic Cost Frontier Models: A Comparative Analysis," Working Paper 03-11, Department of Economics, University of Lugano, Switzerland.

Fernandez, C., G. Koop, and M. Steel, 1999, "Multiple Output Production with Undesirable Outputs: An Application to Nitrogen Surplus in Agriculture," Working Paper 99-17, Department of Economics, University of Edinburgh, Scotland.

Fernandez, C., G. Koop, and M. Steel, 2000, "A Bayesian Analysis of Multiple Output Production Frontiers," *Journal of Econometrics*, 98, pp. 47–79.

Fernandez, C., G. Koop, and M. Steel, 2002, "Multiple Output Production with Undesirable Outputs: An Application to Nitrogen Surplus in Agriculture," *Journal of the American Statistical Association*, 97, pp. 432–442.

Fernandez, C., G. Koop, and M. Steel, 2005, "Alternative Efficiency Measures for Multiple Output Production," Journal of Econometrics, 126, 2, 2005, pp. 411–444.

Fernandez, C. J. Osiewalski, and M. Steel, 1997, "On the Use of Panel Data in Stochastic Frontier Models with Improper Priors," *Journal of Econometrics*, 79, pp. 169–193.

Ferrier, G., and K. Lovell, 1990, "Measuring Cost Efficiency in Banking: Econometric and Linear Programming Evidence," *Journal of Econometrics*, 46, pp. 229–245.

Førsund, F., 1992, "A Comparison of Parametric and Nonparametric Efficiency Measures: The Case of Norwegian Ferries," *Journal of Productivity Analysis*, 3, pp. 25–44.

Førsund, F., and L. Hjalmarsson, 1974, "On the Measurement of Productive Efficiency," *Swedish Journal of Economics*, 76, pp. 141–154.

Førsund, F., and L. Hjalmarsson, 1979, "Frontier Production Functions and Technical Progress: A Study of General Milk Processing in Swedish Dairy Plants," *Econometrica*, 47, pp. 883–900.

Førsund, F., and E. Jansen, 1977, "On Estimating Average and Best Practice Homothetic Production Functions via Cost Functions," *International Economic Review*, 18, pp. 463–476.

Førsund, F., K. Lovell, and P. Schmidt, 1980, "A Survey of Frontier Production Functions and of Their Relationship to Efficiency Measurement," *Journal of Econometrics*, 13, pp. 5–25.

Gabrielsen, A., 1975, "On Estimating Efficient Production Functions," Working Paper No. A-85, Chr. Michelsen Institute, Department of Humanities and Social Sciences, Bergen, Norway.

Gong, B., and R. Sickles, 1989, "Finite Sample Evidence on the Performance of Stochastic Frontier Models Using Panel Data," *Journal of Productivity Analysis*, 1, pp. 119–261.

Good, D., I. Nadiri, L. Roller, and R. Sickles, 1993, "Efficiency and Productivity Growth Comparisons of European and U.S. Air Carriers: A First Look at the Data," *Journal of Productivity Analysis*, 4, pp. 115–125.

Good, D., L. Roller, and R. Sickles, 1993, "U.S. Airline Deregulation: Implications for European Transport," *Economic Journal*, 103, pp. 1028–1041.

Good, D., L. Roller, and R. Sickles, 1995, "Airline Efficiency Differences Between Europe and the U.S.: Implications for the Pace of E.C. Integration and Domestic Regulation," *European Journal of Operational Research*, 80, pp. 508–518.

Good, D., and Sickles, R., 1995, "East Meets West: A Comparison of the Productive Performance of Eastern European and Western European Air Carriers," Working Paper, Department of Economics, Rice University, Houston, TX.

Gravelle H, R. Jacobs, A. Jones, and A. Street, 2002a, "Comparing the Efficiency of National Health Systems: Econometric Analysis Should Be Handled with Care," Working Paper, Health Economics Unit, University of York, UK.

Gravelle H, R. Jacobs, A. Jones, and A. Street, 2002b, "Comparing the Efficiency of National Health Systems: A Sensitivity Approach," Working Paper, Health Economics Unit, University of York, UK.

Greene, W., 1980a, "Maximum Likelihood Estimation of Econometric Frontier Functions," *Journal of Econometrics*, 13, pp. 27–56.

Greene, W., 1980b, "On the Estimation of a Flexible Frontier Production Model," *Journal of Econometrics*, 3, pp. 101–115.

Greene, W., 1983, "Simultaneous Estimation of Factor Substitution, Economies of Scale, and Non-neutral Technological Change," in *Econometric Analyses of Productive Efficiency*, Dogramaci, A., ed., Nijoff Publishing Co., Dordrecht, The Netherlands.

Greene, W., 1990, "A Gamma Distributed Stochastic Frontier Model," *Journal of Econometrics*, 46, pp. 141–163.

Greene, W., 1993, "The Econometric Approach to Efficiency Analysis," in *The Measurement of Productive Efficiency*, H. Fried, K. Lovell, and S. Schmidt, eds., Oxford University Press, Oxford.

Greene, W., 1997, "Frontier Production Functions," in *Handbook of Applied Econometrics*, Vol. 2, *Microeconomics*, H. Pesaran and P. Schmidt, eds., Oxford University Press, Oxford.

Greene, W., 2000, "LIMDEP Computer Program: Version 8.0," Econometric Software, Plainview, NY.

Greene, W., 2003a, *Econometric Analysis*, 5th ed., Prentice Hall, Upper Saddle River, NJ.

Greene, W., 2003b, "Simulated Likelihood Estimation of the Normal-Gamma Stochastic Frontier Function," *Journal of Productivity Analysis*, 19, pp. 179–190.

Greene, W., 2004a, "Fixed and Random Effects in Stochastic Frontier Models," *Journal of Productivity Analysis*, 23, pp. 7–32.

Greene, W., 2004b, "Distinguishing Between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems," *Health Economics*, 13, pp. 959–980.

Greene, W., 2005, "Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model," *Journal of Econometrics*, 126, pp. 269–303.

Greene, W., and S. Misra, 2003, "Simulated Maximum Likelihood Estimation of General Stochastic Frontier Regressions," Working Paper, William Simon School of Business, University of Rochester, NY.

Griffin, J., and Steel, M., 2004, "Semiparametric Bayesian Inference for Stochastic Frontier Models," *Journal of Econometrics*, 123, pp. 121–152.

Griffiths, W.E., C. O'Donnell, A. Tan, and R. Cruz, 2000, "Imposing Regularity Conditions on a System of Cost and Cost Share Equations: A Bayesian Approach," *Australian Journal of Agricultural Economics*, 44, pp. 107–127.

Guermat, C., and K. Hadri, 1999, "Heteroscedasticity in Stochastic Frontier Models: A Monte Carlo Analysis," Working Paper, Department of Economics, City University, London, http://http://www.ex.ac.uk/∼cguermat/het_mar99.pdf.

Hadri, K., 1999, "Estimation of a Doubly Heteroscedastic Stochastic Frontier Cost Function," *Journal of Business and Economics and Statistics*, 17, pp. 359–363.

Hadri, K., C. Guermat, and J. Whittaker, 2003a, "Estimating Farm Efficiency in the Presence of Double Heteroscedasticity Using Panel Data," *Journal of Applied Economics*, 6, pp. 255–268.

Hadri, K., C. Guermat, and J. Whittaker, 2003b, "Estimation of Technical Inefficiency Effects Using Panel Data and Doubly Heteroscedastic Stochastic Production Frontiers," *Empirical Economics*, 28, pp. 203–222.

Hausman, J., and W. Taylor, 1981, "Panel Data and Unobservable Individual Effects," *Econometrica*, 49, pp. 1377–1398.

Heshmati, A., and S. Kumbhakar, 1994, "Farm Heterogeneity and Technical Efficiency: Some Results from Swedish Dairy Farms," *Journal of Productivity Analysis*, 5, pp. 45–61.

Hicks, J., 1935, The Theory of Monopoly: A Survey," *Econometrica*, 3, pp. 1–20.

Hildebrand, G., and T. Liu, 1965, *Manufacturing Production Functions in the United States*, Cornell University Press, Ithaca, NY.

Hildreth, C., and C. Houck, 1968, "Some Estimators for a Linear Model with Random Coefficients," *Journal of the American Statistical Association*, 63, pp. 584–595.

Hjalmarsson, L., S. Kumbhakar, and A. Heshmati, 1996, "DEA, DFA and SFA: A Comparison," *Journal of Productivity Analysis*, 7, pp. 303–327.

Hollingsworth, J., and B. Wildman, 2002, "The Efficiency of Health Production: Re-estimating the WHO Panel Data Using Parametric and Nonparametric Approaches to Provide Additional Information," *Health Economics*, 11, pp. 1–11.

Holloway, G., D. Tomberlin, and X. Irz, 2005, "Hierarchical Analysis of Production Efficiency in a Coastal Trawl Fishery," in *Simulation Methods in Environmental and Resource Economics*, R. Scarpa and A. Alberini, eds., Springer Publishers, New York, 2005.

Horrace, W., and S. Richards, 2005, "Bootstrapping Efficiency Probabilities in Parametric Stochastic Frontier Models," Working Paper 2005-004, Maxwell School, Syracuse University, Syracuse, NY.

Horrace, W., and P. Schmidt, 1996, "Confidence Statements for Efficiency Estimates from Stochastic Frontier Models," *Journal of Productivity Analysis*, 7, pp. 257–282.

Horrace, W., and P. Schmidt, 2000, "Multiple Comparisons with the Best, with Economic Applications," *Journal of Applied Econometrics*, 15, pp. 1–26.

Hotelling, H., 1932, "Edgeworth's Taxation Paradox and the Nature of Supply and Demand Functions," *Journal of Political Economy*, 40, pp. 577–616.

Huang, C., and T. Fu, 1999, "An Average Derivative Estimator of a Stochastic Frontier," *Journal of Productivity Analysis*, 12, pp. 49–54.

Huang, C., and J. Liu, 1994, "Estimation of a Non-neutral Stochastic Frontier Production Function," *Journal of Productivity Analysis*, 5, pp. 171–180.

Huang, R., 2004, "Estimation of Technical Inefficiencies with Heterogeneous Technologies," *Journal of Productivity Analysis*, 21, pp. 277–296.

Huang, T., and M. Wang, 2004, "Comparisons of Economic Inefficiency Between Output and Input Measures of Technical Inefficiency Using the Fourier Flexible Cost Function," *Journal of Productivity Analysis*, 22, pp. 123–142.

Humphrey, D., and L. Pulley, 1997, 'Banks' Responses to Deregulation: Profits, Technology and Efficiency," *Journal of Money, Credit and Banking*, 29, pp. 73–93.

Hunt, J., Y. Kim, and R. Warren, 1986, "The Effect of Unemployment Duration on Re-employment Earnings: A Gamma Frontier Approach," Working Paper, Department of Economics, University of Georgia, Athens, GA.

Hunt-McCool, J., and R. Warren, 1993, "Earnings Frontiers and Labor Market Efficiency," in *The Measurement of Productive Efficiency*, H. Fried, K. Lovell, and S. Schmidt, eds., Oxford University Press, New York.

Johnston, J. 1959, *Statistical Cost Analysis*, McGraw-Hill, New York.

Jondrow, J., K. Lovell, I. Materov, and P. Schmidt, 1982, "On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model," *Journal of Econometrics*, 19, pp. 233–238.

Kalirajan, K., and M. Obwona, "Frontier Production Function: The Stochastic Coefficients Approach," *Oxford Bulletin of Economics and Statistics*, 56, 1, 1994, pp. 87–96.

Kalirajan, K., and R. Shand, 1999, "Frontier Production Functions and Technical Efficiency Measures," *Journal of Economic Surveys*, 13, pp. 149–172.

Kim, J., 2002, "Limited Information Likelihood and Bayesian Analysis," *Journal of Econometrics*, 107, pp. 175–193.

Kim, Y., and P. Schmidt, 2000, "A Review and Empirical Comparison of Bayesian and Classical Approaches to Inference on Efficiency Levels in Stochastic Frontier Models with Panel Data," *Journal of Productivity Analysis*, 14, pp. 91–98.

Kleit, A., and D. Terrell, 2001, "Measuring Potential Efficiency Gains from Deregulation of Electricity Generation: A Bayesian Approach," *Review of Economics and Statistics*, 83, pp. 523–530.

Klotz, B., R. Madoo, and R. Hansen, 1980, "A Study of High and Low Labor Productivity Establishments in U.S. Manufacturing," in *Studies in Income and Wealth*, Vol. 44, *New Developments in Productivity Measurement and Analysis*, *Research Conference on Research in Income and Wealth*, J. Kendrick and B. Vaccara, eds., National Bureau of Economic Research, University of Chicago Press, Chicago.

Koop, G., 2001, "Comparing the Performance of Baseball Players: A Multiple Output Approach," Working Paper, University of Glasgow, http://http://www.gla.ac.uk/Acad/PolEcon/Koop.

Koop, G., J. Osiewalski, and M. Steel, 1994, "Bayesian Efficiency Analysis with a Flexible Form: The AIM Cost Function," *Journal of Business and Economic Statistics*, 12, pp. 339–346.

Koop, G., J. Osiewalski, and M. Steel, 1997, "Bayesian Efficiency Analysis Through Individual Effects: Hospital Cost Frontiers," *Journal of Econometrics*, 76, pp. 77–106.

Koop, G., J. Osiewalski, and M. Steel, 1999, "The Components of Output Growth: A Stochastic Frontier Approach," *Oxford Bulletin of Economics and Statistics*, 61, pp. 455–486.

Koop, G., and M. Steel, 2001, "Bayesian Analysis of Stochastic Frontier Models," in *Companion to Theoretical Econometrics*, B. Baltagi, ed., Blackwell Publishers, Oxford, UK.

Koop, G., M. Steel, and J. Osiewalski, 1995, "Posterior Analysis of Stochastic Frontier Models Using Gibbs Sampling," *Computational Statistics*, 10, pp. 353–373.

Kopp, R., and J. Mullahy, 1989, "Moment-Based Estimation and Testing of Stochastic Frontier Models," Discussion Paper No. 89-10, Resources for the Future, Washington, DC.

Kotzian, P., 2005, "Productive Efficiency and Heterogeneity of Health Care Systems: Results of a Measurement for OECD Countries," Working Paper, London School of Economics, London.

Kumbhakar, S., 1989, "Estimation of Technical Efficiency Using Flexible Functional Forms and Panel Data," *Journal of Business and Economic Statistics*, 7, pp. 253–258.

Kumbhakar, S., 1990, "Production Frontiers and Panel Data, and Time Varying Technical Inefficiency," *Journal of Econometrics*, 46, pp. 201–211.

Kumbhakar, S., 1991a, "Estimation of Technical Inefficiency in Panel Data Models with Firm- and Time-Specific Effects," *Economics Letters*, 36, pp. 43–48.

Kumbhakar, S., 1991b, "The Measurement and Decomposition of Cost Inefficiency: The Translog Cost System," *Oxford Economic Papers*, 43, pp. 667–683.

Kumbhakar, S., 1993, "Production Risk, Technical Efficiency, and Panel Data," *Economics Letters*, 41, pp. 11–16.

Kumbhakar, S., 2001, "Estimation of Profit Functions When Profit Is Not Maximum," *American Journal of Agricultural Economics*, 83, pp. 1715–1736.

Kumbhakar, S., and Bhattacharyya, A., 1992, "Price Distortions and Resource Use Efficiency in Indian Agriculture: A Restricted Profit Function Approach," *Review of Economics and Statistics*, 74, pp. 231–239.

Kumbhakar, S., S. Ghosh, and J. McGuckin, 1991, "A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in U.S. Dairy Farms," *Journal of Business and Economic Statistics*, 9, pp. 279–286.

Kumbhakar, S., and A. Heshmati, 1995, "Efficiency Measurement in Swedish Dairy Farms 1976–1988 Using Rotating Panel Data," *American Journal of Agricultural Economics*, 77, pp. 660–674.

Kumbhakar, S., and L. Hjalmarsson, 1995, "Labor Use Efficiency in Swedish Social Insurance Offices," *Journal of Applied Econometrics*, 10, pp. 33–47.

Kumbhakar, S., G. Karagiannis, and E. Tsionas, 2004, "A Distance Function Approach for Estimating Technical and Allocative Inefficiency," *Indian Economic Review*, 1, pp. 31–54.

Kumbhakar, S., and K. Lovell, 2000, *Stochastic Frontier Analysis*, Cambridge University Press, Cambridge, UK.

Kumbhakar, S., B. Park, L. Simar, and E. Tsionas, 2005, "Nonparametric Stochastic Frontiers: A Local Maximum Likelihood Approach," Working Paper, Department of Economics, State University of New York, Binghamton.

Kumbhakar, S., and E. Tsionas, 2002, "Scale and Efficiency Measurement Using Nonparametric Stochastic Frontier Models," Working Paper, Department of Economics, State University of New York, Binghamton.

Kumbhakar, S., and E. Tsionas, 2004, "Estimation of Technical and Allocative Inefficiency in a Translog Cost System: An Exact Maximum Likelihood Approach," Working Paper, Department of Economics, State University of New York, Binghamton.

Kumbhakar, S., and E. Tsionas, "The Joint Measurement of Technical and Allocative Inefficiency: An Application of Bayesian Inference in Nonlinear Random Effects Models," *Journal of the American Statistical Association*, 100, 2005a, pp. 736–747.

Kumbhakar, S., and E. Tsionas, 2005b, "Estimation of Stochastic Frontier Production Functions with Input-Oriented Technical Efficiency," Working Paper, Department of Economics, State University of New York, Binghamton.

Kumbhakar, S., and E. Tsionas, "Measuring Technical and Allocative Efficiency in the Translog Cost System: A Bayesian Approach," *Journal of Econometrics*, 126, 2005, pp. 355–384.

Kurkalova, L., and A. Carriquiry, 2003, "Input and Output Oriented Technical Efficiency of Ukranian Collective Farms, 1989–1992: Bayesian Analysis of a Stochastic Frontier Production Model," *Journal of Productivity Analysis*, 20, pp. 191–212.

Lang, G., and P. Welzel, 1998, "Technology and Cost Efficiency in Universal Banking: A 'Thick Frontier' Analysis of German Banking Industry," *Journal of Productivity Analysis*, 10, pp. 63–84.

Lee, L., 1983, "A Test for Distributional Assumptions for the Stochastic Frontier Function," *Journal of Econometrics*, 22, pp. 245–267.

Lee, L., 1993, "Asymptotic Distribution of the MLE for a Stochastic Frontier Function with a Singular Information Matrix," *Econometric Theory*, 9, pp. 413–430.

Lee, L., and M. Tyler, 1978, "The Stochastic Frontier Production Function and Average Efficiency," *Journal of Econometrics*, 7, pp. 385–390.

Lee, Y., and Schmidt, P., 1993, "A Production Frontier Model with Flexible Temporal Variation in Technical Efficiency," in *The Measurement of Productive Efficiency*, H. Fried, K. Lovell, and S. Schmidt, eds., Oxford University Press, Oxford, UK.

Leibenstein, H., 1966, "Allocative Efficiency vs. X-Efficiency," *American Economic Review*, 56, pp. 392–415.

Leibenstein, H., 1975, "Aspects of the X-Efficiency Theory of the Firm," *Bell Journal of Economics*, 6, pp. 580–606.

Linna, M., 1998, "Measuring Hospital Cost Efficiency with Panel Data Models," *Health Economics*, 7, pp. 415–427.

Lovell, K., 1993, "Production Frontiers and Productive Efficiency," in *The Measurement of Productive Efficiency*, H. Fried, K. Lovell, and S. Schmidt, eds., Oxford University Press, Oxford, UK.

Lovell, K., and R. Sickles, 1983, "Testing Efficiency Hypotheses in Joint Production: A Parametric Approach," *Review of Economics and Statistics*, 65, pp. 51–58.

Lozano-Vivas, A., 1997, "Profit Efficiency for Spanish Savings Banks," *European Journal of Operational Research*, 98, pp. 381–394.

Martinez-Budria, E., S. Jara-Diaz, and F. Ramos-Real, 2003, "Adopting Productivity Theory to the Quadratic Cost Function: An Application to the Spanish Electric Sector," *Journal of Productivity Analysis*, 20, pp. 213–229.

Meeusen, W., and J. van den Broeck, 1977, "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error," *International Economic Review*, 18, pp. 435–444.

Mester, L., 1994, "Efficiency of Banks in the Third Federal Reserve District," Working Paper 94-13, Wharton School, University of Pennsylvania, Philadelphia.

Migon, H., and E. Medici, 2001, "Bayesian Hierarchical Models for Stochastic Production Frontier," Working Paper, UFRJ, Brazil.

MRC, 2005, "The BUGS Project," http://http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml, Biostatistics Unit, Cambridge University, Cambridge, UK.

Murillo-Zamorano, L., 2004, "Economic Efficiency and Frontier Techniques," *Journal of Economic Surveys*, 18, pp. 33–77.

Murillo-Zamorano, L., and R. Vega-Cervera, 2001, "The Use of Parametric and Nonparametric Frontier Methods to Measure the Productive Efficiency in the Industrial Sector: A Comparative Study," *International Journal of Production Economics*, 69, pp. 265–275.

Nerlove, M., 1963, "Returns to Scale in Electricity Supply," in *Measurement in Economics*, C. Christ et al., eds., Stanford University Press, Stanford, CA.

O'Donnell, J., and T. Coelli, 2005, "A Bayesian Approach to Imposing Curvature on Distance Functions," *Journal of Econometrics*, 126, pp. 493–523.

O'Donnell, C., and W. Griffiths, 2004, "Estimating State Contingent Production Frontiers," Working Paper Number 911, Department of Economics, University of Melbourne.

Orea, C., and S. Kumbhakar, 2004, "Efficiency Measurement Using a Latent Class Stochastic Frontier Model," *Empirical Economics*, 29, pp. 169–184.

Osiewalski, J., and M. Steel, 1998, "Numerical Tools for the Bayesian Analysis of Stochastic Frontier Models," *Journal of Productivity Analysis*, 10, pp. 103–117.

Pagan, A., and A. Hall, 1983, "Diagnostic Tests as Residual Analysis," *Econometric Reviews*, 2, pp. 159–218.

Paris, Q., and M. Caputo, 2004, "Efficient Estimation by the Joint Estimation of All the Primal and Dual Relations," Working Paper, Department of Agricultural and Resource Economics, University of California, Davis.

Park, B., R. Sickles, and L. Simar, 1998, "Stochastic Panel Frontiers: A Semiparametric Approach," *Journal of Econometrics*, 84, pp. 273–301.

Park, B., and L. Simar, 1992, "Efficient Semiparametric Estimation in Stochastic Frontier Models," *Working Paper*, Department of Economics, CORE, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

Park, B., and L. Simar, 1994, "Efficient Semiparametric Estimation in a Stochastic Frontier Model," *Journal of the American Statistical Association*, 89, pp. 929–936.

Pestieau, P., and H. Tulkens, 1993, "Assessing the Performance of Public Sector Activities: Some Recent Evidence from the Productive Efficiency Viewpoint," *Finanz Archiv*, 50, pp. 293–323.

Pitt, M., and L. Lee, 1981, "The Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry," *Journal of Development Economics*, 9, pp. 43–64.

QMS, Inc., 2005, *EViews Reference Guide*, http://http://www.eviews.com, Irvine, CA.

Ray, S., and K. Mukherjee, 1995, "Comparing Parametric and Nonparametric Measures of Efficiency: A Reexamination of the Christensen and Greene Data," *Journal of Quantitative Economics*, 11, pp. 155–168.

Reifschnieder, D., and R. Stevenson, 1991, "Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency," *International Economic Review*, 32, pp. 715–723.

Reinhard, S., K. Lovell, and G. Thijssen, 1999, "Econometric Applications of Technical and Environmental Efficiency: An Application to Dutch Dairy Farms," *American Journal of Agricultural Economics*, 81, pp. 44–60.

Richmond, J., 1974, "Estimating the Efficiency of Production," *International Economic Review*, 15, pp. 515–521.

Ritter, C., and L. Simar, 1997, "Pitfalls of Normal-Gamma Stochastic Frontier Models," *Journal of Productivity Analysis*, 8, pp. 167–182.

Rosko, M., 2001, "Cost Efficiency of US Hospitals: A Stochastic Frontier Approach," *Health Economics*, 10, pp. 539–551.

Salvanes, K., and S. Tjotta, 1998, "A Note on the Importance of Testing for Regularities for Estimated Flexible Functional Forms," *Journal of Productivity Analysis*, 9, pp. 133–143.

Samuelson, P., 1938, *Foundations of Economic Analysis*, Harvard University Press, Cambridge, MA.

SAS Institute, Inc., 2005, *SAS Reference Guide*, http://http://www.sas.com, Cary, NC.

Schmidt, P., 1976, "On the Statistical Estimation of Parametric Frontier Production Functions," *Review of Economics and Statistics*, 58, pp. 238–239.

Schmidt, P., 1985, "Frontier Production Functions," *Econometric Reviews*, 4, pp. 289–328.

Schmidt, P., and T. Lin, 1984, "Simple Tests of Alternative Specifications in Stochastic Frontier Models," *Journal of Econometrics*, 24, pp. 349–361.

Schmidt, P., and R. Sickles, 1984, "Production Frontiers and Panel Data," *Journal of Business and Economic Statistics*, 2, pp. 367–374.

Shephard, R., 1953, *Cost and Production Functions*, Princeton University Press, Princeton, NJ.

Sickles, R., 1987, "Allocative Inefficiency in the Airline Industry: A Case for Deregulation," in *Studies in Productivity Analysis*, Vol. 7, A. Dogramaci, ed., Kluwer-Nijoff, Boston.

Sickles, R., 2005, "Panel Estimators and the Identification of Firm Specific Efficiency Levels in Parametric, Semiparametric and Nonparametric Settings, *Journal of Econometrics*, 126, 2005, pp. 305–324.

Sickles, R., D. Good, and L. Getachew, 2002, "Specification of Distance Functions Using Semi- and Nonparametric Methods with an Application to the Dynamic Performance of Eastern and Western European Air Carriers," *Journal of Productivity Analysis*, 17, pp. 133–156.

Sickles, R., D. Good, and R. Johnson, 1986, "Allocative Distortions and the Regulatory Transition of the Airline Industry," *Journal of Econometrics*, 33, pp. 143–163.

Sickles, R., and M. Streitweiser, 1992, "Technical Inefficiency and Productive Decline in the U.S. Interstate Natural Gas Pipeline Industry under the Natural Gas Policy Act," *Journal of Productivity Analysis*, 3, pp. 119–134.

Simar, L., 1992, "Estimating Efficiencies from Frontier Models with Panel Data: A Comparison of Parametric, Nonparametric, and Semiparametric Methods with Bootstrapping," *Journal of Productivity Analysis*, 3, pp. 167–203.

Simar, L., 1996, "Aspects of Statistical Analysis in DEA-Type Frontier Models," *Journal of Productivity Analysis*, 7, 177–185.

Simar, L., K. Lovell, and P. van den Eeckhaut, 1994, "Stochastic Frontiers Incorporating Exogenous Influences on Efficiency," Discussion Paper No. 9403, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

Simar, L., and P. Wilson, 1998, "Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models," *Management Science*, 44, pp. 49–61.

Simar, L., and P. Wilson, 1999, "Of Course We Can Bootstrap DEA Scores! But, Does It Mean Anything?" *Journal of Productivity Analysis*, 11, pp. 67–80.

Smith, M., 2004, "Stochastic Frontier Models with Correlated Error Components," Working Paper, Department of Econometrics and Business Statistics, University of Sydney.

Stata, Inc., 2005, *Stata Reference Guide*, http://http://www.stata.com, College Station, TX.

Stevenson, R., 1980, "Likelihood Functions for Generalized Stochastic Frontier Estimation," *Journal of Econometrics*, 13, pp. 58–66.

Swamy, P., and G. Tavlas, 2001, "Random Coefficient Models," in *Companion to Theoretical Econometrics*, B. Baltagi, ed., Blackwell Publishers, Oxford, UK.

Timmer, P., 1971, "Using a Probabilistic Frontier Production Function to Measure Technical Efficiency," *Journal of Political Economy*, 79, pp. 776–794.

Train, K., 2003, *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, UK.

Trethaway, M., and R. Windle, 1983, "U.S. Airline Cross Section: Sources of Data," Working Paper, Department of Economics, University of Wisconsin, Madison.

Tsionas, E. 2000a, "Combining DEA and Stochastic Frontier Models: An Empirical Bayes Approach," Working Paper, Department of Economics, Athens University of Economics and Business.

Tsionas, E., 2000b, "Full Likelihood Inference in Normal-Gamma Stochastic Frontier Models," *Journal of Productivity Analysis*, 13, pp. 183–206.

Tsionas, E., 2001a, "An Introduction to Efficiency Measurement Using Bayesian Stochastic Frontier Models," *Global Business and Economics Review*, 3, pp. 287–311.

Tsionas, E., 2001b, "Combining DEA and Stochastic Frontier Models: An Empirical Bayes Approach," Working Paper, Department of Economics, Athens University of Business and Economics.

Tsionas, E., 2002, "Stochastic Frontier Models with Random Coefficients," *Journal of Applied Econometrics*, 17, pp. 127–147.

Tsionas, E., 2003, "Inference in Dynamic Stochastic Frontier Models," Working Paper, Department of Economics, Athens University of Economics and Business.

Tsionas, E., 2004, "Maximum Likelihood Estimation of Nonstandard Stochastic Frontiers by the Fourier Transform," Working Paper, Department of Economics, Athens University of Economics and Business.

Tsionas, E., and W. Greene (2003), "Non-Gaussian Stochastic Frontier Models," Working Paper, Department of Economics, Athens University of Economics and Business.

TSP International, 2005, *TSP Reference Guide*, http://http://www.tspintl.com, Palo Alto, CA.

van den Broeck, J., G. Koop, J. Osiewalski, and M. Steel, 1994, "Stochastic Frontier Models: A Bayesian Perspective," *Journal of Econometrics*, 61, pp. 273–303.

Vitaliano, D., 2003, "The Thrift Industry and the Community Reinvestment Act: Assessing the Cost of Social Responsibility," Working Paper No. 0312, Department of Economics, Rensellaer Polytechnic Institute, Troy, NY.

Vuong, Q. 1989, "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica* 57, pp. 307–334.

Wagenvoort, R., and P. Schure, 2005, "A Recursive Thick Frontier Approach to Estimating Production Efficiency," Working Paper EWP0503, Department of Economics, University of Victoria, Australia.

Waldman, D., 1982, "A Stationary Point for the Stochastic Frontier Likelihood," *Journal of Econometrics*, 18, pp. 275–279.

Wang, H.-J., 2002, "Heteroscedasticity and Non-monotonic Efficiency Effects of Stochastic Frontier Model," Institute of Economics, Academica Sinica, Taiwan, http://www.sinica.edu.tw/∼wanghj/jpa02b.pdf.

Wang, H., and P. Schmidt, 2002, "One Step and Two Step Estimation of the Effects of Exogenous Variables on Technical Efficiency Levels," *Journal of Productivity Analysis*, 18, pp. 129–144.

Weinstein, M., 1964, "The Sum of Values from a Normal and a Truncated Normal Distribution," *Technometrics*, 6, pp. 104–105, 469–470.

WHO, 2000, *The World Health Report, 2000, Health Systems: Improving Performance*, World Health Organization, Geneva.

Winsten, C., 1957, "Discussion on Mr. Farrell's Paper," *Journal of the Royal Statistical Society, Series A, General*, 120, pp. 282–284.

Xue, M., and P. Harker, 1999, "Overcoming the Inherent Dependency of DEA Efficiency Scoring: A Bootstrap Approach," Working Paper, Wharton Financial Institutions Center, University of Pennsylvania, Philadelphia.

Zellner, A., J. Kmenta, and J. Dreze, 1966, "Specification and Estimation of Cobb-Douglas Production Functions," *Econometrica*, 34, pp. 784–795.

Zellner, A., and N. Revankar, 1969, "Generalized Production Functions," *Review of Economic Studies*, 36, pp. 241–250.

Zellner, A., and J. Tobias, 2001, "Further Results on Bayesian Method of Moments Analysis of the Multiple Regression Model," *International Economic Review*, 42, pp. 121–140.