4 The Estimation and Decomposition of Cost Efficiency

4.1 INTRODUCTION

In Chapter 3 we considered various approaches to the estimation of technical efficiency. The standard against which technical efficiency was estimated was provided by the production frontier, and we adopted an *output-oriented* approach to the estimation of technical efficiency. In this chapter we consider various approaches to the estimation of cost efficiency. The standard against which cost efficiency is estimated is provided by the cost frontier, and we adopt an *input-oriented* approach to the estimation of cost efficiency.

Several significant differences between the estimation of output-oriented technical efficiency and the estimation of input-oriented cost efficiency should be noted.

The first difference concerns data requirements. The estimation of technical efficiency requires information on input use and output provision, whereas the estimation of cost efficiency requires information on input prices, output quantities, and total expenditure on the inputs used, and depending on the model, perhaps input quantities or input cost shares as well. The data requirements for the estimation of cost efficiency are more onerous in some situations and less onerous in others.

The second difference concerns the number of outputs. Estimation of a cost frontier can be accomplished in situations in which producers produce multiple outputs, whereas estimation of a production frontier requires that producers produce a single output. To use

quantity data on multiple inputs and multiple outputs to estimate technical efficiency requires the estimation of either of the two distance functions introduced in Section 2.2.3. An output distance function, which is dual to a revenue frontier, can be used to estimate output-oriented technical efficiency, as in Section 3.2.3. Alternatively, an input distance function, which is dual to a cost frontier, can be used to estimate input-oriented technical efficiency.

The third difference concerns quasi-fixed inputs. In the estimation of a stochastic production frontier, efficiency measurement is output oriented and all inputs are treated equally. No distinction is made between variable inputs and quasi-fixed inputs, and so knowledge that some inputs are not variable during the time period under consideration is not exploited. However in the estimation of a stochastic cost frontier, efficiency measurement is input oriented and it is possible to treat variable and quasi-fixed inputs differently. In this way knowledge concerning quasi-fixity of some inputs is exploited by replacing a cost frontier with a variable cost frontier.

The fourth difference concerns behavioral assumptions. The estimation of technical efficiency does not require the imposition of a behavioral objective on producers, whereas the estimation of cost efficiency does. Such an objective may be inappropriate, although it is hard to conceive of many situations in which it is. It may also be unrealistic, if producers are constrained in their ability to freely adjust their use of inputs, and such constraints are not explicitly modeled. However if not all inputs are variable, due perhaps to short-run fixity or to contractual arrangement, then as we just mentioned a variable cost frontier can be used to estimate variable cost efficiency. In our view the (total or variable) cost minimization objective is an appropriate objective in many environments. It is particularly appropriate in competitive environments in which input prices (rather than input quantities) are exogenous, and in which output is demand driven, and so also can be considered to be exogenous. The more competitive the operating environment, the more appropriate the cost efficiency criterion becomes. Ironically, many regulated industries also satisfy these exogeneity criteria, despite the facts that they are generally noncompetitive and that the regulatory constraints are rarely incorporated into models of cost efficiency. Moreover, in many industries (such as electricity generation, for example) output is not storable, and so the output maximization objective that underlies the estimation of output-oriented technical efficiency would be inappropriate.

The final difference concerns the information that can be obtained from the efficiency estimation exercise. Whereas technical efficiency cannot be decomposed, cost efficiency can be decomposed, and in many circumstances it is desirable to do so. As we indicated in Section 2.4.1, any departure from cost efficiency has two potential sources, input-oriented technical inefficiency and input allocative inefficiency. To the extent that technical and allocative inefficiency have different causes, a determination of which of the two constitutes the main source of cost inefficiency can be a very useful exercise. Thus there are two significant differences between the performance evaluation explored in this chapter and that explored in Chapter 3. First, since input-oriented technical efficiency is necessary, but not sufficient, for cost efficiency, the degree of cost efficiency is not greater than the magnitude of input-oriented technical efficiency, the difference being the extent of input allocative efficiency. Second, measures of input-oriented technical efficiency can differ from measures of output-oriented technical efficiency. As we indicated in Section 2.3.1, the two measures are equal if either measure equals one, so that production is technically efficient, or if production is technically inefficient and production technology satisfies constant returns to scale. If neither condition holds, then input-oriented technical efficiency is greater than or less than output-oriented technical efficiency according as returns to scale are increasing or decreasing over the relevant region of production technology. What this means is that any comparison of producer performance based on technical efficiency estimates obtained from the models developed in Chapter 3, and technical efficiency estimates obtained from the models to be developed in this chapter, must be treated with caution.

This chapter is organized as follows.

In Section 4.2 we consider the estimation of cost efficiency when only cross-sectional data are available.

In Section 4.2.1 we develop single-equation cost frontier models. These models are based on expenditure, output quantity, and input price data; they do not utilize input quantity data in their estimation. They can be used to estimate cost efficiency, or they can be used to estimate input-oriented technical efficiency if an assumption of input allocative efficiency is maintained. In the latter case a direct link between the material in Chapter 3 and that in Chapter 4 is established, subject to the qualification concerning the different orientations. It is not possible to decompose estimated cost efficiency

into estimates of technical and allocative efficiency with a single-equation model. In Section 4.2.1.1 we consider the case in which producers produce a single output, and we use a Cobb–Douglas cost frontier as the standard against which to estimate cost efficiency. The advantage of the Cobb–Douglas specification is its self-duality property, which enables us to go back and forth between the cost frontier and the production frontier. In Section 4.2.1.2 we consider the case in which producers produce multiple outputs, and we use a translog cost frontier as the standard against which to measure cost efficiency. The translog functional form is not self-dual; indeed it has no known dual. However it has the advantages of flexibility, which reduces the likelihood of confounding the structure of the cost frontier with variation in cost efficiency, and an ability to incorporate multiple outputs into the analysis. In Section 4.2.1.3 we consider the case in which producers use multiple inputs to produce a single output, and in which some of the inputs are quasi-fixed. We use a translog variable cost frontier as the standard against which to measure variable cost efficiency. The techniques developed in Chapter 3 for the estimation of technical efficiency can be applied with only minor modification to the problem of estimating cost efficiency within a single-equation framework.

In Section 4.2.2 we develop simultaneous-equation cost frontier models. These models are based on expenditure, output quantity, input price, and either input quantity or input cost share data. The chief advantage of these models is that, since they exploit additional information, they can be used to decompose estimated cost efficiency into estimates of the cost of input-oriented technical efficiency and the cost of input allocative efficiency. In Section 4.2.2.1, which extends Section 4.2.1.1, we consider the case in which producers produce a single output, and we use a Cobb–Douglas cost frontier as the standard against which to estimate and decompose cost efficiency. As a result of the self-duality property of the Cobb–Douglas functional form, a variety of equation systems can be developed for the purpose of estimating and decomposing cost efficiency. One such system, which we explore in some detail, consists of the first-order conditions for cost minimization. In Section 4.2.2.2, which extends Section 4.2.1.2, we consider the case in which producers produce multiple outputs, and we use a translog cost frontier as the standard against which to estimate and decompose cost efficiency. Since the translog functional form is not self-dual, the choice of equation systems to estimate is limited. In this case the system of equations to be estimated typically consists of the cost frontier and $(N - 1)$ of the input cost share equations.

In Section 4.2.3 we consider the problem of decomposing cost efficiency in greater detail than in previous sections. A decomposition of estimated cost efficiency into estimates of the cost of input-oriented technical efficiency and the cost of input allocative efficiency requires that either input quantity or input cost share data be available, in addition to input price, output quantity, and expenditure data. The decomposition is based on a simultaneous-equation model that utilizes the additional data in the estimation stage.

In Section 4.3 we consider the estimation of cost efficiency when panel data are available. As was the case with the estimation of output-oriented technical efficiency in Chapter 3, the availability of panel data provides many advantages in the estimation of cost efficiency.

In Section 4.3.1 we develop single-equation panel data cost frontier models. As in a cross-sectional context, these models can be used to estimate cost efficiency, or if an assumption of allocative efficiency is maintained they can be used to estimate the magnitude and cost of technical efficiency. These panel data models are structurally similar to the single-equation models developed for use with cross-sectional data in Section 4.2.1. The material presented in this section largely parallels the material presented in Section 3.3.

In Section 4.3.2 we develop simultaneous-equation panel data cost frontier models. These models are structurally similar to the simultaneous-equation models developed for use with cross-sectional data in Section 4.2.2. The material presented in this section has no counterpart in Chapter 3.

In Section 4.4 we discuss a pair of very different approaches to the estimation of cost efficiency. These two approaches are less ambitious, and less formally structured, than the approaches developed in Sections 4.2 and 4.3, but they both generate a modest amount of useful information with a minimum of effort. The first approach is labeled "thick frontier analysis," and can be applied to either cross-sectional data or panel data. The second approach is referred to as a "distribution-free approach," and requires panel data.

We do not analyze the problem of heteroskedasticity in this

chapter, despite the fact that it is apt to be at least as serious a problem in a cost frontier context as in a production frontier context. The techniques developed in Section 3.4 can be applied to the problem of heteroskedasticity in a single-equation cost frontier model, with nothing more than a few sign changes. In addition, Kumbhakar (1996d) provides a detailed analytical and empirical treatment of heteroskedasticity within a simultaneous-equation panel data cost frontier model of the type we discuss in Section 4.3.2.

Section 4.5 concludes with a guide to the relevant literature.

4.2 CROSS-SECTIONAL COST FRONTIER MODELS

A cost frontier can be treated as a single-equation model, just as a production frontier was in Chapter 3. In this case it is possible to obtain estimates of the parameters describing the structure of the cost frontier, as well as producer-specific estimates of cost efficiency. However if input quantity data or input cost share data are available, and if Shephard's lemma is invoked, a cost frontier can be treated as a component of a simultaneous-equation model. In this case it is possible to obtain estimates of the parameters describing the structure of the cost frontier and producer-specific estimates of cost efficiency, as in the single-equation case, and it is also possible to obtain producer-specific estimates of the magnitude and cost of technical efficiency and the magnitude and cost of input allocative efficiency as well. Thus moving from a single-equation model to a simultaneous-equation model requires more data and involves a more complicated estimation problem, but it offers the possibility of gaining more insight into the nature of cost efficiency. We consider single-equation cost frontier models in Section 4.2.1, and we consider simultaneous-equation cost frontier models in Section 4.2.2. In Section 4.2.3 we show how the availability of input quantity data enables one to decompose cost efficiency into its two components.

4.2.1 Single-Equation Cost Frontier Models

In this section we assume that cross-sectional data on the prices of inputs employed, the quantities of outputs produced, and total expenditure are available for each of I producers. The analysis is based on a cost frontier, which can be expressed as

$$E_i \geq c(y_i, w_i; \beta), \qquad i = 1, \ldots, I, \tag{4.2.1}$$

where $E_i = w_i^T x_i = \Sigma_n w_{ni} x_{ni}$ is the expenditure incurred by producer i, $y_i = (y_{1i}, \ldots, y_{Mi}) \geq 0$ is a vector of outputs produced by producer i, $w_i = (w_{1i}, \ldots, w_{Ni}) > 0$ is a vector of input prices faced by producer i, $c(y_i, w_i; \beta)$ is the cost frontier common to all producers, and β is a vector of technology parameters to be estimated. Notice that the input vector x_i used by producer i is not necessarily observed. If it is not observed, cost efficiency cannot be decomposed into the cost of input-oriented technical inefficiency and the cost of input allocative inefficiency. If it is observed, the decomposition can be achieved. We defer discussion of the decomposition of cost efficiency to Sections 4.2.2 and 4.2.3.

In Chapter 2 we expressed the cost efficiency of a producer as $CE(y, x, w)$. Here we write cost efficiency as CE_i, replacing the arguments with a producer identifier. Since CE_i is the cost efficiency of producer i, we have from equation (4.2.1)

$$CE_i = \frac{c(y_i, w_i; \beta)}{E_i}, \tag{4.2.2}$$

which defines cost efficiency as the ratio of minimum feasible cost to observed expenditure. Since $E_i \geq c(y_i, w_i; \beta)$, it follows that $CE_i \leq 1$. $CE_i = 1$ if, and only if, $x_{ni} = x_{ni}(y_i, w_i; \beta)$ $\forall n$ so that $E_i = \Sigma_n w_{ni} x_{ni}(y_i, w_i; \beta)$ attains its minimum feasible value of $c(y_i, w_i; \beta)$. Otherwise $CE_i < 1$ provides a measure of the ratio of minimum cost to observed expenditure.

In equation (4.2.1) the cost frontier $c(y_i, w_i; \beta)$ is deterministic, and so in equation (4.2.2) the entire excess of observed expenditure over minimum feasible cost is attributed to cost inefficiency. Such a formulation ignores the fact that expenditure may be affected by random shocks not under the control of producers. A stochastic cost frontier can be written as

$$E_i \geq c(y_i, w_i; \beta) \cdot \exp\{v_i\}, \tag{4.2.3}$$

where $[c(y_i, w_i; \beta) \cdot \exp\{v_i\}]$ is the stochastic cost frontier. The stochastic cost frontier consists of two parts: a deterministic part $c(y_i, w_i; \beta)$ common to all producers and a producer-specific random part $\exp\{v_i\}$,

which captures the effects of random shocks on each producer. If the cost frontier is specified as being stochastic, the appropriate measure of cost efficiency becomes

$$CE_i = \frac{c(y_i, w_i; \beta) \cdot \exp\{v_i\}}{E_i}, \tag{4.2.4}$$

which defines cost efficiency as the ratio of minimum cost attainable in an environment characterized by $\exp\{v_i\}$ to observed expenditure. $CE_i \leq 1$, with $CE_i = 1$ if, and only if, $E_i = c(y_i, w_i; \beta) \cdot \exp\{v_i\}$. Otherwise $CE_i < 1$ provides a measure of the ratio of minimum feasible cost to observed expenditure.

The estimation of cost efficiency can be based on either equation (4.2.1) or equation (4.2.3). Estimation based on equation (4.2.1) would follow procedures analogous to those developed in Section 3.2.1 for the estimation of technical efficiency relative to a deterministic production frontier. Goal programming, corrected OLS, and modified OLS have all been used to estimate deterministic cost frontiers. However since these procedures have already been developed, but more importantly because we are not enamored of deterministic frontiers of any type, we do not consider the estimation of deterministic cost frontiers. We focus our attention on the estimation of stochastic cost frontier models based on equation (4.2.3). In Section 4.2.1 we consider single-equation models in which cost efficiency is estimated, but cannot be decomposed. The techniques for the estimation of these single-equation models are analogous to the techniques developed in Section 3.2.2 for the estimation of output-oriented technical efficiency. Maximum likelihood and method of moments approaches can both be applied to the estimation of cost efficiency in a single-equation model. In Section 4.2.2 we consider simultaneous-equation models in which cost efficiency can be estimated and decomposed. There is no analogue to these models in Chapter 3. In Section 4.2.3 we consider the decomposition of cost inefficiency in greater detail. We consider simultaneous-equation models in which a decomposition of cost inefficiency is made possible by the availability of input quantity or input cost share data.

4.2.1.1 The Single-Output Cobb–Douglas Cost Frontier

If we assume that the deterministic kernel $c(y_i, w_i; \beta)$ of the single-output cost frontier takes the log-linear Cobb–Douglas functional

form, then the stochastic cost frontier model given in equation (4.2.3) can be written as

$$\begin{aligned} \ln E_i &\geq \beta_o + \beta_y \ln y_i + \sum_n \beta_n \ln w_{ni} + v_i \\ &= \beta_o + \beta_y \ln y_i + \sum_n \beta_n \ln w_{ni} + v_i + u_i, \end{aligned} \tag{4.2.5}$$

where v_i is the two-sided random-noise component, and u_i is the nonnegative cost inefficiency component, of the composed error term $\varepsilon_i = v_i + u_i$. Since a cost frontier must be linearly homogeneous in input prices, $c(y_i, \lambda w_i; \beta) = \lambda c(y_i, w_i; \beta)$, $\lambda > 0$, and either the parameter restriction $\beta_k = 1 - \Sigma_{n \neq k} \beta_n$ must be imposed prior to estimation, or equation (4.2.5) must be reformulated as

$$\ln\left(\frac{E_i}{w_{ki}}\right) = \beta_o + \beta_y \ln y_i + \sum_{n \neq k} \beta_n \ln\left(\frac{w_{ni}}{w_{ki}}\right) + v_i + u_i. \tag{4.2.6}$$

Using equation (4.2.4), a measure of cost efficiency is provided by

$$CE_i = \exp\{-u_i\}. \tag{4.2.7}$$

In both formulations of the stochastic cost frontier, the error term $\varepsilon_i = v_i + u_i$ is asymmetric, being positively skewed since $u_i \geq 0$. Apart from the homogeneity restriction on the β_ns and the direction of the skewness of the error term, the stochastic cost frontier model given by equation (4.2.5) or equation (4.2.6) is structurally indistinguishable from the stochastic production frontier model given by equation (3.2.18). Thus apart from some sign changes, the entire analysis of Section 3.2.2 applies with equal force to the estimation of a stochastic cost frontier. If maximum likelihood techniques are employed to obtain estimates of β and the parameters of the two error components, the same distributional assumptions can be made for the error components in equation (4.2.5) or equation (4.2.6). It is also possible to use method of moments estimation techniques to obtain estimates of β and the parameters of the two error components. In either case the JLMS decomposition can be used to separate noise from cost inefficiency in the residuals. The estimated cost inefficiency component can then be substituted into equation (4.2.7) to obtain producer-specific estimates of cost efficiency.

We now illustrate the use of maximum likelihood techniques to estimate the stochastic Cobb–Douglas cost frontier given in equation (4.2.6). We make the following distributional assumptions:

(i) $v_i \sim$ iid $N(0, \sigma_v^2)$.

(ii) $u_i \sim$ iid $N^+(0, \sigma_u^2)$.

(iii) v_i and u_i are distributed independently of each other, and of the regressors.

The density function of $u \geq 0$ is given in equation (3.2.12). The density function of v is given in equation (3.2.20). The marginal density function of $\varepsilon = v + u$ is

$$\begin{aligned} f(\varepsilon) &= \int_o^\infty f(u,\varepsilon)\,du \\ &= \int_o^\infty \frac{2}{2\pi\sigma_u\sigma_v}\cdot\exp\left\{\frac{-u^2}{2\sigma_u^2}-\frac{(\varepsilon-u)^2}{2\sigma_v^2}\right\}du \\ &= \frac{2}{\sqrt{2\pi}\sigma}\cdot\left[1-\Phi\left(\frac{-\varepsilon\lambda}{\sigma}\right)\right]\cdot\exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\} \\ &= \frac{2}{\sigma}\cdot\phi\left(\frac{\varepsilon}{\sigma}\right)\cdot\Phi\left(\frac{\varepsilon\lambda}{\sigma}\right), \end{aligned} \tag{4.2.8}$$

where $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$, $\lambda = \sigma_u/\sigma_v$, and $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cumulative distribution and density functions. As $\lambda \to 0$ either $\sigma_v^2 \to +\infty$ or $\sigma_u^2 \to 0$, and the symmetric error component dominates the one-sided error component in the determination of ε. As $\lambda \to +\infty$ either $\sigma_u^2 \to +\infty$ or $\sigma_v^2 \to 0$ and the one-sided error component dominates the symmetric error component in the determination of ε. In the former case the stochastic cost frontier model collapses to an OLS cost function model with no variation in cost efficiency, whereas in the latter case the model collapses to a deterministic cost frontier model with no noise. As in Chapter 3, it is possible to conduct a likelihood ratio test of the hypothesis that $\lambda = 0$.

The marginal density function $f(\varepsilon)$ is asymmetrically distributed, with mean and variance

$$\begin{aligned} E(\varepsilon) &= E(u) = \sigma_u\sqrt{\frac{2}{\pi}}, \\ V(\varepsilon) &= \frac{\pi-2}{\pi}\sigma_u^2 + \sigma_v^2. \end{aligned} \tag{4.2.9}$$

Geometrically, $f(\varepsilon)$ looks just like the densities appearing in Figure 3.3, except that the direction of the skewness is reversed.

Using equation (4.2.8), the log likelihood function for a sample of I producers is

$$\ln L = \text{constant} - I\ln\sigma + \sum_i \ln\Phi\left(\frac{\varepsilon_i\lambda}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_i \varepsilon_i^2. \tag{4.2.10}$$

The log likelihood function can be maximized with respect to the parameters to obtain maximum likelihood estimates of all parameters.

The next step is to obtain estimates of the cost efficiency of each producer. We have estimates of $\varepsilon_i = v_i + u_i$, which obviously contain information on u_i. If $\varepsilon_i < 0$, chances are that u_i is not large [since $E(v_i) = 0$], which suggests that this producer is relatively cost efficient, whereas if $\varepsilon_i > 0$, chances are that u_i is large, which suggests that this producer is relatively cost inefficient. The problem is to extract the information that ε_i contains on u_i. A solution to the problem is obtained from the conditional distribution of u_i given ε_i, which contains whatever information ε_i contains concerning u_i. Adapting the JLMS procedure to the estimation of cost efficiency when $u_i \sim N^+(0, \sigma_u^2)$, the conditional distribution of u given ε is

$$\begin{aligned} f(u|\varepsilon) &= \frac{f(u,\varepsilon)}{f(\varepsilon)} \\ &= \frac{1}{\sqrt{2\pi}\sigma_*}\cdot\exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\}\Bigg/\left[1-\Phi\left(\frac{-\mu_*}{\sigma_*}\right)\right], \end{aligned} \tag{4.2.11}$$

where $\mu_* = \varepsilon\sigma_u^2/\sigma^2$ and $\sigma_*^2 = \sigma_u^2\sigma_v^2/\sigma^2$. Since $f(u|\varepsilon)$ is distributed as $N^+(\mu_*, \sigma_*^2)$, either the mean or the mode of this distribution can serve as a point estimator for u_i. They are given by

$$\begin{aligned} E(u_i|\varepsilon_i) &= \mu_{*i} + \sigma_*\left[\frac{\phi(-\mu_{*i}/\sigma_*)}{1-\Phi(-\mu_{*i}/\sigma_*)}\right] \\ &= \sigma_*\left[\frac{\phi(\varepsilon_i\lambda/\sigma)}{1-\Phi(-\varepsilon_i\lambda/\sigma)} + \left(\frac{\varepsilon_i\lambda}{\sigma}\right)\right], \end{aligned} \tag{4.2.12}$$

and

$$M(u_i|\varepsilon_i) = \begin{cases} \varepsilon_i\left(\dfrac{\sigma_u^2}{\sigma^2}\right) & \text{if } \varepsilon_i \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{4.2.13}$$

respectively. Once point estimates of u_i are obtained, estimates of the cost efficiency of each producer can be obtained by substituting either $E(u_i|\varepsilon_i)$ or $M(u_i|\varepsilon_i)$ into equation (4.2.7). It is also possible to adapt the Battese and Coelli (1988) point estimator

$$CE_i = E(\exp\{-u_i\}|\varepsilon_i) = \left[\frac{1-\Phi(\sigma_* - \mu_{*i}/\sigma_*)}{1-\Phi(-\mu_{*i}/\sigma_*)}\right] \cdot \exp\left\{-\mu_{*i} + \frac{1}{2}\sigma_*^2\right\}. \tag{4.2.14}$$

The point estimators of CE_i obtained by substituting equations (4.2.12) and (4.2.14) into equation (4.2.7) can give different results, since $\exp\{E(u_i|\varepsilon_i)\} \neq E[\exp\{u_i\}|\varepsilon_i]$. We prefer the Battese and Coelli point estimator to the JLMS point estimator for the same reasons we did in Chapter 3; the latter is a first-order approximation to the former. Regardless of which estimator is used, however, the estimates of cost efficiency are inconsistent because the variation associated with the distribution of $(u_i|\varepsilon_i)$ is independent of i. It is also possible to obtain confidence intervals for the point estimates of cost efficiency, by exploiting the fact that the density of $(u_i|\varepsilon_i)$ is known to be that of an $N^+(\mu_*, \sigma_*^2)$. Horrace and Schmidt (1995, 1996), Bera and Sharma (1996), and Hjalmarsson, Kumbhakar, and Heshmati (1996) have derived upper and lower bounds on $(u_i|\varepsilon_i)$, which imply lower and upper bounds on $(\exp\{u_i\}|\varepsilon_i)$.

Three final points deserve mentioning. First, equations (4.2.8)–(4.2.14) are the same as the corresponding equations in Chapter 3, apart from some sign changes to reflect the fact that here $\varepsilon_i = v_i + u_i$, whereas in Chapter 3 $\varepsilon_i = v_i - u_i$. Second, it is not necessary to specify the deterministic kernel of the stochastic cost frontier as having a Cobb–Douglas functional form. This form was used for illustrative purposes only, and other forms may be used in its place. Third, regardless of the functional form used, the efficiency information that emerges from the analysis is limited to producer-specific estimates of the cost of inefficiency. With a single-equation model, and without input quantity or input cost share data, it is not possible to decompose these estimates into estimates of the cost of input-oriented technical inefficiency and the cost of input allocative inefficiency. A decomposition requires additional data and a simultaneous-equation model.

4.2.1.2 The Multiple-Output Translog Cost Frontier

A great virtue of the Cobb–Douglas functional form is that its simplicity enables us to focus our attention where it belongs, on the error term, which contains information on the cost of inefficiency. As an empirical matter, however, the simplicity of the Cobb–Douglas functional form creates two problems. As Hasenkamp (1976) noted long ago, in a commentary on Klein's (1947) famous railroad study, a function (or frontier) having the Cobb–Douglas form cannot accommodate multiple outputs without violating the requisite curvature properties in output space. In addition, if the true structure of (single-output) production technology is more complex than its Cobb–Douglas representation, the unmodeled complexity will show up in the error term, perhaps leading to biased estimates of the cost of inefficiency. For these reasons we now introduce the translog functional form, due originally to Christensen, Jorgenson, and Lau (1971). The translog cost frontier has several virtues: (i) It accommodates multiple outputs without necessarily violating curvature conditions; (ii) it is flexible, in the sense that it provides a second-order approximation to any well-behaved underlying cost frontier at the mean of the data; and (iii) it forms the basis of much of the empirical estimation and decomposition of cost efficiency based on a system of equations.

If we assume that the deterministic kernel $c(y_i, w_i; \beta)$ of the multiple-output cost frontier takes the log-quadratic translog functional form, then the stochastic cost frontier model given in equation (4.2.5) can be written as

$$\begin{aligned}
\ln E_i \geq \beta_o &+ \sum_m \alpha_m \ln y_{mi} + \sum_n \beta_n \ln w_{ni} + \frac{1}{2}\sum_m \sum_j \alpha_{mj} \ln y_{mi} \ln y_{ji} \\
&+ \frac{1}{2}\sum_n \sum_k \beta_{nk} \ln w_{ni} \ln w_{ki} + \sum_n \sum_m \gamma_{nm} \ln w_{ni} \ln y_{mi} + v_i \\
= \beta_o &+ \sum_m \alpha_m \ln y_{mi} + \sum_n \beta_n \ln w_{ni} + \frac{1}{2}\sum_m \sum_j \alpha_{mj} \ln y_{mi} \ln y_{ji} \\
&+ \frac{1}{2}\sum_n \sum_k \beta_{nk} \ln w_{ni} \ln w_{ki} + \sum_n \sum_m \gamma_{nm} \ln w_{ni} \ln y_{mi} + v_i + u_i,
\end{aligned} \tag{4.2.15}$$

where Young's theorem requires that the symmetry restrictions $\alpha_{nk} = \alpha_{kn}$ and $\beta_{mj} = \beta_{jm}$ be imposed, and homogeneity of degree +1 in input prices requires imposition of the additional restrictions $\Sigma_n \beta_n = 1, \Sigma_n \beta_{nk} = 0\ \forall k$, and $\Sigma_n \gamma_{nm} = 0\ \forall m$. As usual, v_i is the two-sided noise component, and u_i is the nonnegative cost inefficiency component, of the composed error term $\varepsilon_i = v_i + u_i$. The one-sided error component u_i captures the composite cost of input-oriented technical inefficiency and input allocative inefficiency; if technical inefficiency had an output orientation, it would also interact with the regressors in equation (4.2.15). If $M = 1$, equation (4.2.15) collapses to a single-output translog cost frontier. If, in addition, $\beta_{nk} = \gamma_n = 0\ \forall n, k$, then the translog cost frontier collapses to the Cobb–Douglas cost frontier given in equation (4.2.5). These and other parametric restrictions are testable.

It is possible to estimate a translog cost frontier and to obtain producer-specific estimates of cost efficiency, by following the procedures described in Section 4.2.1.1 for the estimation of a Cobb–Douglas cost frontier. Nothing changes, except for the functional form of $c(y_i, w_i; \beta)$. However if either M or N is large, a large sample size will be required. Moreover, multicollinearity among the regressors is likely to lead to imprecise estimates of many parameters in the model, possibly including those characterizing the two error components. Thus the benefit of flexibility is likely to be offset by the cost of statistically insignificant parameter estimates. For this reason the translog cost frontier is infrequently estimated as a single-equation model. We return to the translog cost frontier in a simultaneous-equation setting in Section 4.2.2.2.

4.2.1.3 *The Single-Output Translog Variable Cost Frontier*

Suppose that a sample of $i = 1, \ldots, I$ producers each use a vector of variable inputs $x_i = (x_{1i}, \ldots, x_{Ni}) > 0$, available at prices $w_i = (w_{1i}, \ldots, w_{Ni}) > 0$, and a vector of quasi-fixed inputs $z_i = (z_{1i}, \ldots, z_{Qi}) > 0$, to produce a single output $y_i > 0$. Producers incur variable expense $VE_i = \Sigma_n w_{ni} x_{ni}$ in the process. Then the relevant frontier against which to measure their efficiency is the stochastic variable cost frontier $vc(y_i, w_i, z_i; \beta) \cdot \exp\{v_i + u_i\}$, where $v_i \sim N(0, \sigma_v^2)$ captures the effects of statistical noise and $u_i \geq 0$ reflects the cost of inefficiency in the allocation of variable inputs. If the deterministic

kernel of this frontier takes the translog functional form, then we have

$$\begin{aligned}\ln VE_i = {} & \beta_o + \beta_y \ln y_i + \sum_n \alpha_n \ln w_{ni} + \sum_q \beta_q \ln z_{qi} + \frac{1}{2}\beta_{yy}(\ln y_i)^2 \\ & + \frac{1}{2}\sum_n \sum_k \alpha_{nk} \ln w_{ni} \ln w_{ki} + \frac{1}{2}\sum_q \sum_r \beta_{qr} \ln z_{qi} \ln z_{ri} \\ & + \sum_n \sum_q \gamma_{nq} \ln w_{ni} \ln z_{qi} + \sum_n \alpha_{yn} \ln y_i \ln w_{ni} \\ & + \sum_q \beta_{yq} \ln y_i \ln z_{qi} + v_i + u_i. \end{aligned} \tag{4.2.16}$$

The usual symmetry and linear homogeneity parameter restrictions can be imposed prior to estimation. The remaining regularity conditions, including those involving z_i, can be tested after estimation. The regularity conditions involving z_i depend on the properties satisfied by GR discussed in Chapter 2. Under the strong monotonicity property $G6$, $vc(y_i, w_i, z_i; \beta)$ is nonincreasing in z_i. Under the convexity property $G7$, $vc(y_i, w_i, z_i; \beta)$ is a convex function in (y_i, z_i). Finally if GR is a cone [so that technology exhibits constant returns to scale in (y_i, x_i, z_i)], $vc(y_i, w_i, z_i; \beta)$ is linearly homogeneous in (y_i, z_i).

If independence assumptions are maintained and if a distributional assumption (e.g., half normal) is imposed on u_i, equation (4.2.16) can be estimated by maximum likelihood. Estimation proceeds exactly as in Sections 4.2.1.1 and 4.2.1.2, so we do not repeat the details here. We do note, however, that the independence assumptions become more restrictive in this model, since now v_i and u_i must be distributed independently of each other and of the regressors, which now include the quasi-fixed input quantities.

The great advantage of equation (4.2.16) is that it exploits information (quasi-fixity of some inputs), which a stochastic production frontier cannot. Moreover, after estimation of equation (4.2.16), shadow prices of each quasi-fixed input can be calculated by means of

$$\frac{-\partial VE_i}{\partial z_{qi}} = \frac{\widehat{VE}_i}{z_{qi}} \cdot \left(\hat{\beta}_q + \sum_r \hat{\beta}_{qr} \ln z_{ri} + \sum_n \hat{\gamma}_{qn} \ln w_{ni} + \hat{\beta}_{yq} \ln y_i \right), \quad q = 1, \ldots, Q, \tag{4.2.17}$$

where $\widetilde{VE}_i = \widehat{VE}_i \cdot \exp\{-\hat{u}_i\}$ is the predicted cost-efficient value of VE_i. If prices $p_i = (p_{1i}, \ldots, p_{Qi}) > 0$ of the quasi-fixed inputs are known, a comparison of predicted shadow prices with actual prices provides an indication of which quasi-fixed inputs are over- or underutilized, given the observed values of (y_i, w_i, z_i). Quasi-fixed input z_{qi} is efficiently utilized if $(-\partial VE_i/\partial z_{qi}) = p_{qi}$, and overutilized (underutilized) if $(-\partial VE_i/\partial z_{qi}) < (>)\ p_{qi}$.

Since total cost is minimized when both variable cost and quasi-fixed cost are minimized, misallocation of quasi-fixed inputs constitutes another type of cost inefficiency. The cost of over- or underutilization of quasi-fixed inputs can be determined as follows. First, set the predicted quasi-fixed input shadow prices given in equation (4.2.17) equal to their actual prices, a necessary condition for total cost minimization. Next, solve this system of Q equations for optimal values z_{qi}^* of the quasi-fixed inputs. Finally, calculate the ratio of (or the difference between) the actual cost $\Sigma_q p_{qi} z_{qi}$ of the quasi-fixed inputs to the cost of an efficient combination of quasi-fixed inputs $\Sigma_q p_{qi} z_{qi}^*$.

4.2.2 Simultaneous-Equation Cost Frontier Models

Single-equation cost frontier models are easy to estimate, but they generate limited information. If all that is desired is producer-specific estimates of cost efficiency, single-equation models are adequate for the task, although degrees of freedom problems are likely to plague the estimation of a flexible cost frontier. However if a decomposition of cost inefficiency into its technical and allocative components is desired, it is necessary to employ data on either input quantities or input cost shares in the estimation of a system of equations. The question is: which system? The obvious answer would be to invoke Shephard's lemma and estimate a system consisting of the cost frontier and the associated cost-minimizing input demand equations, or the natural logarithm of the cost frontier and the associated cost-minimizing input share equations. This is indeed the approach adopted when the translog functional form is specified, but when the Cobb–Douglas functional form is specified, its property of self-duality makes other approaches feasible. We begin by considering the single-output Cobb–Douglas cost frontier in Section 4.2.2.1, and we then move on to the multiple-output translog cost frontier in Section 4.2.2.2.

4.2.2.1 Single-Output Cobb–Douglas Cost Systems

We begin by following Schmidt and Lovell (1979), who first developed a stochastic cost frontier model designed to provide estimates of input-oriented technical efficiency and input allocative efficiency. Their approach exploits the self-duality of the Cobb–Douglas functional form. The stochastic Cobb–Douglas production frontier is

$$\ln y = \beta_o + \sum_n \beta_n \ln x_n + v - u, \tag{4.2.18}$$

where producer subscripts are omitted for the moment. As always, $[\beta_o + \Sigma_n \beta_n \ln x_n]$ represents the deterministic kernel of the stochastic production frontier $[\beta_o + \Sigma_n \beta_n \ln x_n + v]$, and $u \geq 0$ represents output-oriented technical inefficiency. If the producer is assumed to seek to minimize the cost $E = \Sigma_n w_n x_n$ of producing its chosen rate of output, then the first-order conditions for the cost minimization problem can be expressed as the system of equations consisting of equation (4.2.18) and the $(N-1)$ first-order conditions

$$\ln\left(\frac{x_1}{x_n}\right) = \ln\left(\frac{\beta_1 w_n}{\beta_n w_1}\right), \qquad n = 2, \ldots, N. \tag{4.2.19}$$

Input allocative inefficiency can be introduced by converting equation (4.2.19) to

$$\ln\left(\frac{x_1}{x_n}\right) = \ln\left(\frac{\beta_1 w_n}{\beta_n w_1}\right) + \eta_n, \qquad n = 2, \ldots, N. \tag{4.2.20}$$

The terms η_n represent input allocative inefficiency for the input pair x_1 and x_n. Since an input can be over- or underutilized relative to input x_1, individual η_ns can take on positive, zero, or negative values. The direction and magnitude of the input allocative inefficiency involving inputs x_n and x_k is given by the ratio (η_n/η_k). A producer is allocatively efficient in its input use if, and only if, $\eta_n = 0$, $n = 2, \ldots, N$.

Equations (4.2.18) and (4.2.20) incorporate both output-oriented technical inefficiency and input allocative inefficiency, and they can be used to solve for the input demand equations, which are given by

$$\ln x_1 = \ln k_1 + \frac{1}{r}\ln y + \frac{1}{r}\sum_{n>1}\beta_n \ln\left(\frac{w_n}{w_1}\right) + \sum_{n>1}\left(\frac{\beta_n}{r}\right)\eta_n - \frac{1}{r}(v-u)$$

$$\vdots$$

$$\ln x_n = \ln k_n + \frac{1}{r}\ln y + \frac{1}{r}\sum_{n>1}\beta_n \ln\left(\frac{w_n}{w_1}\right) + \sum_{n>1}\left(\frac{\beta_n}{r}\right)\eta_n - \eta_n - \frac{1}{r}(v-u), \qquad n=2,\ldots,N, \tag{4.2.21}$$

where $r = \Sigma_n \beta_n$ provides a measure of returns to scale in production and

$$k_n = \beta_n\left[\exp\{\beta_o\}\prod_n \beta_n^{\beta_n}\right]^{-1/r}, \qquad n=1,\ldots,N.$$

The first three terms on the right-hand sides of equations (4.2.21) are the deterministic kernels $\ln x_1(\ln y, \ln w; \beta)$ and $\ln x_n(\ln y, \ln w; \beta)$, $n = 2, \ldots, N$, respectively, of the stochastic cost-minimizing input demand equations. The stochastic cost-minimizing input demand equations are given by $[\ln x_1(\ln y, \ln w; \beta) - v/r]$ and $[\ln x_n(\ln y, \ln w; \beta) - v/r]$, $n = 2, \ldots, N$, respectively. Actual input demands differ from stochastic cost-minimizing input demands due to the presence of both technical inefficiency and input allocative inefficiency.

The impact of technical inefficiency on input demands is given by the terms $(+u/r)$ in each demand equation. Since $u \geq 0$, technical inefficiency increases demand for each input by $(+u/r)$ percent. Technical inefficiency being neutral with respect to input usage, its impact is uniform across input demands. Notice that the output-oriented technical inefficiency $(-u)$ appearing in the production frontier introduced in equation (4.2.18) has been converted to input-oriented technical inefficiency $(+u/r)$ in the input demand equations derived in equations (4.2.21), with the conversion factor being provided by the reciprocal of the magnitude of scale economies, as measured

by the degree of homogeneity of the production frontier. The change of sign reflects the fact that production of *less* than maximum output corresponds to usage of *more* than minimum inputs.

The impact of input allocative inefficiency on input demand is given by the term $[\Sigma_{n>1}(\beta_n/r)\eta_n]$ in the demand equation for x_1, and by the term $[\Sigma_{n>1}(\beta_n/r)\eta_n - \eta_n]$ in the demand equations for the remaining inputs. Since the signs of individual η_ns are not known a priori, it is not possible to say whether demand for a particular input will be increased or reduced by input allocative inefficiency. This can be determined only by estimating the η_ns.

From the input demand equations (4.2.21) we derive an expression for total expenditure $E = \Sigma_n w_n x_n$. This expression incorporates both the cost of technical inefficiency and the cost of input allocative inefficiency, and is given by

$$\ln E = K + \frac{1}{r}\ln y + \sum_n\left(\frac{\beta_n}{r}\right)\ln w_n - \frac{1}{r}(v-u) + (A - \ln r), \tag{4.2.22}$$

where

$$K = \ln\left[\sum_n k_n\right] = \ln r - \frac{\beta_o}{r} - \frac{1}{r}\ln\left[\prod_n \beta_n^{\beta_n}\right],$$

$$A = \sum_{n>1}\left(\frac{\beta_n}{r}\right)\eta_n + \ln\left[\beta_1 + \sum_{n>1}\beta_n \exp\{-\eta_n\}\right].$$

The first three terms on the right-hand side of equation (4.2.22) constitute the deterministic kernel $\ln c(\ln y, \ln w; \beta)$ of the stochastic cost frontier $[\ln c(\ln y, \ln w; \beta) - v/r]$. Actual expenditure exceeds minimum cost for either or both of two reasons. The term $(+u/r) \geq 0$ measures the cost of output-oriented technical inefficiency, and attains its minimum value if, and only if, $u = 0$. The term $(A - \ln r) \geq 0$ measures the cost of input allocative inefficiency, and attains its minimum value if, and only if, $\eta_2 = \cdots = \eta_N = 0$.

We now consider how to estimate the magnitudes and costs of technical and input allocative inefficiency in the Cobb–Douglas model. It is possible to estimate the cost frontier given in equation (4.2.22), and to use the JLMS technique to estimate the separate effects on expenditure of *total* cost inefficiency $[(u/r) + (A - \ln r)]$ and noise. But it is not possible to disentangle the contribution of technical inefficiency from that of input allocative inefficiency; the

composed error term is simply intractable. It is also possible to estimate the system of input demand equations given in equations (4.2.21), but the error terms in these equations are also so complicated as to make it impossible to disentangle the effects of technical and input allocative inefficiency. The procedure suggested by Schmidt and Lovell is to estimate the system of first-order conditions for cost minimization given in equations (4.2.18) and (4.2.20), since the error terms in these equations are simple. Adding a producer subscript and rearranging equations (4.2.20), this system becomes

$$\ln y_i = \beta_o + \sum_n \beta_n \ln x_{ni} + v_i - u_i$$

$$\ln\left(\frac{x_{ni}}{x_{1i}}\right) = \ln\left(\frac{\beta_n}{\beta_1}\right) - \ln\left(\frac{w_{ni}}{w_{1i}}\right) - \eta_{ni},$$

$$n = 2, \ldots, N, \; i = 1, \ldots, I. \tag{4.2.23}$$

This system is estimated using maximum likelihood techniques, since the error term in the production frontier is composed of noise and technical inefficiency, and as we saw in Chapter 3, decomposition of this error term is facilitated by the imposition of distributional assumptions on its two components. Four features of this system are noteworthy: (i) Inputs are endogenous, and output and input prices are exogenous; (ii) technical inefficiency is producer specific, and input allocative efficiency is also producer specific for each input pair; (iii) technical inefficiency is not transmitted to the last $(N-1)$ first-order conditions; and (iv) noise does not appear in the input mix equations. Although technical inefficiency increases the demand for all inputs, it does so equiproportionately, so that the ratios of input demands in the first-order conditions are unaffected by technical inefficiency. Consequently we shall assume that the error term in the production frontier is statistically independent of the error terms in the first-order conditions.

We make the following distributional assumptions on the error terms:

(i) $v_i \sim$ iid $N(0, \sigma_v^2)$.

(ii) $u_i \sim$ iid $N^+(0, \sigma_u^2)$.

(iii) $\eta_i = (\eta_{2i}, \ldots, \eta_{Ni})' \sim$ iid $N(0, \Sigma)$.

(iv) v_i is distributed independently of u_i, and each of them is distributed independently of the elements of η_i.

With these distributional assumptions, the joint density function of the error vector $[(v-u), \eta']'$ for a single observation can be written as $f(\varepsilon, \eta) = f_\varepsilon(\varepsilon) \cdot f_\eta(\eta)$, where $\varepsilon = v - u$, the density function $f_\varepsilon(\varepsilon)$ is given in equation (3.2.23), and $f_\eta(\eta)$ is the density function of a multivariate normal variable given by

$$f_\eta(\eta) = (2\pi)^{-(N-1)/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\eta' \Sigma^{-1} \eta)\right\}. \tag{4.2.24}$$

Thus the log likelihood function is

$$\begin{aligned} \ln L &= \text{constant} + \sum_i \ln f_\varepsilon(\varepsilon_i) + \sum_i \ln f_\eta(\eta_i) + I \ln r \\ &= \text{constant} - I \ln \sigma - \frac{I}{2} \ln |\Sigma| + I \ln r \\ &\quad - \frac{1}{2} \sum_i \left[\eta_i' \Sigma^{-1} \eta_i + \left(\frac{1}{\sigma^2}\right) \varepsilon_i^2 \right] + \sum_i \left[1 - \Phi\left(\frac{\varepsilon_i \lambda}{\sigma}\right) \right], \end{aligned} \tag{4.2.25}$$

where

$$\eta_i = \begin{bmatrix} -\ln\left(\frac{x_{2i}}{x_{1i}}\right) + \ln\left(\frac{\beta_2}{\beta_1}\right) - \ln\left(\frac{w_{2i}}{w_{1i}}\right) \\ \vdots \\ -\ln\left(\frac{x_{Ni}}{x_{1i}}\right) + \ln\left(\frac{\beta_N}{\beta_1}\right) - \ln\left(\frac{w_{Ni}}{w_{1i}}\right) \end{bmatrix},$$

$$\varepsilon_i = \ln y_i - \beta_o - \sum_n \beta_n \ln x_{ni},$$

and $r = \Sigma_n \beta_n$ is the Jacobian of the transformation from (ε, η) to $(\ln x_1, \ldots, \ln x_N)$. As in Chapter 3, $\lambda = \sigma_u/\sigma_v$, $\sigma = (\sigma_u^2 + \sigma_v^2)^{1/2}$, and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal variable. This log likelihood function can be maximized with respect to the parameters to obtain maximum likelihood estimates of all parameters in the model.

Once the parameters have been estimated, information of two sorts is provided. The structure of production technology is characterized by estimated values of β_o and the β_ns, and the nature of inefficiency is characterized by estimates of u_i and the η_{ni}s. Producer-specific estimates of technical inefficiency can be extracted from the residuals of the production frontier in equations (4.2.23). Following

JLMS, either the mean or the mode of $(u_i|\varepsilon_i)$ can be used to provide an estimate of technical inefficiency, which is then substituted into $TE_i = \exp\{-u_i\}$ as usual. Producer-specific estimates of the cost of technical inefficiency are obtained from equation (4.2.22) by means of $CTE_i = \exp\{u_i/r\}$. Producer-specific estimates of allocative inefficiency are obtained from the residuals of the input mix equations in equations (4.2.23). These residuals are then substituted into the input demand equations (4.2.21) to provide estimates of the impact of allocative inefficiency on the usage of each input. Producer-specific estimates of the cost of allocative inefficiency are obtained by substituting these same residuals into the expression for $(A - \ln r)$ in equation (4.2.22) to obtain $CAE_i = \exp\{A - \ln r\}$.

In the system given by equations (4.2.23), the input allocative inefficiencies are treated as elements of a random error vector having zero mean. Thus it is expected that mean input allocative inefficiencies are zero, and this is in contrast to the expectation that mean technical inefficiency is positive, since $E(u) = \sqrt{2/\pi}\sigma_u$. It is possible to generalize this model to allow the input mix system error vector to have nonzero mean, in which case the input allocative inefficiencies (and their cost) have nonzero expectation. One motivation for such a generalization would be that the data were generated by an environment in which it is to be expected that input allocative inefficiency of a particular type might be present. An example is provided by utilities subject to rate of return regulation, which according to the Averch–Johnson (1962) hypothesis induces them to overinvest in their rate base (capital) inputs relative to their other inputs. Another example is provided by agriculture, in which evidence suggests that in a wide variety of environments farmers use excessive amounts of fertilizers and pesticides relative to other inputs. In each of these situations, and in others as well, it is desirable to allow input allocative inefficiency to be systematic, and then to test the hypothesis that it is not.

We retain the system of equations (4.2.23), and we retain the distributional assumptions imposed previously, with one exception. We replace (iii) with

(iiis) $\eta_i \sim$ iid $N(\mu, \Sigma)$,

and we follow the same estimation strategy as before. The log likelihood function given in equation (4.2.25) becomes

$$\ln L = \text{constant} - I\ln\sigma - \frac{I}{2}\ln|\Sigma| + I\ln r - \frac{1}{2}\sum_i\left[(\eta_i-\mu)'\Sigma^{-1}(\eta_i-\mu)+\left(\frac{1}{\sigma^2}\right)\varepsilon_i^2\right]+\sum_i\left[1-\Phi\left(\frac{\varepsilon_i\lambda}{\sigma}\right)\right], \tag{4.2.26}$$

which collapses to the log likelihood function given in equation (4.2.25) if $\mu = 0$. This log likelihood function can be maximized with respect to the parameters to obtain maximum likelihood estimates of all parameters, now including the elements of the systematic input misallocation vector μ. The hypothesis that $\mu = 0$, or that any subset of the μs is zero, can be tested by computing a likelihood ratio test statistic in the usual manner. The magnitudes and costs of technical and systematic input allocative inefficiency are calculated as before, recalling that now the η_ns have means of μ_n, $n = 2, \ldots, N$.

In both of the preceding models we assumed that the two error components in the production frontier were statistically independent of each other, and of the error terms in the input mix equations. Since v represents the influence of factors beyond the control of producers, it makes sense to assume that v is independent of u and η. However independence of u and η is a different matter. It is possible to relax this assumption also, and to test the independence between u and η, if there is reason to believe that producers who are relatively technically inefficient are also relatively inefficient in their allocation of inputs, and vice versa. Schmidt and Lovell (1980) developed a model in which technical and input allocative inefficiencies are allowed to be correlated. In such a model distributional assumptions (i) and (ii) are retained, but (iiis) [or its restricted version (iii)] and (iv) are replaced by

$$\text{(v)}\quad \begin{bmatrix}|u|\\ \eta\end{bmatrix} \sim N\left[\begin{pmatrix}0\\ \mu\end{pmatrix}, \Sigma\right],$$

where

$$\Sigma = \begin{bmatrix}\sigma_u^2 & \Sigma_{u\eta}\\ \Sigma_{u\eta}' & \Sigma_{\eta\eta}\end{bmatrix}.$$

Notice that u is allowed to be correlated not with the η_ns, but with the absolute values of the η_ns. If $\Sigma_{u\eta} \neq 0$, then $|u|$ and η are

correlated, which implies that u is uncorrelated with η_n. However u is positively correlated with $|\eta_n|$, and this positive correlation holds regardless of the sign of the nth element of $\Sigma_{u\eta}$. If $\Sigma_{u\eta} = 0$, then u is uncorrelated with the $|\eta_n|$, and we are back to the previous model in which technical and allocative inefficiencies are independently distributed over the sample. This is a testable hypothesis.

In this framework producers who are relatively technically inefficient are also relatively allocatively inefficient, in the sense that their input mixes are farther off their least cost expansion paths, in either direction, than are those of producers who are relatively technically efficient. What matters in this framework is not the directions of allocative inefficiencies, but their magnitudes, because it is the magnitudes of allocative inefficiencies which, like technical inefficiency, raise cost.

4.2.2.2 *The Multiple-Output Translog Cost System*

In the single-output Cobb–Douglas cost system we did not exploit duality theory in the efficiency estimation exercise; we estimated the system of first-order conditions for cost minimization, which were obtained directly from the production frontier. Instead we exploited the self-duality property of the Cobb–Douglas functional form in the efficiency estimation exercise. The ability to express the impact of technical and allocative inefficiency on input demand and total expenditure was the consequence of our ability to express the expenditure and input demand relationships in closed form in terms of the parameters of the production relationship. Very few functional forms have a dual that can be obtained directly from them, and so the strategy adopted in Section 4.2.2.1 is not generally applicable.

In this section we develop a different strategy, one that is based on the (nonfrontier) strategy originally proposed by Christensen and Greene (1976) for estimating the parameters of a translog cost *function*. We begin by reviewing their strategy, which involves estimating a system of equations based on the cost function and its associated input share equations, and then using the estimated cost function parameters to draw inferences concerning the structure of the underlying, but unknown, production technology. The underlying production technology is unknown because, like most flexible functional forms, the translog cost function has no closed-form dual production or transformation function. We then develop a strategy

for converting their cost function to a cost frontier by introducing technical and input allocative inefficiency into the system.

The translog cost function and its associated input cost share equations can be written in deterministic form as

$$\begin{aligned}
\ln E_i &= \beta_o + \sum_m \alpha_m \ln y_{mi} + \sum_n \beta_n \ln w_{ni} + \frac{1}{2}\sum_m \sum_j \alpha_{mj} \ln y_{mi} \ln y_{ji} \\
&\quad + \frac{1}{2}\sum_n \sum_k \beta_{nk} \ln w_{ni} \ln w_{ki} + \sum_n \sum_m \gamma_{nm} \ln w_{ni} \ln y_{mi}, \\
S_{ni} &= \beta_n + \sum_k \beta_{nk} \ln w_{ki} + \sum_m \gamma_{nm} \ln y_{mi}, \qquad n = 1, \ldots, N,
\end{aligned} \tag{4.2.27}$$

where $S_{ni} = \partial \ln E_i / \partial \ln w_{ni} = w_{ni}x_{ni}/E_i$ from Shephard's lemma.

Estimation of the system, rather than just the cost function, adds degrees of freedom and results in more efficient parameter estimates. The system can be estimated by imposing the symmetry and linear homogeneity restrictions listed beneath equation (4.2.15), deleting one input cost share equation (since they sum to unity for each producer), appending additive jointly normally distributed error terms for each remaining equation, and iterating on seemingly unrelated regressions (SUR) until convergence is achieved. Kmenta and Gilbert (1968) have demonstrated that this procedure generates maximum likelihood estimates, and Barten (1969) has demonstrated that these estimates are invariant to which share equation is deleted.

The problem now is to reformulate the system given in equations (4.2.27) into a frontier context. We begin by deleting the first input cost share equation and rewriting the system of N equations as

$$\begin{aligned}
\ln E_i &= \ln c(y_i, w_i; \beta) + v_i + u_i, \\
S_{ni} &= S_{ni}(y_i, w_i; \beta) + \eta_{ni}, \qquad n = 2, \ldots, N
\end{aligned} \tag{4.2.28}$$

where $\ln c(y_i, w_i; \beta)$ is the deterministic kernel of the stochastic translog cost frontier, the $S_{ni}(y_i, w_i; \beta)$ are the deterministic kernels of the stochastic translog input cost share equations, and β represents the set of all technology parameters appearing in the cost frontier in equations (4.2.27). The error component u_i captures the effect on expenditure of inefficiency, either technical inefficiency or input allocative inefficiency or both, depending on how the N error terms u_i and the η_{ni} are interpreted and related. We begin by making the following distributional assumptions on the error terms:

(i) $v_i \sim$ iid $N(0, \sigma_v^2)$.

(ii) $u_i \sim$ iid $N^+(0, \sigma_u^2)$.

(iii) $\eta_i = (\eta_{2i}, \ldots, \eta_{Ni})' \sim N(0, \Sigma)$.

(iv) v_i and u_i are distributed independently of each other, and of the elements of η_i.

These assumptions make sense only if allocative efficiency is assumed, so that v_i and η_i represent statistical noise. This is because if η_i represents allocative inefficiency, it cannot be distributed independently of u_i, since allocative inefficiency raises cost. However if η_i represents statistical noise, then u_i captures the cost, which is equivalent to the magnitude, of input-oriented technical inefficiency. Consequently under this assumption the system of equations (4.2.28) provides no more information than does its cost frontier component. Including the share equations provides more efficient parameter estimates, but they may very well be biased by an inappropriate assumption of allocative efficiency. On the other hand, if it is assumed that η_i represents allocative inefficiency, then u_i captures the cost of both technical and allocative inefficiency. Unfortunately, in this case the distributional assumption (iv) makes no sense, because the cost of allocative inefficiency must vary directly with its magnitude, in which case u_i and η_i cannot be statistically independent. It is of course possible to maintain the independence assumption and estimate the model, but failure of independence is likely to lead to inconsistent parameter estimates. And even then, it is not possible to decompose the cost of inefficiency into its two sources.

The dilemma in the preceding paragraph was first noted by Greene (1980b), and so the problem of specifying a sensible translog system that incorporates both technical and allocative inefficiency has come to be known as "the Greene problem." Schmidt (1984) was the first to propose a solution to the problem. He interpreted $\eta \sim N(0, \Sigma)$ as reflecting allocative inefficiency, and he interpreted $u = u_T + u_A$ as reflecting the cost of technical (u_T) and allocative (u_A) inefficiency, with $u_T \sim N^+(0, \sigma_T^2)$. Rather than making a separate distributional assumption on u_A, he specified the cost of allocative inefficiency as a function of η. Schmidt proposed $u_A = \eta' A \eta$, with A an $N \times N$ positive semidefinite matrix. In this formulation $u_A = 0$ when $\eta = 0$, $u_A > 0$ when $\eta \neq 0$, and u_A is positively correlated with the absolute value of each element of η. No distributional assumption needs to be made for u_A, since an estimate of u_A is obtained from estimates of η and estimates of the elements of A. (Although only $N-1$ share equations are included in the system to be estimated, the elements of η sum to zero, so the missing element can be derived residually.) Schmidt proposed $A = D^{1/(N-1)}\Sigma^+$, where D is the product of the nonzero eigenvalues of Σ, and Σ^+ is the generalized inverse of Σ. With this specification of A, maximum likelihood techniques can be used to provide estimates of the parameters of the cost frontier, the magnitudes of allocative inefficiency, and the cost of technical and allocative inefficiency. Schmidt did not provide an empirical application of his procedure.

Melfi (1984) and Bauer (1985) implemented Schmidt's procedure empirically, but to do so they had to simplify the specification of A. Melfi assumed $A = I$, so that u_A is the sum of squared input share equation errors. An obvious drawback of this specification is that u_A is forced toward zero. Bauer generalized Melfi's specification by allowing A to be a positive semidefinite diagonal matrix whose elements are treated as $(N-1)$ additional parameters to be estimated; in this formulation u_A is a weighted sum of squared input share equation errors. Once distributional assumptions are made for v, u_T, and η, the translog cost frontier system can be estimated by maximum likelihood; the log likelihood function appears in Bauer (1990b).

Kumbhakar (1991) suggested a specification similar to Bauer's, but which incorporates no additional parameters to be estimated. He required that $\eta_n = \partial u_A / \partial \ln w_n$, $n = 2, \ldots, N$, so that input share equation errors are related to the cost of input allocative inefficiency, just as the input share equations themselves are related to the cost frontier, by way of Shephard's lemma. He then showed that the relationship $u_A = \eta' A \eta$ can be written as $u_A = \eta^{*\prime} K \eta^*$, where $\eta^* = (\eta_1, \ldots, \eta_{N-1})$ and K is a symmetric $(N-1) \times (N-1)$ matrix. Now $\partial(\eta^{*\prime} K \eta^*)/\partial \ln w_n = \eta_n$ if

$$-\sum_n k_{kn}\beta^*_{kn} = -\frac{1}{2} \quad \text{and} \quad \sum_n k_{hn}\beta^*_{kn} = 0, \qquad h \neq k, \tag{4.2.29}$$

where $[k_{hn}]$ is the K matrix and $[\beta^*_{kn}]$ is the matrix of coefficients on $[\ln w_k \ln w_n]$ in the translog cost frontier, with the last row and column deleted. These conditions can be expressed in matrix form as $K\beta^* =$

$-(1/2)I$, where β^* is the $[\beta^*_{kn}]$ matrix and I is the identity matrix of order $(N-1)$. Thus the solution for K is $K = -(1/2)\beta^{*-1}$ and we have $u_A = -(1/2)\eta^{*\prime}\beta^{*-1}\eta^*$. In contrast to the previous specifications, there are no additional parameters to be estimated. However for K to be a positive semidefinite matrix, β^{*-1} must be negative semidefinite, and even concavity of the cost frontier in input prices does not guarantee that β^{*-1} is negative semidefinite.

An obvious drawback of all three specifications is that η represents pure allocative inefficiency; there is no noise in the input share equations. Ferrier and Lovell (1990) provided an extension of Bauer's specification by allowing $\eta_{ni} = \eta_n + \xi_{ni}$. In their specification η_{ni} is the error in the nth share equation, η_n represents allocative inefficiency in the use of the nth input, and ξ_{ni} represents noise in the nth input share equation. Unfortunately in this specification allocative inefficiency varies across inputs, but not across producers; this is the price to be paid for introducing noise into the input share equations. Consequently in this specification the cost of allocative inefficiency, $u_A = \eta' A \eta$, is the same for all producers. If distributional assumptions are made for v, u_T, and η, and if A is specified as an $N \times N$ positive semidefinite diagonal matrix, maximum likelihood techniques can be used to obtain estimates of all parameters in the model, and to estimate magnitudes and costs of producer-specific technical inefficiency and allocative inefficiency, which is common to all producers. Ferrier and Lovell assumed that $u_{Ti} \sim N^+(0, \sigma_T^2)$ and $\eta_{ni} \sim N(\eta_n, \sigma_\eta^2)$. An alternative way of introducing persistent allocative inefficiency into the model would be to delete the cross-equation parameter equality restrictions on the β_n in equations (4.2.27). This would leave the β_n in the expenditure equation, but the intercept in the nth input cost share equation would become $(\beta_n + \eta_n)$, $n = 2, \ldots, N$. It would then be possible to test the hypothesis of no allocative inefficiency by testing the hypothesis that the $N-1$ input cost share equation intercepts are jointly equal to the corresponding expenditure equation slope parameters.

4.2.3 Decomposing Cost Inefficiency

In Section 4.2.1 we showed that in a single-equation cost frontier model it is possible to estimate the cost of overall (the sum of technical and input allocative) inefficiency, but that it is not possible

to decompose estimated cost inefficiency into its technical and allocative components. In Section 4.2.2 we showed that, under certain conditions, in a simultaneous-equation cost frontier model it is possible to estimate the magnitude as well as the cost of input allocative inefficiency, and to estimate both the cost and the magnitude of technical inefficiency. The conditions sufficient for a decomposition include either restrictions on the number of outputs and the functional form of the cost frontier (the self-dual, single-output Cobb–Douglas form considered in Section 4.2.2.1) or restrictions on the structure of the disturbance terms in the model (the multiple-output translog cost frontier considered in Section 4.2.2.2). In both cases the ability to decompose the estimated cost of overall inefficiency into its sources comes not from having a system of equations rather than a single equation, but from imposing restrictive assumptions and having additional data, the input quantities or the input cost shares. This suggests that with such data in hand, it may *always* be possible to decompose cost inefficiency, without having to make such restrictive assumptions. Färe and Primont (1996) have demonstrated that this is indeed true analytically, and several authors have achieved such an analytical decomposition within a translog stochastic cost frontier framework. However successful empirical implementation of the decomposition has proved elusive.

The argument that the availability of input quantity or cost share data does permit the decomposition of cost inefficiency originated with Kopp and Diewert (1982) and Zieschang (1983), and was revisited by Mensah (1994). Their argument is intuitively appealing and analytically correct. However as a practical matter their argument is flawed. We will return to the flaw after outlining their analytical argument.

Figure 4.1 provides an illustration of the problem. A producer uses input vector x to produce output vector y, and faces input price vector w. The producer's actual cost is $E = w^T x$, and the producer's minimum cost is $c(y, w; \beta) = w^T x^*$. The producer's cost efficiency is thus $CE = w^T x^*/w^T x = c(y, w; \beta)/w^T x$. The problem is to decompose cost inefficiency into its technical and allocative components. The cost (and magnitude) of input-oriented technical inefficiency is $CTI = w^T x'/w^T x$, and the cost of allocative inefficiency is $CAI = w^T x^*/w^T x' = c(y, w; \beta)/w^T x'$. Thus the decomposition problem boils down to one of finding the two unobserved input quantity vectors x^* and x'.

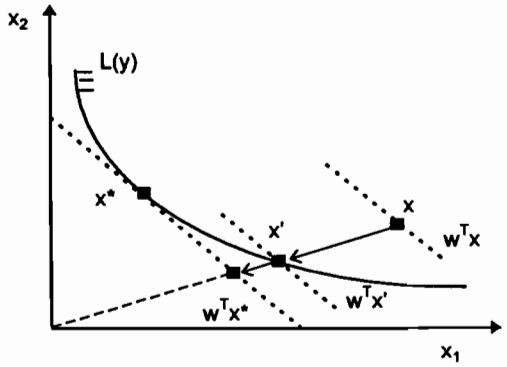

Figure 4.1 The Decomposition of Cost Inefficiency ($N = 2$)

Finding the cost-efficient input quantity vector x^* is easy. From Shephard's lemma, $x^* = \nabla_w c(y, w; \beta)$, and this can be determined even in a single-equation model and even without input quantity data. Finding the technically efficient, but allocatively inefficient, input quantity vector x' is the problem. However it is clear from Figure 4.1 that $x' = \nabla_{w'} c(y, w'; \beta)$ for some unobserved input price vector w'. Now the decomposition problem boils down to one of finding the unobserved input price vector w', from which Shephard's lemma generates the unobserved input quantity vector x'.

Kopp and Diewert demonstrated that w' and x' can be derived by solving the system of $(2N - 1)$ equations

$$\frac{x'_n}{x'_1} - \frac{x_n}{x_1} = 0, \qquad n = 2, \ldots, N,$$
$$x' - \nabla_{w'} c(y, w'; \beta) = 0 \tag{4.2.30}$$

in the $2N$ variables (w', x'). The first set of conditions requires x' to have the same input mix as x, and the second set of conditions requires $x' \in \text{Isoq}\, L(y)$. The system is closed with a normalizing condition on w', such as $w'_1 = 1$, since only relative input prices are required to determine x'. Kopp and Diewert obtained a solution to the system of equations (4.2.30) by using numerical techniques to find the values of the unknown variables (w', x') that minimize the sum of squared differences.

Zieschang simplified the system of $(2N - 1)$ equations to the system of $(N - 1)$ equations

$$\frac{\partial c(y, w'; \beta)/\partial w'_n}{\partial c(y, w'; \beta)/\partial w'_1} - \frac{x_n}{x_1} = 0, \qquad n = 2, \ldots, N, \tag{4.2.31}$$

which is to be solved for (w'_n/w'_1). Once the (w'_n/w'_1) are determined, it follows that $w' = (1, w'_2/w'_1, \ldots, w'_N/w'_1)/[x_1 + \Sigma_{n>1}(w'_n/w'_1)x_n)]$. Finally, $x' = \nabla_{w'} c(y, w'; \beta)$. The computational advantage of the system (4.2.31) over the system (4.2.30) increases with N.

Mensah showed that the unknown input price vector w', and hence the unknown input quantity vector x', could be obtained by solving the system of N equations

$$w'_n - w_n \cdot \left[\frac{\partial \ln c(y, w'; \beta)/\partial \ln w'_n}{w_n x_n / w^T x} \right] = 0, \qquad n = 1, \ldots, N. \tag{4.2.32}$$

Thus the Mensah approach involves inferring w' from w by scaling each element of w by the ratio of its cost-minimizing input share when input prices are w' to its actual cost share when input prices are w.

Each of these solution procedures is analytically elegant, although each procedure requires the numerical solution of a system of nonlinear equations to derive an estimate of the unobserved input price vector w'. Particularly as N and M become large, concerns about the rate at which the system converges to a solution, about the nonnegativity of the solution, and about the existence of multiple local optimal solutions become relevant. There is some empirical evidence that solutions for w' can contain negative elements, or can generate negative costs of either technical or allocative inefficiency, neither of which makes any economic sense. An example of the former problem is reported by Kopp and Diewert, and several examples of the latter problem are reported by Berger and Humphrey (1991). As a way of avoiding such difficulties, Mensah proposed a linear approximation to the nonlinear system (4.2.32). The linear approximation is given by

$$w'_n - w_n \cdot \left[\frac{\partial \ln c(y, w; \beta)/\partial \ln w_n}{w_n x_n / w^T x} \right] \cong 0, \qquad n = 1, \ldots, N. \tag{4.2.33}$$

The approximation replaces cost-minimizing input cost shares when input prices are w' with cost-minimizing input cost shares when input prices are w. Mensah reports empirical evidence that the approximate solution for w' obtained from equation (4.2.33), and the implied

decomposition of cost inefficiency into its technical and allocative components, are very close to the nonlinear solution for w' obtained from equation (4.2.32) and its implied decomposition.

Unfortunately each of the procedures outlined previously has an additional empirical shortcoming: Each is based on a knowledge of the cost frontier $c(y, w; \beta)$, perhaps with w' replacing w. However $c(y, w; \beta)$ is unknown, and must be estimated prior to the decomposition of cost inefficiency, and estimation of $c(y, w; \beta)$ when both technical and allocative inefficiency are present is a formidable problem, as we have indicated. Indeed this problem may be the source of the numerical difficulties mentioned previously. Thus although these "solutions" are analytically correct, they do not solve the fundamental econometric problem of formulating and estimating a translog cost system in the presence of both types of inefficiency because they fail to incorporate statistical noise in an econometrically consistent fashion. The Greene problem is an econometric problem, not an analytical problem. These three studies demonstrate that the analytical problem has been solved. However the econometric problem remains.

In the models discussed in Section 4.2.2.2 allocative inefficiency is modeled in an ad hoc fashion. Each of these specifications appeals to the fact that allocative inefficiency increases cost, and a quadratic expression involving allocative inefficiency is added to the expenditure equation to represent the cost increase due to allocative inefficiency. In these specifications the cost of allocative inefficiency is independent of output quantities and input prices. Kumbhakar (1997) has introduced allocative inefficiency in a theoretically and econometrically consistent manner, by adapting the Schmidt and Lovell (1979) Cobb–Douglas production frontier specification to the translog cost frontier framework. Although he worked out the details for a translog cost frontier system, his approach is applicable to any cost frontier.

A cost frontier incorporating both technical and allocative inefficiency can be expressed as

$$\ln E = \ln c(y, w; \beta) + v + u_T + u_A, \tag{4.2.34}$$

where $c(y, w; \beta)$ is the deterministic kernel of the stochastic cost frontier, v is a stochastic noise error component, the error component $u_T \geq 0$ represents the cost of input-oriented technical inefficiency, and

the error component $u_A \geq 0$ represents the cost of input allocative inefficiency. The latter can be modeled as departures of marginal rates of substitution from input price ratios, and so

$$\frac{f_n(x;\beta)}{f_1(x;\beta)} = \frac{w_n}{w_1} \cdot \exp\{\eta_n\} = \frac{w_n^*}{w_1}, \qquad n = 2, \ldots, N, \tag{4.2.35}$$

where $f(x; \beta)$ is the production frontier dual to $c(y, w; \beta)$. It can be shown that

$$u_A = \ln c\left(y, w^*; \beta\right) - \ln c(y, w; \beta) + \ln G, \tag{4.2.36}$$

where $w^* = (w_1, w_2^*, \ldots, w_N^*)$ and

$$G = \sum_n \left[\frac{\partial \ln c(y, w^*; \beta)}{\partial \ln w_n^*} \right] \cdot \exp\{-\eta_n\} = \sum_n S_n(y, w^*; \beta) \cdot \exp\{-\eta_n\}.$$

The input cost share equations associated with the expenditure equation (4.2.34) are

$$\frac{w_n x_n}{E} = \frac{S_n(y, w^*; \beta)}{G \cdot \exp\{\eta_n\}} = S_n(y, w; \beta) + A\eta_n, \tag{4.2.37}$$

where

$$A\eta_n = \frac{S_n(y, w^*; \beta)}{G \cdot \exp\{\eta_n\}} - S_n(y, w; \beta) \qquad n = 1, \ldots, N.$$

If $c(y, w^*; \beta)$ is assumed to take the translog form, then

$$\begin{aligned} u_A = \ln G &+ \sum_n \beta_n \eta_n + \sum_n \sum_k \beta_{nk} \ln w_k \eta_n \\ &+ \frac{1}{2} \sum_n \sum_k \beta_{nk} \eta_n \eta_k + \sum_n \sum_m \gamma_{nm} \ln y_m \eta_n, \end{aligned} \tag{4.2.38}$$

where

$$\begin{aligned} A\eta_n &= \left[S_n(y, w; \beta) \cdot (1 - G \cdot \exp\{\eta_n\}) + \sum_m \beta_{nm} \eta_m \right] \Big/ G \cdot \exp\{\eta_n\}, \\ &\quad n = 1, \ldots, N, \\ G &= \sum_n \left(\beta_n + \sum_k \beta_{nk} \ln w_k^* + \sum_m \gamma_{nm} \ln y_m \right) \cdot \exp\{-\eta_n\}. \end{aligned} \tag{4.2.39}$$

Although this translog cost system looks like the other translog cost systems considered previously, it differs in that it treats allocative inefficiency in a consistent fashion. Several features of this system should be noted. (1) Both the impact of allocative inefficiency on input cost shares [the $A\eta_n$s in equations (4.2.39)] and the cost of allocative inefficiency [u_A in equation (4.2.37)] are influenced by output quantities and input prices. This implies that if the η_n are assumed to be random, all error terms are heteroskedastic unless parameter restrictions sufficient to collapse the cost frontier to a Cobb–Douglas form are imposed. (2) Since u_A depends on output quantities and input prices, an assumption that overall cost inefficiency ($u_T + u_A$) is iid in a single-equation translog cost frontier model would be inappropriate. (3) The magnitudes of allocative inefficiencies (the η_ns), the impacts of allocative inefficiencies on input cost shares (the $A\eta_n$s), and the cost of allocative inefficiency (u_A) are all identified. It is not necessary to assume a self-dual production frontier, as Schmidt and Lovell did, to identify each of these effects. (4) It is straightforward to show that $u_A = 0 \Leftrightarrow \eta = 0$ and that $u_A > 0$ if $\eta \neq 0$. These parametric restrictions are testable.

Estimation remains a problem. If allocative inefficiency is assumed to be random, and allowed to be both input and producer specific, the translog cost system becomes extremely difficult to estimate. The difficulty arises from the fact that the $A\eta_n$s are highly nonlinear functions of the η_ns, and it is impossible to derive distributions for the $A\eta_n$s starting from a joint distribution for the allocative inefficiency vector η. It is possible, however, to overcome this difficulty by making the restrictive assumption that the magnitudes of allocative inefficiency are input specific but do not vary across producers. The advantage of this assumption, noted by Ferrier and Lovell, is that stochastic noise components ξ_{ni} can be added to the nth input cost share equation for the ith producer. Thus the effects of random noise can be accommodated in the expenditure equation and the input cost share equations, which become

$$\frac{w_n x_n}{E} = S_n(y, w; \beta) + A\eta_n + \xi_n, \qquad n = 1, \ldots, N, \tag{4.2.40}$$

where the $A\eta_n$ are defined in equations (4.2.39). Note that the $A\eta_n$ are functions of data and parameters to be estimated and that the ξ_n are both input and producer specific.

The translog cost system can be estimated using maximum likelihood techniques under the distributional assumptions:

(i) $v \sim$ iid $N(0, \sigma_v^2)$.

(ii) $u_T \sim$ iid $N^+(0, \sigma_u^2)$.

(iii) $\xi \sim$ iid $N(0, \Sigma_\xi)$.

Here $\xi = (\xi_1, \ldots, \xi_N)'$ and Σ_ξ is an $N \times N$ positive semidefinite matrix. Although these distributional assumptions were also made by Ferrier and Lovell, the present model incorporates allocative inefficiency in a more sophisticated fashion than does their model.

Once the parameters are estimated, estimates of the cost of allocative inefficiency can be obtained from u_A in equation (4.2.37). Note that although the magnitudes of allocative inefficiency are assumed to be invariant across producers, estimates of the cost of allocative inefficiency are producer specific. This is because u_A depends on output quantities and input prices. Estimates of u_T, which is also producer specific, can be obtained from estimates of $(v + u_T)$ using the JLMS decomposition.

A less efficient but computationally simpler way of estimating the translog cost system is to use a two-step procedure. In the first step the input cost share equations are estimated using iterated SUR. This procedure does not require the normality assumption on ξ. However the input cost share equations do not contain all the parameters of the translog cost frontier. The remaining parameters, and the parameters associated with the distributions of v and u_T, are estimated in the second step. In the second step distributional assumptions are imposed on v and u_T, and maximum likelihood techniques are used to estimate the remaining parameters.

We conclude this section on a somewhat pessimistic note. It appears that the Greene problem has yet to be satisfactorily resolved. The strategy initiated by Schmidt (1984), and modified by Melfi (1984), Bauer (1985), and Kumbhakar (1991), provides a solution to the problem, but at the cost of assuming that there is no noise in the input share equations. The strategy proposed by Ferrier and Lovell (1990) also provides a solution to the problem, but at the considerable cost of assuming that allocative inefficiency and its cost vary across inputs but not across producers. The strategy proposed by Kopp and Diewert (1982), Zieschang (1983), and Mensah (1994) does

not embed the translog system within a sensible stochastic framework, and requires the numerical solution of a system of nonlinear equations. The procedure developed by Kumbhakar (1997) is analytically elegant and econometrically tractable. At the cost of assuming that the magnitudes of allocative inefficiency are invariant across producers, the model accommodates random noise in the input cost share equations as well as in the cost frontier equation. And although the magnitudes of allocative inefficiency are invariant across producers, their impacts on input cost shares and on expenditure do vary across producers.

4.3 PANEL DATA COST FRONTIER MODELS

The disadvantages of having only cross-sectional data with which to estimate technical efficiency relative to a stochastic production frontier were noted at the beginning of Section 3.3. Each of these disadvantages carries over to the estimation of cost efficiency relative to a stochastic cost frontier in a single-equation model. The fundamental problem is that in a single cross section we get to observe each producer only once, and this severely limits the confidence we have in our (technical or cost) efficiency estimates. In the cross-sectional environment of Section 4.2, data limitations required the imposition of two types of assumptions and still left us with problems:

(i) Maximum likelihood estimation of a stochastic cost frontier, and the subsequent decomposition of the residual into cost inefficiency and statistical noise, both rest on strong distributional assumptions on each error component.
(ii) Maximum likelihood estimation also requires an assumption that the cost inefficiency error component be independent of the regressors – output quantities and input prices, and perhaps quasi-fixed input quantities as well. However it is frequently argued that a principal cause of cost inefficiency is large size; if this is true, it would call into question the validity of the independence assumption.
(iii) The JLMS technique can be applied to the estimation of cost efficiency, but the estimator is not consistent as $I \rightarrow +\infty$.

It is possible to overcome each of these problems if we have access to panel data. The expanded range of estimation procedures discussed in Section 3.3 is equally applicable to the estimation of stochastic cost frontier models. Maximum likelihood estimation remains feasible, and it is a popular option, its distributional and independence assumptions notwithstanding. However fixed-effects and random-effects approaches are also available in a panel data context, and they have certain advantages over maximum likelihood techniques.

Strategies for the estimation of single-equation stochastic cost frontier models in the presence of panel data are virtually identical to those developed for the estimation of stochastic production frontier models. All that is required is to change the sign of the appropriate error component. Consequently Section 4.3.1 is brief, since all essential details appear in Section 3.3. There is no counterpart in Chapter 3 to simultaneous-equation cost frontier models. Consequently Section 4.3.2 is new, although the extension of cross-sectional cost frontier systems to panel data cost frontier systems is straightforward.

4.3.1 Single-Equation Cost Frontier Models

We assume that we have observations on a panel of I producers through T time periods. The panel need not be balanced, although to conserve on notation we shall assume that it is. Also to conserve on notation, we assume that the deterministic kernel of the stochastic cost frontier takes the single-output Cobb–Douglas form. The extension to multiple outputs and to flexible functional forms is straightforward. Finally, we assume initially that cost efficiency is time invariant. Armed with these assumptions, we express the cost frontier model as

$$\ln E_{it} = \beta_o + \beta_y \ln y_{it} + \Sigma_n \beta_n \ln w_{nit} + v_{it} + u_i, \qquad (4.3.1)$$

where v_{it} represents random statistical noise, $u_i \geq 0$ represents time-invariant cost inefficiency, and $\Sigma_n \beta_n = 1$ ensures homogeneity of degree +1 of the cost frontier in input prices.

If we assume that the v_{it} are iid $(0, \sigma_v^2)$ and are uncorrelated with the regressors, none of which is time invariant, and if we make no

distributional or independence assumption on the u_i, equation (4.3.1) can be estimated by means of a fixed-effects approach. Rewriting equation (4.3.1) as

$$\ln E_{it} = \beta_{oi} + \beta_y \ln y_{it} + \Sigma_n \beta_n \ln w_{nit} + v_{it}, \tag{4.3.2}$$

where the $\beta_{oi} = \beta_o + u_i$ are producer-specific intercepts, the model can be estimated by LSDV. After estimation the cost frontier intercept is estimated as $\hat{\beta}_o = \min_i\{\hat{\beta}_{oi}\}$ and the u_i are estimated from $\hat{u}_i = \hat{\beta}_{oi} - \hat{\beta}_o \geq 0$. Finally, producer-specific estimates of cost efficiency are obtained from $CE_i = \exp\{-\hat{u}_i\}$. In the fixed-effects model at least one producer has $CE_i = 1$, and the remaining producers have $CE_i < 1$. Estimates of cost efficiency are consistent as $I \to +\infty$ and $T \to +\infty$.

If instead we assume that the u_i are randomly distributed with constant mean and variance, but are uncorrelated with the v_{it} and with the regressors, and if we assume that the v_{it} have zero expectation and constant variance, we can incorporate time-invariant regressors into the model and use a random-effects approach to estimate equation (4.3.1). Rewriting equation (4.3.1) as

$$\ln E_{it} = \beta_o^* + \beta_y \ln y_{it} + \Sigma_n \beta_n \ln w_{nit} + v_{it} + u_i^*, \tag{4.3.3}$$

where $\beta_o^* = [\beta_o + E(u_i)]$ and $E(u_i^*) = E[u_i - E(u_i)] = 0$, the model can be estimated by GLS. After estimation of equation (4.3.3), an estimate of u_i^* is obtained from the regression residuals by means of $\hat{u}_i^* = (1/T)\Sigma_t(\ln E_{it} - \hat{\beta}_o^* - \hat{\beta}_y \ln y_{it} - \Sigma_n \hat{\beta}_n \ln w_{nit})$, from which we obtain $\hat{u}_i = \hat{u}_i^* - \min_i\{\hat{u}_i^*\} \geq 0$ and $CE_i = \exp\{-\hat{u}_i\}$. These estimates are also consistent as $I \to +\infty$ and $T \to +\infty$.

The GLS estimator hinges on the assumption that the random effects (the u_i) are uncorrelated with the regressors, whereas the LSDV estimator does not. However since GLS allows for the inclusion of time-invariant regressors, it is desirable to test the independence hypothesis. As in Chapter 3, a Hausman–Taylor (1981) test procedure can be applied here. If an independence assumption is warranted, and if one is willing to make distributional assumptions on v and u, maximum likelihood techniques can be employed to estimate the parameters of equation (4.3.1). The appeal of MLE is that it should produce more efficient parameter estimates than either LSDV or GLS, since it exploits distributional information that the other estimators do not. MLE proceeds exactly as in Section 3.3.1 (with many sign changes, since here $\varepsilon = v + u$, whereas there

$\varepsilon = v - u$) and exactly as in Section 4.2.1 (with time subscripts added to producer subscripts where appropriate).

We now sketch the MLE procedure, omitting details. We make the following assumptions on the error components in the stochastic cost frontier model given in equation (4.3.1):

(i) $v_{it} \sim$ iid $N(0, \sigma_v^2)$

(ii) $u_i \sim$ iid $N^+(0, \sigma_u^2)$

(iii) u_i and v_{it} are distributed independently of each other, and of the regressors.

The marginal density function $f(\varepsilon) = f(v + u)$ is the same as the marginal density function $f(\varepsilon) = f(v - u)$ given in equation (3.3.15), apart from a change in sign in the definition of μ_*. The log likelihood function for a sample of I producers, each observed for T periods of time, becomes

$$\ln L = \text{constant} - \frac{I(T-1)}{2}\ln\sigma_v^2 - \frac{I}{2}\ln(\sigma_v^2 + T\sigma_u^2) + \sum_i \ln\left[1 - \Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right)\right] - \left(\frac{\varepsilon'\varepsilon}{2\sigma_v^2}\right) + \frac{1}{2}\sum_i \left(\frac{\mu_{*i}}{\sigma_*}\right)^2, \tag{4.3.4}$$

where $\mu_{*i} = T\sigma_u^2 \bar{\varepsilon}_i/(\sigma_v^2 + T\sigma_u^2)$ and $\sigma_*^2 = \sigma_u^2\sigma_v^2/(\sigma_v^2 + T\sigma_u^2)$. This log likelihood function can be maximized with respect to the parameters to obtain maximum likelihood estimates of β, σ_v^2, and σ_u^2.

The conditional distribution of $(u|\varepsilon)$ is

$$f(u|\varepsilon) = \frac{1}{(2\pi)^{1/2}\sigma_*[1 - \Phi(-\mu_*/\sigma_*)]} \cdot \exp\left\{-\frac{(u-\mu_*)^2}{2\sigma_*^2}\right\}, \tag{4.3.5}$$

which is the density function of a variable distributed as $N^+(\mu_*, \sigma_*^2)$. Either the mean or the mode of this distribution can be used as a point estimator of cost efficiency, and we have

$$E(u_i|\varepsilon_i) = \mu_{*i} + \sigma_*\left[\frac{\phi(-\mu_{*i}/\sigma_*)}{1 - \Phi(-\mu_{*i}/\sigma_*)}\right] \tag{4.3.6}$$

and

$$M(u_i|\varepsilon_i) = \begin{cases} \mu_{*i} & \text{if } \varepsilon_i \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{4.3.7}$$

respectively. These estimators are consistent as $T \to +\infty$. Either can be substituted into $CE_i = \exp\{-u_i\}$ to obtain producer-specific estimates of time-invariant cost efficiency. An alternative estimator is provided by the minimum squared error predictor

$$E(\exp\{-u_i\}|\varepsilon_i) = \frac{1-\Phi[\sigma_* - (\mu_{*i}/\sigma_*)]}{1-\Phi(-\mu_{*i}/\sigma_*)} \cdot \exp\left\{-\mu_{*i} + \frac{1}{2}\sigma_*^2\right\}. \qquad (4.3.8)$$

Confidence intervals for any of these estimators can be constructed exactly as in the cross-sectional maximum likelihood model, with the appropriate changes in notation.

The longer the panel, the less tenable is the assumption that cost efficiency is time invariant. All three estimation procedures, LSDV, GLS, and MLE, can be modified to accommodate time-varying cost efficiency, exactly as they were in Section 3.3.2 in the estimation of technical efficiency. The only changes involve raising instead of lowering intercepts for cost-inefficient producers in the case of LSDV and GLS, and some sign changes in the case of MLE. We leave the details to the reader.

4.3.2 Simultaneous-Equation Cost Frontier Models

As we noted at the outset of Section 4.2.2 in a cross-sectional context, simultaneous-equation cost frontier models offer the possibility of generating more information than single-equation cost frontier models. While single-equation models can provide estimates of the cost of input-oriented inefficiency, simultaneous-equation models can *in principle* provide estimates of the separate costs of technical and allocative inefficiency. However we also noted that a fully satisfactory *econometric* specification of a simultaneous-equation model remains to be developed. We also noted at the outset of Section 4.3 that the advantage of having access to panel data is that we get to observe producers more than once. This enables us to relax restrictive distributional and independence assumptions, and to obtain consistent estimates of cost efficiency. Consequently having access to panel data with which to estimate a simultaneous-equation cost frontier system combines the best of both worlds, and offers the best opportunity to solve the Greene problem.

We consider two models, the single-output Cobb–Douglas model discussed in Section 4.2.2.1 and the multiple-output translog model discussed in Section 4.2.2.2. The Cobb–Douglas model given in equations (4.2.23) is not entirely satisfactory because the input mix equation errors consist entirely of allocative inefficiency; no statistical noise appears in these equations. None of the translog systems is embedded in a stochastic framework that is both econometrically sensible and estimable. Perhaps having access to panel data will help to resolve either or both difficulties.

Cobb–Douglas With panel data covering I producers through T time periods, the Cobb–Douglas system (4.2.23) can be slightly rearranged and written as

$$\ln y_{it} = \beta_o + \sum_n \beta_n \ln x_{nit} + v_{it} - u_i,$$
$$\ln x_{1it} - \ln x_{nit} + \ln\left(\frac{\beta_n}{\beta_1}\right) = \ln\left(\frac{w_{nit}}{w_{1it}}\right) + \eta_{ni} + \xi_{nit}, \qquad n = 2, \ldots, N. \qquad (4.3.9)$$

The essential difference between equations (4.3.9) and equations (4.2.23), apart from the addition of time subscripts, is the appearance of statistical noise, captured by the ξ_{nit}, in the input mix equations contained in equations (4.3.9). With panel data it is possible to separate allocative inefficiency from noise. Notice that both the cost of technical inefficiency (represented by the u_i) and the magnitudes of allocative inefficiency (represented by the η_{ni}) are producer specific but time invariant.

This model can be estimated in two ways, depending on whether the u_i are assumed to be fixed or random effects. If the u_i are assumed to be fixed effects, no distributional assumption is needed, and a fixed-effects approach is feasible. Solving equations (4.3.9) for the natural logarithms of the inputs, we obtain

$$\ln x_{kit} = \alpha_k + \sum_{n>1}\left(\frac{\beta_n}{r} - \delta_{nk}\right)\eta_{ki} + \frac{1}{r}\ln y_{it} + \sum_{n>1}\left(\frac{\beta_n}{r}\right)\ln\left(\frac{w_{nit}}{w_{kit}}\right)$$
$$+ \sum_{n>1}\left(\frac{\beta_n}{r} - \delta_{nk}\right)\xi_{nit} + \frac{1}{r}u_i - \frac{1}{r}v_{it}, \qquad k = 1, \ldots, N, \qquad (4.3.10)$$

where $r = \Sigma_n \beta_n$ and

$$\alpha_k = \ln\beta_k - \frac{1}{r}\left[\beta_o + \sum_n \beta_n \ln\beta_n\right],$$

$$\delta_{nk} = \begin{cases} 1 & \text{if } k = n, \\ 0 & \text{otherwise.} \end{cases}$$

Applying a within transformation to equations (4.3.10) eliminates the time-invariant terms α_k, η_{ki}, and u_i. This allows nonlinear SUR to be applied to the transformed system of input demand equations to obtain estimates of the β_ks. These estimates can then be inserted into equations (4.3.9) to generate estimates of the η_{ni} and u_i by means of

$$\hat{\eta}_{ni} = \overline{\ln\left(\frac{x_{1i}}{x_{ni}}\right)} + \overline{\ln\left(\frac{w_{1i}}{w_{ni}}\right)} + \ln\left(\frac{\hat{\beta}_n}{\hat{\beta}_1}\right),$$

$$\hat{u}_i = \max_i(\bar{e}_i) - \bar{e}_i, \tag{4.3.11}$$

where $\bar{e}_i = (1/T)\Sigma_t(\ln y_{it} - \hat{\beta}_o - \Sigma_n\hat{\beta}_n \ln x_{nit})$ and a bar over a variable indicates the temporal mean of the variable.

If the u_i are assumed to be random effects, we can use MLE to obtain estimates of the parameters of the model. We make the following assumptions:

(i) $v_{it} \sim$ iid $N(0, \sigma_v^2)$.

(ii) $u_i \sim$ iid $N^+(0, \sigma_u^2)$.

(iii) $\xi_{nit} \sim$ iid $N(0, \Sigma)$.

(iv) u_i, v_{it}, and ξ_{nit} are distributed independently of each other, and of the regressors.

With these distributional assumptions, the log likelihood function for equations (4.3.9) is

$$\begin{aligned} \ln L = \text{constant} &- \frac{1}{2}I(T-1)\ln\sigma_v^2 - \frac{I}{2}\ln(\sigma_v^2 + T\sigma_u^2) \\ &- \frac{1}{2\sigma_v^2}\sum_i(\varepsilon_i' A\varepsilon_i) + \sum_i \ln\Phi(a_i) + IT\ln r \\ &- \frac{1}{2}IT\ln|\Sigma| - \frac{1}{2}\sum_i\sum_t z_{it}'\Sigma^{-1}z_{it}, \end{aligned} \tag{4.3.12}$$

where

$$A = I_T - \frac{\sigma_u^2 \iota\iota'}{\sigma_v^2 + T\sigma_u^2},$$

$\iota_{N\times 1} = (1, \ldots, 1)'$,

I_T is a $T \times T$ identity matrix,

$$\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})' = \varpi_i - \iota u_i, \qquad i = 1, \ldots, I,$$

$$a_i = \frac{\sigma_u}{\sigma_v}\frac{\sum_t \varepsilon_{it}}{\sqrt{(\sigma_v^2 + T\sigma_u^2)}}, \qquad i = 1, \ldots, I,$$

$$z_{it} = \begin{bmatrix} \ln x_{1it} - \ln x_{2it} - \ln w_{2it} + \ln w_{1it} - \ln\beta_1 + \ln\beta_2 - \eta_{2i} \\ \vdots \\ \ln x_{1it} - \ln x_{Nit} - \ln w_{Nit} + \ln w_{1it} - \ln\beta_1 + \ln\beta_N - \eta_{Ni} \end{bmatrix}.$$

The computational burden of obtaining parameter estimates can be substantially reduced by concentrating the log likelihood function with respect to η_{ni} and Σ. At the maximum of $\ln L$ we have

$$\eta_{ni} = \overline{\ln x_{1i}} - \overline{\ln x_{ni}} - \overline{\ln w_{ni}} + \overline{\ln w_{1i}} - \ln\beta_1 + \ln\beta_n, \qquad n = 1, \ldots, N,$$

$$\Sigma = \frac{1}{I \cdot T}\sum_i\sum_t z_{it}^* z_{it}^{*\prime},$$

where $z_{it}^* = z_{it} - \iota\bar{z}_i$. Substituting these values back into the log likelihood function yields the concentrated log likelihood function

$$\begin{aligned} \ln L = \text{constant} &- \frac{1}{2}I(T-1)\ln\sigma_v^2 - \frac{I}{2}\ln(\sigma_v^2 + T\sigma_u^2) - \frac{1}{2\sigma_v^2}\sum_i(\varepsilon_i' A\varepsilon_i) \\ &+ \sum_i \ln\Phi(a_i) + IT\ln r - \frac{1}{2}IT\ln\left|\frac{1}{IT}\sum_i\sum_t z_{it}^* z_{it}^{*\prime}\right|, \end{aligned} \tag{4.3.13}$$

which can be maximized to obtain estimates of $\beta_o, \beta_1, \ldots, \beta_N, \sigma_v^2$, and σ_u^2. The simplification of the concentrated log likelihood function results from the fact that the z_{it}^* do not depend on any parameters, a typical element of z_{it}^* having the form $z_{kit}^* = \overline{\ln x_{1it}} - \overline{\ln x_{kit}} + \overline{\ln w_{1it}} - \overline{\ln w_{kit}}$. Consequently the last term in the concentrated log likelihood function is a constant.

The next step is to obtain estimates of technical efficiency. A minor modification of analogous results in Section 3.3.1 can be used to demonstrate that the conditional distribution of $(u_i | \varepsilon_{it})$ is $N^+(\mu_{*i}, \sigma_*^2)$, where

$$\mu_{*i} = \frac{\sigma_u^2}{\sigma_v^2 + T\sigma_u^2} \sum_t \varepsilon_{it},$$

$$\sigma_*^2 = \frac{\sigma_u^2 \sigma_v^2}{\sigma_v^2 + T\sigma_u^2}.$$

Thus either the mean or the mode of $(u_i | \varepsilon_i)$, where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$, can be used as a point estimator of u_i, and we have

$$E(u_i|\varepsilon_{it}) = \sigma_* \left[\frac{\mu_{*i}}{\sigma_*} + \frac{\phi(\mu_{*i}/\sigma_*)}{\Phi(\mu_{*i}/\sigma_*)} \right],$$

$$M(u_i|\varepsilon_{it}) = \begin{cases} \mu_{*i} & \text{if } \sum_t \varepsilon_{it} \geq 0, \\ 0 & \text{otherwise.} \end{cases} \qquad (4.3.14)$$

The final step is to obtain estimates of allocative inefficiency. This requires estimates of the η_{ni}, which are provided in equation (4.3.11).

Translog With panel data covering I producers through T time periods, the translog cost system given in equations (4.2.34) and (4.2.38) can be written as

$$\ln E_{it} = \ln c(y_{it}, w_{it}; \beta) + v_{it} + u_{Ti} + u_{Ait},$$

$$\frac{w_{nit} x_{nit}}{E_{it}} = S_{nit}(y_{it}, w_{it}; \beta) + A\eta_{ni} + \xi_{nit}, \qquad n = 2, \ldots, N, \qquad (4.3.15)$$

where u_{Ait} generalizes the expression in equation (4.2.37) and the $A\eta_{ni}$ generalize the expressions in equation (4.2.39). The essential difference between the translog cost system in equations (4.2.34) and (4.2.38) and the system in equation (4.3.15), apart from the addition of time subscripts, is that in the present panel data formulation the magnitudes of allocative inefficiency (the η_{ni}) are parameters to be estimated, and are both input and producer specific, although they are time invariant. However the impact of allocative inefficiency on input cost shares (the $A\eta_{ni}$) and on cost (u_{Ait}) is time varying, because both expressions depend on time-varying data. This increased flexibility is yet another illustration of the advantage of having access to panel data. It is also possible to introduce some form of time dependence into the magnitudes of allocative inefficiency, for example by specifying $\eta_{nit} = \eta_{ni} \cdot \gamma(t)$, where $\gamma(t)$ is a parametric function of time and the η_{ni} are fixed parameters.

Under the distributional assumptions that $v_{it} \sim$ iid $N(0, \sigma_v^2)$, $u_{Ti} \sim$ iid $N^+(0, \sigma_u^2)$, and $\xi_{it} \sim$ iid $N(0, \Sigma_\xi)$, the log likelihood function can be written as

$$\ln L = \sum_i \ln f(\varepsilon_i) + \sum_i \ln f(\xi_i), \qquad (4.3.16)$$

where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$, $\varepsilon_{it} = v_{it} + u_{Ti}$, and

$$f(\xi_i) = (2\pi)^{-NT/2} \cdot |\Sigma_\xi|^{-T/2} \cdot \exp\left\{ -\frac{1}{2} \sum_t \xi_{it}' \Sigma_\xi^{-1} \xi_{it} \right\},$$

and $\xi_{it} = (\xi_{i1t}, \ldots, \xi_{iNt})'$. Estimation of u_{Ti} for each producer is the same as in equation (4.3.6) or (4.3.7), and estimates of CE_i can be obtained from equation (4.3.8). Finally, estimates of u_{Ait} can be obtained from the estimated values of the parameters in the definition of u_{Ait}.

4.4 TWO ADDITIONAL APPROACHES TO THE ESTIMATION OF COST EFFICIENCY

Two novel approaches to the estimation of cost efficiency have recently been developed. Both can be based on a translog system consisting of a cost equation and its associated input cost share equations. However neither approach attempts to decompose estimated cost inefficiency into its technical and allocative components. The first approach is an admittedly ad hoc approach dubbed "thick frontier analysis," TFA for short. TFA does not require a one-sided error term, and so is not really a frontier approach to the estimation of cost efficiency. However in contrast to some of the more sophisticated approaches discussed in previous sections, TFA is easy to implement, using either cross-sectional data or panel data. The price to be paid for its simplicity is the extreme paucity of information it generates. We consider TFA in Section 4.4.1. The second approach is dubbed a "distribution-free approach," DFA for short, because although it contains a one-sided error term representing cost inefficiency, it imposes no distributional assumptions on it. DFA requires panel data, and is structurally similar to the GLS approach discussed in Section 4.3. We consider DFA in Section 4.4.2.

4.4.1 Thick Frontier Analysis

Thick frontier analysis was developed by Berger and Humphrey (1991, 1992) as a way of avoiding the restrictive assumptions required in conventional approaches to the estimation of cost efficiency. TFA is much less structured than conventional approaches, and so it generates less information, but it always "works." It can be employed within a single cross section or a panel. We discuss TFA within a cross-sectional context.

We begin by assuming that we have observations on total expenditure E_i incurred, a vector $y_i \geq 0$ of outputs produced, and a vector $w_i > 0$ of input prices faced, by a sample of producers indexed $i = 1, \ldots, I$. Next we identify producers located in the top and bottom quartiles (or quintiles or whatever) of the average cost distribution. (If producers produce multiple outputs, average cost can be proxied by total cost divided by the Euclidean norm of the output vector, or by assets or employment.) Producers located in the bottom quartile are presumed to be relatively cost efficient as a group, and together they define a *thick frontier*. Producers located in the top quartile are presumed to be cost inefficient relative to the thick frontier. If differences in average cost are also thought to be scale related, it is possible to control for this influence by stratifying the sample into size classes (if the sample size permits) before forming the quartiles; this ensures that a range of producers is represented within each quartile. Variation in input prices is handled in a different way, described later.

The next step is to estimate separate cost functions (not frontiers) for the top and bottom average cost quartiles. Variation in residuals *within* each quartile is assumed to reflect only random statistical noise, whereas differences in the average level of predicted costs *between* the top and bottom quartiles is assumed to reflect only a combination of exogenous influences and cost inefficiency within the top quartile. While it is unlikely that these assumptions hold exactly, and consequently it is unlikely that TFA yields precise estimates of cost efficiency, the objective of TFA is not econometric rigor, but reliable insight into the probable magnitude of the problem.

Suppose that the cost functions within each quartile have translog form. Then we can estimate the structure of technology within each quartile by estimating the system of equations

$$\begin{aligned} \ln E_i &= \ln c(y_i, w_i; \beta) + v_i, \\ \frac{w_{ni} x_{ni}}{E_i} &= S_{ni}(y_i, w_i; \beta) + v_{ni}, \qquad n = 2, \ldots, N. \end{aligned} \tag{4.4.1}$$

This system is estimated twice, once for each quartile. On the assumption that $[v_i, v_{ni}']' \sim N(0, \Sigma)$, this system may be estimated by SUR, as in Christensen and Greene (1976). Denote the estimated parameter vectors for the first and fourth quartiles β^1 and β^4, and denote the predicted average costs at the mean values of (y_i, w_i) within each quartile by $[c(y, w; \beta^1)/y]^{Q1}$ and $[c(y, w; \beta^4)/y]^{Q4}$, where if y is not a scalar one of the proxies mentioned previously is used.

The difference between the two quartile predicted average costs can be expressed and decomposed as

$$\begin{aligned} &\frac{[c(y,w;\beta^4)/y]^{Q4} - [c(y,w;\beta^1)/y]^{Q1}}{[c(y,w;\beta^1)/y]^{Q1}} \\ &= \frac{[c(y,w;\beta^4)/y]^{Q4} - [c(y,w;\beta^1)/y]^{Q4}}{[c(y,w;\beta^1)/y]^{Q1}} \\ &+ \frac{[c(y,w;\beta^1)/y]^{Q4} - [c(y,w;\beta^1)/y]^{Q1}}{[c(y,w;\beta^1)/y]^{Q1}}. \end{aligned} \tag{4.4.2}$$

The left-hand side of equation (4.4.2) provides an estimate of the percentage difference between the predicted average costs of producers located in quartiles $Q4$ and $Q1$. Since cost functions are estimated separately for the two quartiles, estimated parameter vectors β^4 and β^1 are allowed to differ. Differences between β^4 and β^1, particularly but not exclusively differences between the two intercepts, are intended to reflect differences in cost efficiency between the two quartiles. The first term on the right-hand side provides an estimate of the cost inefficiency of producers located in $Q4$. Here cost inefficiency is estimated as the difference between predicted average cost in $Q4$ using the inefficient $Q4$ technology represented by β^4 and predicted average cost in $Q4$ using the efficient $Q1$ technology represented by β^1, expressed as a percentage of predicted average cost in $Q1$. There is no cost efficiency differential if $[c(y, w; \beta^4)/y]^{Q4} = [c(y, w; \beta^1)/y]^{Q4}$, a sufficient condition for which is $\beta^4 = \beta^1$. The second term on the right-hand side provides an estimate of the average cost difference attributable not to cost inefficiency

but to heterogeneity in the markets in which inefficient and efficient producers operate. Equal cost efficiency is enforced by giving producers in both quartiles the same estimated β^1 technology. The market heterogeneity may be reflected in differences between $(y, w)^{Q4}$ and $(y, w)^{Q1}$.

It is important to emphasize that the decomposition in equation (4.4.2) is applied not to individual producers in either quartile, but to hypothetical mean producers in each quartile. It is in this sense that TFA provides limited information. However since TFA associates cost inefficiency with differences between β^4 and β^1, the following expression can be used to provide an estimate of the average cost inefficiency (ACI_i) of individual producers in $Q4$:

$$ACI_i = \frac{[c(y_i, w_i; \beta^4)/y_i]^{Q4} - [c(y_i, w_i; \beta^1)/y_i]^{Q4}}{[c(y_i, w_i; \beta^1)/y_i]^{Q4}}. \tag{4.4.3}$$

ACI_i indicates the percentage by which average cost is raised using inefficient technology represented by β^4 rather than efficient technology represented by β^1.

The TFA approach has some nice features. Unlike an explicit cost frontier approach, TFA does not require restrictive distributional and independence assumptions on error components; indeed it has no one-sided error component. It is based on an estimable version of the translog cost system with a conventional error structure, and so it is not susceptible to the "Greene problem" that plagues the translog cost frontier system. It does not look for evidence of inefficiency in one-sided error components, which can be difficult to identify; instead it associates inefficiency with differences between easily estimated quartile parameter vectors. Finally, Bauer, Berger, and Humphrey (1993) report evidence based on U.S. banking data suggesting that TFA generates cost efficiency estimates that are similar in magnitude to estimates generated by stochastic frontier techniques.

However TFA has some rather serious shortcomings. First, it is arbitrarily based on average cost quartiles, and estimated cost inefficiency would increase if equally arbitrary quintiles were used instead. Second, it uses only half of the data (or 40% of the data if quintiles are used), and not many researchers are so well endowed

with degrees of freedom that they can discard half of their observations. Most importantly, TFA does not generate cost efficiency estimates for each producer in the sample. It generates only one cost efficiency estimate, for the hypothetical mean producer in the high-cost quartile relative to the hypothetical mean producer in the low-cost quartile. (Of course if the observations are stratified into S size classes, TFA would generate S such cost efficiency estimates.) Thus TFA is likely to be useless to management and of limited value to policy-makers.

4.4.2 A Distribution-Free Approach

The distribution-free approach was introduced by Berger (1993). DFA requires panel data, and is structurally similar to the GLS approach discussed in Section 4.3. It is based on a translog system of cost and input cost share equations, and it generates estimates of cost inefficiency for each producer in each time period.

Suppose we observe a sample of producers indexed $i = 1, \ldots, I$ in each of T time periods indexed $t = 1, \ldots, T$. For each producer we observe total expenditure E_{it}, a vector y_{it} of outputs produced, and a vector w_{it} of input prices paid. Then a translog system consisting of a cost equation and its associated input cost share equations can be written as

$$\begin{aligned} \ln E_{it} &= \ln c(y_{it}, w_{it}; \beta^t) + v_{it} + u_i, \\ \frac{w_{nit} x_{nit}}{E_{it}} &= s_{nit}(y_{it}, w_{it}; \beta^t) + v_{nit}, \qquad n = 2, \ldots, N. \end{aligned} \tag{4.4.4}$$

This system is estimated separately for each time period, and so the technology parameter vector has a time superscript. Within each time period the error vector $[v_{it}, v_{nit}']'$ captures the effects of random statistical noise, and the error component $u_i \geq 0$ measures the cost of producer-specific cost inefficiency. Since $E(v_{nit}) = 0$, allocative efficiency is imposed, and so u_i captures the cost of technical inefficiency only.

The system of equations (4.4.4) is estimated using SUR a total of T times, once for each time period. Thus it is assumed that the u_i are random effects distributed independently of the regressors. For each

producer the cost equation residuals $\hat{\varepsilon}_{it} = \hat{v}_{it} + \hat{u}_i$ are averaged over time to obtain $\hat{\varepsilon}_i = (1/T)\Sigma_t \hat{\varepsilon}_{it}$. On the assumption that the random-noise error component v_{it} should tend to average zero over time, $\hat{\varepsilon}_i = (1/T)\Sigma_t \hat{\varepsilon}_{it} \cong \hat{u}_i$ provides an estimate of the cost inefficiency error component. To ensure that estimated cost inefficiency is nonnegative, $\hat{\varepsilon}_i$ is normalized on the smallest value, and we obtain

$$\widehat{CE}_i = \exp\{-[\hat{\varepsilon}_i - \min_i(\hat{\varepsilon}_i)]\}. \tag{4.4.5}$$

This estimator is similar to the GLS panel data estimator in which u_i is treated as a random effect, and this similarity suggests that it is appropriate when I is large relative to T and when the u_i are orthogonal to the regressors. However it differs from GLS in that the structure of the underlying production technology is allowed to vary through time. Berger also noted that since the elements of v_{it} may not fully cancel out through time for each producer, $\hat{\varepsilon}_i$ may contain elements of luck as well as inefficiency. To alleviate this problem, he recommended truncating the distribution of CE_i at its qth and $(1 - q)$th quantiles.

A disadvantage of DFA is the requirement that cost efficiency be time invariant, and this assumption becomes less tenable as T increases. However DFA also has two distinct virtues. First, being based on a sequence of T separate cross-sectional regressions, it allows the structure of production technology to vary flexibly through time (although excessive variation in $\hat{\beta}^t$ would be difficult to explain). Second, it does not impose a distributional assumption on the u_i; it lets the data reveal the empirical distribution of the $\hat{\varepsilon}_i \cong \hat{u}_i$. Although ε_i is truncated, it need not follow any of the specific distributions we have considered for u_i when discussing the MLE approach.

Using three large ($T = 10$, $I \cong 1{,}000$) panels of U.S. banks, Berger examined the empirical distributions of $\hat{\varepsilon}_i \cong \hat{u}_i$. Only two of three distributions exhibited the positive skewness one would expect, and all three skewness coefficients were numerically small (0.36, 0.36, –0.21). Histograms of the three distributions appear approximately normal, and Shapiro–Wilks variance ratio test statistics could not reject the normality hypothesis for each distribution.

An important consideration in the DFA approach concerns the length of the panel. If T is "small" the random-noise terms v_{it} may not average zero, and substantial amounts of random noise will appear in the cost inefficiency error component u_i. On the other hand,

if T is "large" the time-invariant assumption on u_i is likely to be violated. This suggests that there may exist an optimal value of T on which to base the DFA approach.

DeYoung (1997) has developed a diagnostic test for determining the optimal panel length. The procedure begins by expressing ε_i as $\varepsilon_i(T)$ to indicate that it is based on T time periods and writing it in longhand as

$$\varepsilon_i(T) = \frac{1}{T}[(u_{i1} + \cdots + u_{iT}) + (v_{i1} + \cdots + v_{iT})], \tag{4.4.6}$$

where u_i is now allowed to be time varying and to follow the time path

$$u_{it} = \begin{cases} u_{i1} & \text{for } t \le S, \\ u_{it-1} + \tau_i \cdot u_{i1} & \text{for } t > S, \end{cases} \tag{4.4.7}$$

where $\tau_i \cdot u_{i1}$ is the annual "drift" in cost inefficiency for the ith producer and $\tau_i \in [-1, +1]$. Substituting equation (4.4.7) into equation (4.4.6) yields

$$\varepsilon_i(T) = u_{i1} + \frac{1}{2T} \cdot (\tau_i \cdot u_{i1}) \cdot \max\{0, |T - S| \cdot (T - S + 1)\} + \frac{1}{T}\sum_t v_{it}. \tag{4.4.8}$$

The first term on the right-hand side of equation (4.4.8) is the initial level of cost inefficiency for the ith producer. The second term is the average annual accumulated drift in cost inefficiency from $t = 1$ to $t = T$. The third term is the mean of the random-noise error component. The general strategy is to select a value of T that is short enough to limit the distortions caused by the temporal drift in u_{it} and long enough to minimize the mean random-noise error component.

For a given value of T, define the cross-sectional mean and variance of $\varepsilon_i(T)$ as

$$\mu(T) = \frac{1}{N}\sum_i \varepsilon_i(T),$$
$$\sigma^2(T) = \frac{1}{N-1}\sum_i [\varepsilon_i(T) - \mu(T)]^2. \tag{4.4.9}$$

Substituting equation (4.4.8) and the expression for $\mu(T)$ into the expression for $\sigma^2(T)$ yields the expression

$$\sigma^2(T)=\frac{1}{N-1}\sum_i\left[\left(u_{i1}-\frac{1}{N}\sum_i u_{i1}\right)+\frac{1}{2T}\max\{0,|T-S|\cdot(T-S+1)\}\right.$$
$$\left.\times\left(\tau_i\cdot u_{i1}-\frac{1}{N}\sum_i\tau_i\cdot u_{i1}\right)+\frac{1}{T}\left(\sum_t v_{it}-\frac{1}{N}\sum_i\sum_t v_{it}\right)\right]^2. \tag{4.4.10}$$

Interest centers on how the absolute values of the three terms inside the square brackets change as the value of T increases. The first term (the initial level of cost inefficiency) does not vary with T. The second term (the average annual accumulated drift in cost inefficiency) equals zero from $T = 1$ to $T = S$. For $T > S$ this term increases in absolute value as T increases. The absolute value of the third term (random noise) decreases as T increases. Thus until $T = S$ the magnitude of $\sigma^2(T)$ declines as T increases, and as a result the estimates of cost inefficiency $\varepsilon_i(T)$ will approach the true values u_i. However for $T > S$ the magnitude of $\sigma^2(T)$ either increases, remains constant, or decreases with increases in T, depending on whether the marginal reduction in random noise exceeds, equals, or falls short of the marginal drift in cost inefficiency. The optimal value of T is therefore defined as the first value of T for which $\sigma^2(T)$ stops decreasing. DeYoung conducted a test of the hypothesis that $\sigma^2(T)$ has stopped decreasing by performing a series of F tests of the hypothesis that $\sigma^2(T)/\sigma^2(T+1) = 1$, although he noted that the test is not strictly valid because the numerator and denominator are not independent draws from the same population.

4.5 A GUIDE TO THE LITERATURE

The analytical foundations for the definition and decomposition of cost efficiency were laid by Farrell (1957), and much of the material in this chapter derives ultimately from Farrell's insights. Early research into the estimation of cost functions is summarized by Johnston (1960). Other pioneering research includes that of Nerlove (1963), who followed Shephard (1953) by allowing for homotheticity of the dual production technology and by imposing linear homogeneity in input prices on the Cobb–Douglas cost function he estimated, and Christensen and Greene (1976), who were perhaps the first to estimate a flexible translog cost and input cost share equation system.

The estimation of stochastic cost frontiers began with Schmidt and Lovell (1979, 1980), although a deterministic cost frontier was previously estimated by Førsund and Jansen (1977). Other early studies include those of Greene (1980b) and Stevenson (1980), who introduced two-parameter distributions for the one-sided error component. Progress during the past two decades is surveyed by Schmidt (1985–1986), Bauer (1990b), Greene (1993, 1997), and Cornwell and Schmidt (1996).

New developments and applications appear regularly in the *Journal of Econometrics*, the *Journal of Productivity Analysis*, and a wide variety of field journals. One development worthy of mention concerns the use of flexible functional forms other than translog to model the deterministic kernel of a stochastic cost frontier. The Fourier functional form contains the translog form as a special case, and has been used to model a cost frontier by Berger and Mester (1997), and in several other studies they cite.