EFFICIENCY ANALYSIS: A PRIMER ON RECENT ADVANCES

CHRISTOPHER F. PARMETER AND SUBAL C. KUMBHAKAR

Abstract. This paper reviews the econometric literature on the estimation of stochastic frontiers and technical efficiency. Special attention is devoted to current research.

Contents

University of Miami, State University of New York at Binghamton
Date: November 5, 2014.
Key words and phrases. Stochastic Frontier, Nonparametric, Latent Class, Sample Selection, Panel Data, Closed Skew Normal.
Subal Kumbhakar, Corresponding Author, Department of Economics, State University of New York at Binghamton; e-mail: kkar@binghamton.edu. Christopher F. Parmeter, Department of Economics, University of Miami; e-mail: cparmeter@bus.miami.edu. This monograph is an extension of our lecture notes for several short courses presented at Aalto University, the University of Stavanger, and Wageningen University in 2013. The organizers of those workshops are warmly acknowledged. All errors are ours alone.

1. Overview

At its core inefficiency is a nebulous concept. Førsund & Hjalmarsson (1974, p. 152) note that it is an easy term to use but much more difficult to precisely pin down its meaning. A precise definition is lacking mainly given that those who conform to the strict boundaries of price theory believe that output shortfall and rapid growth are related to pricing information and profit maximization; concepts defined as inefficiency can be construed as managerial goals which encapsulate maximizing behavior. In an outstanding review of the development of (or argument over) X-inefficiency, Perelman (2011) notes that both Leibenstein (1966) and Stigler (1976) fail to provide convincing evidence that the other is wrong. That is, while Leibenstein (1966) only provides anecdotal evidence on the existence of firm inefficiency and Stigler (1976) provides cursory discussion demonstrating an alternative view, neither can resoundingly reject the other's views.

And yet, the study of firm inefficiency persists to this day primarily because even though a formal, robust theory which details how inefficiency operates does not exist, many are unsatisfied with the strict optimizing restrictions placed on firms. Further, myriad evidence of productivity differences exists across firms that *ex ante* are close to homogenous and standard views on productivity differences are not applicable. For example, Syverson (2011) finds that within U.S. manufacturing industries at the 4 digit SIC the 90th percentile plant within the productivity distribution produces nearly double the output of the 10th percentile plant with the same inputs. Moving offshore, Hsieh & Klenow (2009) find productivity differences at a ratio of 5 to 1 in both India and China.

Lest concerns over geographical differences, workforce characteristics and the like drive these differences, consider the study of Chew, Clark & Bresnahan (1990) of a large commercial food operation in the U.S.. Chew et al.'s (1990) example is instructive since these plants should be able to transfer knowledge extensively and share best practices easily. Yet, this network was characterized by the almost complete void of knowledge transfer and large differences in productivity. In fact, within this division of the firm, there are over 40 operating units each of which produce a near identical set of outputs with almost all work done manually and free of international influences.

The stark reality of this division is that even with all the advantages of operating multiple units and sharing best practice, the most productive unit produces almost three times as much output for the same amount of inputs as the least productive unit. Chew et al. (1990) recognize that underlying differences could be driving these differences and control for geographic location, the size of the local market that is served, unemployment, unionization, equipment, quality, and local monopoly power. Even after accounting for these differences in what are considered relatively homogenous firms, productivity differences on the order of 2:1 still are pervasive; a clear signal that firm inefficiency is at play.

Our objective here is not to develop a formal theory or definition of inefficiency. Rather, we seek to detail the important econometric area of efficiency estimation; both past approaches as well as new methodology. Beginning with the seminal work of Farrell (1957), myriad approaches to discerning output shortfall have been developed. Amongst the proposed approaches, two main

camps have emerged. Those that estimate maximal output and attribute all departures from this as inefficiency, known as Data Envelopment Analysis (DEA) and those that allow for both unobserved variation in output do to shocks and measurement error as well as inefficiency, known as Stochastic Frontier Analysis (SFA).

Our review here will focus exclusively on SFA. For an exceptionally authoritative review of DEA methods and their statistical underpinnings see Simar & Wilson (2013).[1]The econometric study of efficiency analysis typically begins by constructing a convoluted error term that is composed on noise, shocks, measurement error and a one-sided shock called inefficiency. Early in the development of these methods attention focused on the proposal of distributional assumptions which yielded a likelihood function whereby the parameters of the distributional components of the convoluted error could be recovered. The field evolved to the study of individual specific efficiency scores and the extension of these methods to panel data. Recently, attention has focused on relaxing the stringent distributional assumptions that are commonly imposed, relaxing the functional form assumptions commonly placed on the underlying technology, or some combination of both. All told, exciting and seminal breakthroughs have occurred in this literature on regular bases and reviews of these methods are needed to effectively detail the state of the art.

To explain the generality of SFA we go back to neoclassical production theory. The textbook definition of a production function is: given the input vector $\boldsymbol{x}_i$ for a producer i, the production function $m(\boldsymbol{x}_i;\boldsymbol{\beta})$ is defined by the maximum possible output that can be produced. That is, $m(\boldsymbol{x}_i;\boldsymbol{\beta})$ is the technical maximum (potential). To emphasize on the word maximum we call $m(\boldsymbol{x}_i;\boldsymbol{\beta})$ the frontier production function. Not every producer can produce the maximum possible output, even if $\boldsymbol{x}$ were exactly the same for all of them. Thus, $y_i \leq m(\boldsymbol{x}_i;\boldsymbol{\beta})$ and the ratio $y/m(\boldsymbol{x}_i;\boldsymbol{\beta}) \leq 1$ is defined as technical efficiency ($0 \leq TE \leq 1$), when y is the actual output produced. Quite often we define technical inefficiency ($TI = 1 - TE$) as percentage shortfall of output from its maximum, given the inputs. Thus, $TI = (m(\boldsymbol{x}_i;\boldsymbol{\beta})) - y)/m(\boldsymbol{x}_i;\boldsymbol{\beta}) \geq 0$. This is important when the inequality $y \leq m(\boldsymbol{x}_i;\boldsymbol{\beta})$ is expressed as $\ln y_i = \ln m(\boldsymbol{x}_i;\boldsymbol{\beta}) - u_i$ and $u_i \geq 0$ is interpreted as technical inefficiency.[2]

The above definition of inefficiency fits into the theory in which the role of unforeseen/uncontrollable factors is ignored. However, in reality, randomness, for obvious reasons, is a part and parcel of econometric models. And there are innumerable uncontrollable factors that affect output, given the controllable inputs $\boldsymbol{x}_i$. To accommodate this randomness (v_i), we specify the production frontier as a stochastic relationship and write it as $\ln y_i = \ln m(\boldsymbol{x}_i;\boldsymbol{\beta}) - u_i + v_i$.

The generality of SFA is such that the study of efficiency has gone beyond simple application of frontier methods to study firms and appears across a diverse set of applied milieus. Thus, we also hope that this review will be of appeal to those outside of the efficiency literature seeking to

[1]For comprehensive book length treatments on SFA we suggest one consult Kumbhakar & Lovell (2000) or Kumbhakar, Wang & Horncastle (2014).

[2]Strictly speaking $-u_i \leq 0 \approx \ln TE$ is technical inefficiency.

learn about new methods which might assist them in uncovering phenomena in their applied area of interest.

2. The Benchmark Stochastic Production Frontier Model

The stochastic production frontier was first proposed independently by Aigner, Lovell & Schmidt (1977) and Meeusen & van den Broeck (1977).[3] Their general setup is

$$y_i = m(\boldsymbol{x}_i;\boldsymbol{\beta}) - u_i + v_i = m(\boldsymbol{x}_i;\boldsymbol{\beta}) + \varepsilon_i \tag{2.1}$$

where the key difference from a standard production function is the appearance of the two distinct error terms in the model. The u_i terms captures inefficiency, shortfall from maximal output dictated by the production function, $m(\boldsymbol{x}_i;\boldsymbol{\beta})$, while the v_i terms captures outside influences beyond the control of the producer. At the time this model was proposed this insight was novel. Earlier approaches to model inefficiency typically ignored v_i and this led to solutions to the model that did not have proper statistical properties.[4] The standard neoclassical production function which assumes full efficiency becomes a special case of this SF model when $u_i = 0$ which can tested by hypothesizing that the variance of $u_i = 0$. Thus, SF production model is an extension of the neoclassical production model.

The specification of inefficiency in (2.2) is purposely vague about where the inefficiency is coming from. Is it from input slacks, i.e., inputs not being used fully effectively? For example, workers may not work with their full effort and capital may be improperly used. Similarly, the technology may not be properly used (workers operating machines may not be properly trained although they may working with their full effort). If the source is the inputs, the resulting inefficiency is often labeled as input-oriented inefficiency so that inputs used are worth less than their full potential (inputs in effective units are less than what is actually used and observed). On the other hand, if inputs are effectively used but output is still less than its maximum, inefficiency is often labeled as output-oriented. Since the end result is output shortfall, whatever the sources are, most of the time inefficiency is modeled as output-oriented. We discuss these issues in greater details in later sections.

We will begin our discussion assuming that inputs are exogenously given, in which case we have $u_i \perp \boldsymbol{x}$ and $v_i \perp \boldsymbol{x}\ \forall \boldsymbol{x}$. However, given that u_i leads directly to a shortfall in output, it only reduces output and as such it stems from a one-sided distribution. This implies that even with exogenous inputs, $E[\varepsilon_i|\boldsymbol{x}] \neq 0$. This lack of 0 conditional mean only impacts estimation of the intercept of the production function however. Letting $\zeta = E[u]$, we have

$$y_i = m(\boldsymbol{x}_i;\boldsymbol{\beta}) - \zeta - (u_i - \zeta) + v_i \equiv m^*(\boldsymbol{x}_i;\boldsymbol{\beta}) + \varepsilon_i^* \tag{2.2}$$

and $E[\varepsilon_i^*|x] = 0$. Thus, if we were only interested in recovering the production technology, then the presence of inefficiency would not impact standard estimation, i.e., the ordinary least squares (OLS) could be used once a specification was made for $m(\boldsymbol{x}_i;\boldsymbol{\beta})$.

[3] Outside of these two papers, Battese & Corra (1977) were the first to apply these methods.

[4] For example, see the work of Aigner & Chu (1968), Timmer (1971), Afriat (1972), Dugger (1974), Richmond (1974), Schmidt (1976) and Greene (1980*a*).

Rarely is the focus exclusively on the production technology. More commonly the econometrician has the joint aim of estimating the production technology and recovering information pertaining to inefficiency. With this in mind, more structure is required on the problem. Both Aigner et al.'s (1977) and Meeusen & van den Broeck's (1977) solution was to make distributional assumptions on u_i and v_i (which induces a distribution for ε_i and estimate all of the parameters of the model via maximum likelihood. Both papers assumed that v_i was distributed standard normal with mean 0 and variance σ_v^2. However, Aigner et al. (1977) assumed that u_i was generated from a half normal distribution, $N^+(0, \sigma_u^2)$, whereas Meeusen & van den Broeck (1977) assumed u_i was distributed exponentially, with parameter σ_u.

While these distributions appear quite different, they share several interesting aspects. First, the densities have modes at $u_i = 0$ and monotonically decay as u_i increases. Thus, both distributions assume that the probability of being grossly inefficient is small. Second, both distributions are governed by a single parameter, thus, the mean and variance of u_i are influenced by the same parameter. Third, the skewness of both distributions is independent of the parameter, with the skewness of the exponential distribution being 2 and the half normal distribution having skewness that is approximately 1. Figure 1 plots out the half normal and exponential distributions with identical variance of 1. Even with the same variance, the two densities look drastically different. It should be apparent that the choice of distributional assumption is important and should not be overlooked.

2.1. **Determining the Distribution of ε.** To estimate (2.2) via maximum likelihood, the density of ε must be determined. Once distributional assumptions on v and u have been made, determination of $f(\varepsilon)$ can be determined by noting that the joint density of u and v, $f(u, v)$, can be written as the product of the individual densities $f(u)f(v)$ given the independence of u and v. Further, since $v = \varepsilon + u$, $f(u, \varepsilon) = f(u)f(\varepsilon + u)$. u can be integrated out to obtain $f(\varepsilon)$. We note here that not all distributional assumptions will provide closed form solutions for $f(\varepsilon)$. With either the half normal specification of Aigner et al. (1977) or the exponential specification of Meeusen & van den Broeck (1977), $f(\varepsilon)$ possesses an (approximately) closed form solution, making direct application of maximum likelihood straightforward. For these two distributional assumptions we have

$$f(\varepsilon) = \frac{2}{\sigma}\phi(\varepsilon/\sigma)\Phi(-\varepsilon\lambda/\sigma), \quad \text{(Normal – Half Normal)} \tag{2.3}$$

$$f(\varepsilon) = \frac{1}{\sigma_u}\Phi(-\varepsilon/\sigma_v - \sigma_v/\sigma_u)e^{\varepsilon/\sigma_u + \sigma_v^2/2\sigma_u^2}, \quad \text{(Normal – Exponential)} \tag{2.4}$$

where $\phi(\cdot)$ is the standard normal probability density function, $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$ and $\lambda = \sigma_u/\sigma_v$. The parameterization in (2.3) is quite common and has intuitive appeal. λ can be thought of as a measure of the signal to noise, the amount of variation in ε due to inefficiency versus that which is due to noise. As $\sigma_u \to \infty$, $\lambda \to \infty$ whereas as $\sigma_u \to 0$, $\lambda \to 0$. This measure is only suggestive however. Bear in mind that under the assumption that u is distributed half normal, σ_u^2 is not the variance of inefficiency (that would be

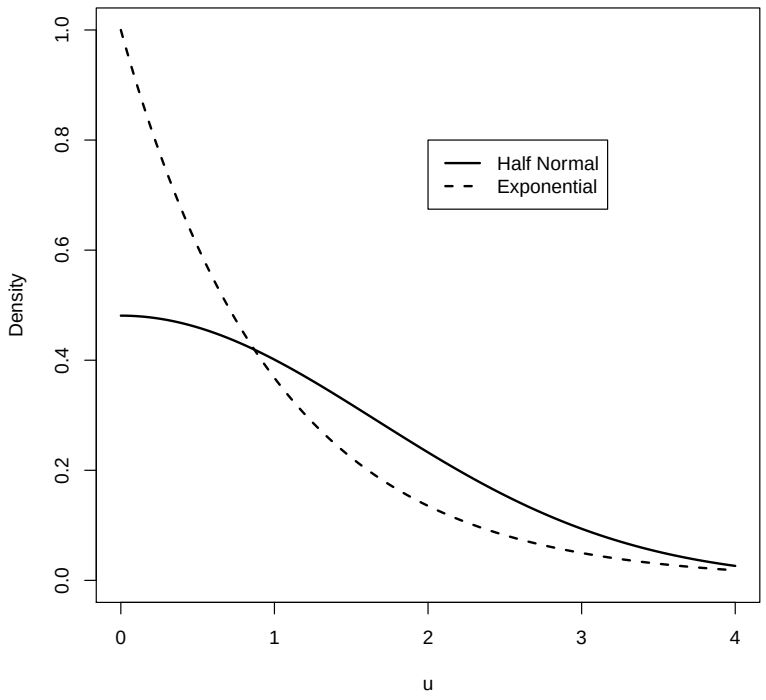

FIGURE 1. Half Normal and Exponential densities both with variance equal to 1

$(1 - 2/\pi)\sigma_u^2$) and the actual signal to noise ratio would be

$$\frac{(1 - 2/\pi)\sigma_u^2}{(1 - 2/\pi)\sigma_u^2 + \sigma_v^2}. \tag{2.5}$$

A consequence of describing σ_u^2 as the variance of inefficiency is that the researcher will overstate the variation of inefficiency by almost 3 ($1/(1 - 2/\pi) \approx 2.75$).

Figure 2 plots out the density of $f(\varepsilon)$ for both the half normal and exponential distributional assumptions with identical variance of 1, and assuming that v is standard normal. Contrary to the depiction in Figure 1, there the shape of the convolved density, $f(\varepsilon)$ looks similar across the two different distribution specifications for u.

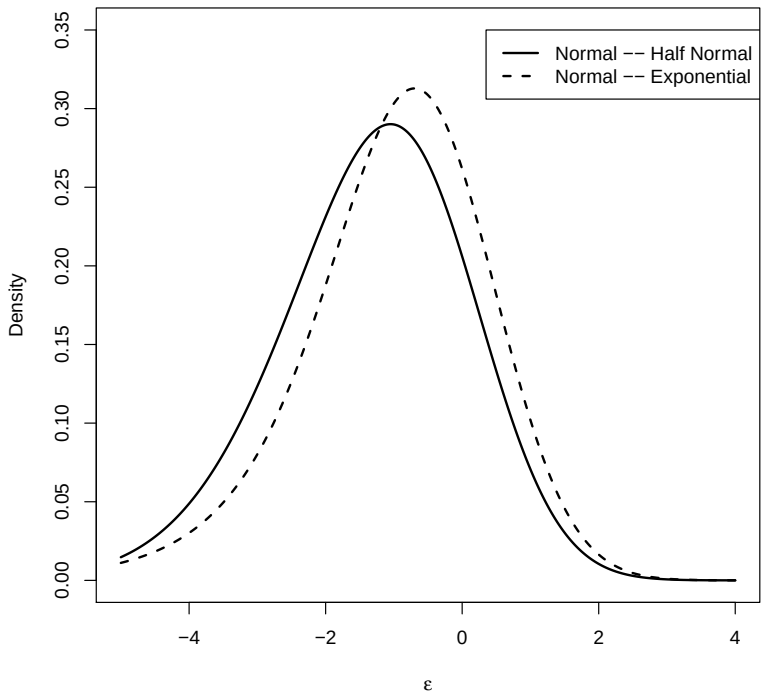

FIGURE 2. Density of $f(\varepsilon)$ when u is distributed as Half Normal or Exponential with variance equal to 1

The corresponding log-likelihood equations can be determined from (2.3) and (2.4) noting that the likelihood is defined as $\mathcal{L} = \prod_{i=1}^{n} f(\varepsilon_i)$, where $\varepsilon_i = y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta})$ yielding

$$\ln \mathcal{L} = -n \ln \sigma + \sum_{i=1}^{n} \ln \Phi(-\varepsilon_i \lambda / \sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \varepsilon_i^2 \tag{2.6}$$

$$\ln \mathcal{L} = -n \ln \sigma_u + n \left(\frac{\sigma_v^2}{2\sigma_u^2} \right) + \sum_{i=1}^{n} \ln \Phi(-\varepsilon_i / \sigma_v - \sigma_v / \sigma_u) + \frac{1}{\sigma_u} \sum_{i=1}^{n} \varepsilon_i \tag{2.7}$$

A close look at (2.3) shows that the pdf ε is nothing but that of a skew normal random variable with location parameter 0, scale parameter σ and skew parameter $-\lambda$. The probability density function of a skew normal random variable x is $f(x) = 2\phi(x)\Phi(\alpha x)$ (O'Hagan & Leonard 1976, Azzalini 1985). The distribution is right skewed if $\alpha > 0$ and is left skewed if $\alpha < 0$. We can also place the normal, truncated normal pair of distributional assumptions in this class. The probability density function of x with location ξ, scale ω, and skew parameter α is $f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right)$. Statistical properties (such as moment, cumulant generating functions, and others) of the skew normal distribution is derived in Azzalini (1985). This connection

has only recently appeared in the efficiency and productivity literature. See Section 7 for more discussion on the skew normal distribution.

2.2. **Alternative Specifications.** While the half normal assumption for the one-sided inefficiency term is undoubtedly the most commonly used in empirical studies of inefficiency, a variety of alternative stochastic frontier models have been proposed using alternative distributions on the one-sided term. Most notably, Stevenson (1980) proposed a generalization of the half normal distribution, the truncated (at 0) normal distribution. The truncated normal distribution depends on two parameters (μ and σ_u^2) and affords the researcher more flexibility in the shape of the distribution of inefficiency. Formally, the truncated normal distribution is

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma_u \Phi(\mu/\sigma_u)} e^{-\frac{(u-\mu)^2}{2\sigma_u^2}}. \tag{2.8}$$

When $\mu = 0$ this reduces to the half normal distribution. Thus, this distributional specification can nest the less flexible distributional assumption and offers avenues for inference on the shape of the density of inefficiency. One aspect of using the truncated normal distribution in practice is the implications it presents regarding inefficiency for the industry as a whole. That is, unlike the half normal and exponential densities, the truncated normal density has mode at 0 only when $\mu \leq 0$, but otherwise has model at μ. Thus, for $\mu > 0$, the implication is that in general producers in the industry are inefficient. This is not necessarily a criticism of using the truncated normal distribution as the specification for inefficiency, more that one needs to make sure they understand what a given distributional assumption translates to in real economic terms.

The density of ε for the normal truncated normal specification is

$$f(\varepsilon) = \frac{1}{\sigma}\phi\left(\frac{\varepsilon+\mu}{\sigma}\right)\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\varepsilon\lambda}{\sigma}\right)\Big/\Phi(\mu/\sigma_u). \tag{2.9}$$

The corresponding log-likelihood function is

$$\ln \mathcal{L} = -n\ln\sigma - \sum_{i=1}^{n}\left(\frac{\varepsilon_i+\mu}{\sigma}\right)^2 - n\ln\Phi(\mu/\sigma_u) + \sum_{i=1}^{n}\ln\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\varepsilon_i\lambda}{\sigma}\right). \tag{2.10}$$

Beyond the truncated normal specification for the distribution of u, a variety of alternatives have been proposed. Greene (1980*a*, 1980*b*) and Stevenson (1980) both proposed a gamma distribution for inefficiency. The gamma density takes the form

$$f(u) = \frac{\sigma_u^{-P}}{\Gamma(P)} u^{P-1} e^{-u/\sigma_u}, \tag{2.11}$$

where the gamma function is defined as $\Gamma(P) = \int_0^\infty t^{P-1}e^{-t}dt$ and when P is an integer $\Gamma(P) = (P-1)!$. When $P = 0$, the gamma density is equivalent to the exponential density. As with the truncated normal distribution, there are two parameters which govern the shape of the density and the exponential density can be tested. Stevenson (1980) only considered a special set of gamma distributions, namely those that have the Erlang form (where P is an integer). This will produce a

tractable formulation for $f(\varepsilon)$. However, for non-integer values of P more care is required. Beckers & Hammond (1987) formally derived the log-likelihood function for $f(\varepsilon)$ without restricting P to be an integer. Greene (1990) followed this by demonstrating that the normal gamma likelihood function can be written as the sum of the normal exponential likelihood plus several additional terms. Notably, one of these additional terms is the fractional moment of a truncated normal random deviate. In general this fractional moment will not have a closed form solution and so direct integration methods are needed when deploying maximum likelihood. Given that the likelihood needs to be evaluated numerically and the possibility for approximation error is high, Ritter & Simar (1997) advocate for using the normal gamma specification in practice. Further, they note that large samples are required to reliably estimate P, which has a larger impact on the shape of the density of inefficiency than σ_u. Greene (2003) developed a more general approach for application of the normal gamma specification based on simulated maximum likelihood which made evaluation of the likelihood simpler. This avoided one of the main critiques of Ritter & Simar (1997) thus, still keeping open the possibility of use of the normal gamma specification in empirical work.

Lee (1983) proposed a four parameter Pearson density for the specification of inefficiency. This four parameter specification generalized many of the proposed inefficiency distributions and thus allowed testing of the correct shape of the distribution of inefficiency. Unfortunately, this distribution is intractable for applied work and until now has not appeared to gain popularity. Other attempts to develop the stochastic frontier model by changing the distribution of inefficiency have appeared. Li (1996) proposed uniform distribution for inefficiency. The impact of the assumption of a uniform density on the density of the composed error was that this density could be positively skewed (which for the previous distributions discussed $f(\varepsilon)$ is always negatively skewed. This was an interesting insight and one that we will return when we discuss implications of the skewness of ε on identification. Further, the uniform assumption can be seen as beginning an efficiency analysis as agnostic given that no shape is imposed on the distribution of inefficiency. A somewhat odd specification for inefficiency appears in Carree (2002). Inefficiency is assumed to have a binomial distribution. In this setting Carree (2002) does not derive the density of $f(\varepsilon)$ nor the log likelihood function. Rather, a methods of moments approach is presented to recover the unknown distributional parameters based off of the OLS residuals. For inefficiency defined as a percentage (scaled between 0 and 1), Gagnepain & Ivaldi (2002) specify inefficiency as being Beta distributed. The Beta distribution does not impose strong restrictions on the shape of the distribution of inefficiency; for certain parameterizations the distribution is symmetric and can be hump or U-shaped. Another alternative was recently proposed by Almanidis, Qian & Sickles (2014), specifying inefficiency as a doubly truncated normal distribution. While we have that the one-sided nature of inefficiency leads to truncation at 0 (for either the half normal or the truncated normal), Almanidis et al. (2014) also truncate the distribution of inefficiency from above, essentially limiting the magnitude of grossly inefficient firms. Their specification provides a closed form solution for $f(\varepsilon)$ and the log-likelihood.

A common feature of the previous papers is that they focus attention exclusively on the distribution of inefficiency. A small literature has shed light on the features of $f(\varepsilon)$ when both the density

of v and the density of u are changed. Specifically, Horrace & Parmeter (2014) study the behavior of the composed error when v is distributed as Laplace and u is distributed as truncated Laplace. Nguyen (2010) considers the Laplace-Exponential distributional pair as well as the Cauchy-Half Cauchy pair for the two error terms of the composed error.[5] These alternative distributional pairs provide different insights into the behavior of inefficiency as well as the properties of the composed error. However, given the relative nascence of these methods more work is required to see if they withstand empirical scrutiny.

2.2.1. *Estimation via Maximum Simulated Likelihood.* A nice feature of nearly all of the proposed distributional assumptions on u is that they yield tractable (almost) closed form solutions for the likelihood that needs to be optimized. However, as we will discuss later on, many of the newer models that are appearing in the literature, specifically those either focusing on sample selection (Section 6) or panel data (Section 7) do not yield tractable likelihoods as uni- or multivariate integrals must be solved. This can make optimization difficult. An alternative route is to use maximum simulated likelihood (MSL) estimation (McFadden 1989).

The key to implementing the MSL estimator is to recognize that if u were known, then, the conditional distribution of y on $\boldsymbol{x}$ *and* u would be

$$f(y|\boldsymbol{x},u) = f(v) = \frac{1}{\sigma_v\sqrt{2\pi}}e^{-0.5\left(\frac{y-\boldsymbol{x}'\boldsymbol{\beta}+u}{\sigma_v}\right)^2}, \tag{2.12}$$

which is nothing more than the normal density. Naturally, u is not known and so we must account for its appearance by conditioning it out of the model through integration, which yields

$$f(y|\boldsymbol{x}) = \int_0^\infty \frac{1}{\sigma_v\sqrt{2\pi}}e^{-0.5\left(\frac{y-\boldsymbol{x}'\boldsymbol{\beta}+u}{\sigma_v}\right)^2} f(u)du, \tag{2.13}$$

where $f(u) = N^+(0,\sigma_u) = \frac{\sqrt{2}}{\sigma_u\sqrt{\pi}}e^{-0.5u^2/\sigma_u^2}$ for $u > 0$. Our likelihood function in this case is just

$$\ln\mathcal{L} = \sum_{i=1}^n \ln f(y_i|\boldsymbol{x}_i).$$

Notice that the integral in (2.13) can be viewed as an expectation, which we can evaluate through simulation as opposed to analytically. If u is distributed as half normal with parameter σ_u^2, then we can generate random deviates which follow this distribution simply by drawing standard random normal deviates, U, taking the absolute value and multiplying by σ_u. That is, $u^* = \sigma_u|U|$. Taking many draws, the integral in (2.13) can be approximated as

$$f(y|\boldsymbol{x}) \approx R^{-1}\sum_{r=1}^R \frac{1}{\sigma_v\sqrt{2\pi}}e^{-0.5\left(\frac{y-\boldsymbol{x}'\boldsymbol{\beta}+\sigma_u|U_r|}{\sigma_v}\right)^2}. \tag{2.14}$$

[5]Nguyen (2010) also considers the Normal-Uniform pair, but as mentioned, this was first discussed in Li (1996).

The simulated log likelihood function is then

$$\ln \mathcal{L}_s = \sum_{i=1}^{n} \ln \left(\sum_{r=1}^{R} \frac{1}{\sigma_v \sqrt{2\pi}} e^{-0.5\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma_u |U_{ir}|}{\sigma_v} \right)^2} \right), \tag{2.15}$$

which can be optimized just as easily as the analytically likelihood function in (2.6). This procedure would work equivalently if u was distributed as exponential or Gamma or truncated normal, provided that in the approximation of $f(y|\boldsymbol{x})$ one generated the random deviates appropriately.

2.3. **Modified Ordinary Least Squares Estimation.** Even though the stochastic frontier model is relatively straightforward to estimate once $m(\boldsymbol{x}_i;\beta)$ has been specified, it still requires nonlinear optimization methods. An alternative approach to recover estimates of β and the parameters of the distributions of v and u is to use modified ordinary least squares (MOLS) which was first proposed by Afriat (1972) and Richmond (1974). Given that MOLS type approaches will be deployed later, we discuss its implementation here. Essentially, MOLS amounts to ignoring the structure of ε and recovering $\boldsymbol{\beta}$ by regressing y on $\boldsymbol{x}$. Doing so results in biased estimation of the intercept of the production function as $\widehat{\beta}_0 = \beta_0 - \mu$. The second stage in MOLS uses the maintained distributional assumptions on v and u to produce a set of moment conditions which can be solved for the unknown distributional parameters; in the normal half normal model these would be σ_v^2 and σ_u^2. Note that the OLS residuals, $\widehat{\varepsilon}$ will have mean zero even though in truth $E[\varepsilon] \neq 0$. This is by default.

From the distributional assumptions we have

$$E[u] = \sqrt{2/\pi}\sigma_u; \quad Var(u) = [(\pi - 2)/\pi]\sigma_u^2; \quad E[u^3] = -E[u](1 - 4/\pi)\sigma_u^2.$$

Given that the third moment of v is 0 under the assumption of normality (actually all we need here is to assume symmetry) we can use the third central moment of the OLS residuals, $\widehat{\gamma}_3$ to recover σ_u^2 as

$$\widehat{\sigma}_u^2 = \left(\frac{\widehat{\gamma}_3}{\sqrt{2/\pi}(1 - 4/\pi)} \right)^{2/3}. \tag{2.16}$$

The $-$ sign for $E[u^3]$ does not appear given that the third central moment of $\varepsilon = v - u$ already accounts for this. With an estimate of σ_u^2, we can estimate σ_v^2 from the second central moment of the OLS residuals, $\widehat{\gamma}_2$, given that $Var(\varepsilon) = \sigma_v^2 + [(\pi - 2)/\pi]\sigma_u^2$. This yields

$$\widehat{\sigma}_v^2 = \widehat{\gamma}_2 - [(\pi - 2)/\pi]\widehat{\sigma}_u^2. \tag{2.17}$$

The last step in the MOLS procedure is to correct the intercept. The MOLS estimate of the intercept, $\widehat{\beta}_0^{MOLS}$ is $\widehat{\beta}_0 + \sqrt{2/\pi}\widehat{\sigma}_u$.

The MOLS procedure is simpler to implement than full maximum likelihood estimation, though as pointed out by Olson, Schmidt & Waldman (1980), if $\widehat{\gamma}_3$ is positive (which can happen in practice) then one obtains negative estimates of σ_u^2, which cannot occur; however, this is not a failure only of MOLS, Waldman (1982) demonstrates that when $\widehat{\gamma}_3 > 0$ that maximum likelihood of the stochastic frontier model will produce an estimator of σ_u^2 of 0. A further complication is when $\widehat{\sigma}_u^2$

is so large that this produces negative estimates of $\widehat{\sigma}_v^2$, which also cannot occur. Thus, the MOLS procedure, while straightforward to implement without access to a nonlinear optimization routine, still has issues which can lead to implausible estimates. Another issue with the MOLS approach is that while it does produce consistent estimates of the parameters of the model, these estimators are inefficient relative to the maximum likelihood estimator, which attains the Cramer-Rao lower bound. The benefit of MOLS is that the first stage estimates, aside from $\widehat{\beta}_0$ are robust to distributional assumptions as the OLS estimator is semiparametrically efficient when no distributional assumptions are imposed Chamberlain (1987).

2.4. **Estimation of Inefficiency.** After the model parameters are estimated, we can proceed to estimate observation-specific efficiency, which is one of the main interests of a stochastic frontier model. The estimated efficiency levels can be used to rank producers, identify under-performing producers, and determine firms using best practices; this information is, in turn, useful in helping to design public policy or subsidy programs aimed at improving the overall efficiency level of private and public sectors, for example.

As a concrete illustration, consider firms operating electricity distribution networks who typically possess a natural local monopoly given that the construction of competing networks over the same terrain is prohibitively expensive.[6] It is not uncommon for national governments to establish regulatory agencies which monitor the provision of electricity to ensure that abuse of the inherent monopoly power is not occurring. Regulators face the task of determining an acceptable price for the provision of electricity while having to counter balance the heterogeneity the exists across the firms (in terms of size of the firm and length of the network). Firms which are inefficient may charge too high a price to recoup a profit, but at the expense of operating below capacity. However, given production and distribution shocks, not all departures from the frontier represent inefficiency. Thus, measures designed to account for the noise are required to parse information from ε_i regarding u_i.

Alternatively, further investigation could reveal what it is that makes these establishments attain such high levels of performance. This could then be used to identify appropriate government policy implications and responses or identify processes and/or management practices that should be spread (or encouraged) across the less efficient, but otherwise similar, units. More directly, efficiency rankings are used in regulated industries such that regulators can set the more inefficient companies tougher future cost reduction targets, in order to ensure that customers do not pay for inefficiency.

Currently, we have only discussed estimation of σ_u^2, which provides information regarding the shape of the half-normal distribution on u_i. This information is all we need if the interest is in the average level of technical inefficiency in the sample. This measure is known as the unconditional mean of u_i. However, if interest lies in the level of inefficiency for a given firm, knowledge of σ_u^2 is not enough as it does not contain any individual-specific information.

[6]An example of this type of study is Kuosmanen (2012) which identified best practices of electricity providers in Finland.

The primary solution, first proposed by Jondrow, Lovell, Materov & Schmidt (1982), is to estimate u_i from the expected value of u_i conditional on the composed error of the model, $\varepsilon_i \equiv v_i - u_i$. This conditional mean of u_i given ε_i gives a point estimate of u_i. The composed error contains individual-specific information, and so the conditional expectation yields the observation-specific value of the inefficiency. This is like extracting signal from noise.

Jondrow et al. (1982) show that the conditional density function of u_i given ε_i, $f(u_i|\varepsilon_i)$, is $N^+(\mu_{*i}, \sigma_*^2)$, where

$$\mu_{*i} = \frac{-\varepsilon_i \sigma_u^2}{\sigma^2} \tag{2.18}$$

and

$$\sigma_*^2 = \frac{\sigma_v^2 \sigma_u^2}{\sigma^2}. \tag{2.19}$$

From this, the conditional mean is shown to be:

$$E(u_i|\varepsilon_i) = \mu_{*i} + \frac{\sigma_* \phi(\frac{\mu_{*i}}{\sigma_*})}{\Phi\left(\frac{\mu_{*i}}{\sigma_*}\right)}. \tag{2.20}$$

Maximum likelihood estimates of the parameters are substituted into the equation to obtain estimates of firm level inefficiency. This estimator will produce values that are guaranteed to be non-negative.

Jondrow et al. (1982) also suggested an alternative to the conditional mean estimator, viz., the conditional mode:

$$M(u_i|\varepsilon_i) = \begin{cases} \mu_{*i} & \text{if } \mu_{*i} > 0, \\ 0 & \text{if} \mu_{*i} \leq 0. \end{cases}$$

This modal estimator can be viewed as the maximum likelihood estimator of u_i given ε_i. Note that ε_i is not known and we are replacing it by the estimated value from the model. Since by construction some of the ε_i will be positive (and therefore $\mu_{*i} < 0$) there will be some observations that are fully efficient (i.e., $M(u_i|\varepsilon_i) = 0$). In contrast, none of the observations will be fully efficient if one uses the conditional mean estimator ($E(u_i|\varepsilon_i)$). Consequently, average inefficiency for a sample of firms will be lower if one uses the modal estimator.

Since the conditional distribution of u is known, one can derive moments of any continuous function of $u|\varepsilon$. That is, we can use the same technique to obtain observation-specific estimates of the efficiency index, e^{-u_i}. Battese & Coelli (1988) show that

$$E\left[e^{-u_i}|\varepsilon_i\right] = e^{\left(-\mu_{*i} + \frac{1}{2}\sigma_*^2\right)} \frac{\Phi\left(\frac{\mu_{*i}}{\sigma_*} - \sigma_*\right)}{\Phi\left(\frac{\mu_{*i}}{\sigma_*}\right)}, \tag{2.21}$$

where μ_{*i} and σ_* are defined in (2.18) and (2.19). Maximum likelihood estimates of the parameters are substituted into the equation to obtain point estimates of e^{-u_i}. This estimaor is bounded

between 0 and 1, with a value of 1 indicating a fully efficient firm. Similar expressions for the Jondrow et al. (1982) and Battese & Coelli (1988) efficiency scores can be derived under the assumption that u is exponential (Kumbhakar & Lovell 2000, p. 82), truncated normal (Kumbhakar & Lovell 2000, p. 86), and Gamma (Kumbhakar & Lovell 2000, p. 89).

2.4.1. *Inference about the Distribution of Inefficiency.* The JLMS efficiency estimator is inconsistent as $n \to \infty$. This is not surprising for two reasons. First, in a cross-section, as $n \to \infty$ we have new firms being added to the sample with their own level of inefficiency instead of new observations to help determine a given firms specific level of inefficiency. Second, the JLMS efficiency estimator is not designed to estimate unconditional inefficiency, it is designed to estimate inefficiency conditional on ε, for which it is a consistent estimator. Moreover, the JLMS inefficiency estimator is known as a shrinkage estimator; on average, we overstate the inefficiency level of a firm with small u_i while we understate inefficiency for a firm with large u_i.

Wang & Schmidt (2009) recently studied the distribution of the JLMS inefficiency scores and showed that unless $\sigma_v \to 0$, the distribution of $\widehat{E}(u_i|\varepsilon_i)$ differs from the distribution of u_i. Moreover, an interesting finding from their analysis is that as $\sigma_v^2 \uparrow$, the distribution of $\widehat{E}(u_i|\varepsilon_i)$ effectively converges to $E(u)$; that is, as the variation in v increases, ε contains no useful information to predict inefficiency through the conditional mean.

Wang & Schmidt's (2009) theoretical findings have no impact on conducting inference for $\widehat{E}(u_i|\varepsilon_i)$, rather, their insights pertain primarily to inference on the assumed distribution of u_i. The message is that it is not valid to simply compare the observed distribution of $\widehat{E}(u_i|\varepsilon_i)$ to the assumed distribution for u_i. Doing so will result in misleading insights regarding the appropriateness of a specific distributional assumption given that these two distributions are only the same when $\sigma_v^2 = 0$. Rather, to test the distributional assumptions of the stochastic frontier model, one needs to compare the distribution of $\widehat{E}(u_i|\varepsilon_i)$ to $E(u_i|\varepsilon_i)$. Wang & Schmidt (2009) provide this distribution for the normal half normal stochastic frontier model and Wang, Amsler & Schmidt (2011) propose χ^2 and Komolgorov-Smirnov type test statistics against this distribution.[7] A key point is to note that for a given test, a rejection does not necessarily imply the distributional assumption on u is incorrect, it could be that the normality distributional assumption on v is violated and this is leading to the rejection. One must be careful in interpreting tests on the distribution of ε (or functionals of ε).

2.4.2. *Prediction of Inefficiency.* Unlike the inferential procedures described above regarding the appropriate specification of the distribution of inefficiency, inference regarding a specific level of inefficiency for a firm does not exist in the literature. In fact, there is some debate on the interpretation of constructed confidence intervals. Here we discuss the surrounding issues.

The prediction interval of $E(u_i|\varepsilon_i)$ is derived by Taube (1988), Hjalmarsson, Kumbhakar & Heshmati (1996), Horrace & Schmidt (1996), and Bera & Sharma (1999) based on $f(u_i|\varepsilon_i)$. The

[7]See also Schmidt & Lin (1984).

formulas for the lower bound (L_i) and the upper bound (U_i) of a $(1-\alpha)100\%$ confidence interval are

$$L_i = \mu_{*i} + \Phi^{-1}\left\{1 - \left(1 - \frac{\alpha}{2}\right)\left[1 - \Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right)\right]\right\}\sigma_*, \tag{2.22}$$

$$U_i = \mu_{*i} + \Phi^{-1}\left\{1 - \frac{\alpha}{2}\left[1 - \Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right)\right]\right\}\sigma_*, \tag{2.23}$$

where μ_{*i} and σ_* are defined in (2.18) and (2.19). The lower and upper bounds of a $(1-\alpha)100\%$ confidence interval of $E(e^{-u_i}|\varepsilon_i)$ are, respectively,

$$\mathbb{L}_i = e^{-U_i}, \tag{2.24}$$

$$\mathbb{U}_i = e^{-L_i}. \tag{2.25}$$

The results follow because of the monotonicity of e^{-u_i} as a function of u_i. Recently, Wheat, Greene & Smith (2014) provided minimum width prediction intervals. They noted that the interval provided by Horrace & Schmidt (1996) are based on central two sided intervals; however, given that the distribution of u_i conditional on ε_i is truncated (at 0) normal and thus asymmetric, this form of interval is not minimum width. By solving a Lagrangian for minimizing the width of a $1-\alpha$ prediction interval, Wheat et al. (2014) are able to show that the minimum prediction interval for u_i given ε_i is[8]

$$L_i^* = \mu_{*i} + \sigma_*\Phi^{-1}\left[\left(\frac{\alpha}{2}\right)\left(1 - \Phi\left(\frac{\mu_{*i}}{\sigma_*}\right)\right)\right] \tag{2.26}$$

$$U_i^* = \mu_{*i} + \sigma_*\Phi^{-1}\left[\left(1 - \frac{\alpha}{2}\right)\left(1 - \Phi\left(\frac{\mu_{*i}}{\sigma_*}\right)\right)\right] \tag{2.27}$$

when both L_i^* and U_i^* are greater than 0, otherwise

$$L_i^* = 0 \tag{2.28}$$

$$U_i^* = \mu_{*i} + \sigma_*\Phi^{-1}\left[1 - \alpha\Phi\left(\frac{\mu_{*i}}{\sigma_*}\right)\right]. \tag{2.29}$$

Unlike the central two-sided prediction intervals, a simple monotonic transformation of the minimum width intervals (to predict technical efficiency say) will not necessarily result in a corresponding minimum width interval; rather, numerical methods are needed. The reason for this is that the percentiles that provide the minimum width interval for one distribution are not necessarily those that provide minimum width bounds for a monotone transformation of the original distribution

To understand how much narrower the prediction interval of Wheat et al. (2014) is relative to that of Horrace & Schmidt (1996), consider the setting where $\sigma_u = \sigma_v = 1$. In this case the relative width of the two intervals ranges from 1 (equal width) to about 1.2 (nearly 20% wider). Clearly, for different parameter combinations the relative width will change, but the point remains that if the goal is to accurately predict firm level inefficiency then a narrower interval is preferred.

[8]We mention here that in Wheat et al. (2014), they define $\mu_{*i} = \frac{\varepsilon_i\sigma_u^2}{\sigma^2}$, however, this is a typo as the $-$ sign is missing on the error term. See (2.18) for the correct definition.

It should be noted that the construction of the above prediction intervals assumes that the model parameters are known, while in actuality they are unknown and must be estimated. The above confidence intervals do not take into account this parameter uncertainty. Alternatively, we may bootstrap the confidence interval (Simar & Wilson 2007, Simar & Wilson 2010) or use resampling based on the asymptotic distribution of the parameters (Wheat et al. 2014) to take into account estimation uncertainty.

An alternative to construction of a prediction interval for a point estimate of inefficiency is to instead compare if different firms are identical with respect to their inefficiency. Using a technique known as multiple comparisons with the best, Horrace & Schmidt (2000) develop the technology to perform just such a calculation. The useful aspect of this approach is that rather than derive bounds for the magnitude of a single firm's level of inefficiency, the practitioner can instead determine if some (or all) firms are statistically equal with respect to inefficiency. Further, multiple comparisons with the best allows statements such as"firm i is statistically indistinguishable from the firm with the highest (lowest) estimated level of inefficiency in the sample." This is useful when offering policy prescriptions.

Recently, Horrace (2005) demonstrated that the mean of the conditional distribution[9] is not a fully informative estimate of technical inefficiency. This holds since the mean is only one characterization of many for the distribution. An alternative is to calculate the probability that a firm is fully efficient based on the underlying conditional distribution of all firms in the sample. These probabilities can then be used to identify a firm which has high probability of being efficient.

One shortcoming of Horrace (2005) is that the procedure can only identify a single firm that is efficient. However, in many industries, a single efficient firm is unlikely. Consequently, it may be possible Horrace's (2005) approach to yield no inference on a single firm. To remedy this, Flores-Lagunes, Horrace & Schnier (2007) extend the methodology of Horrace (2005) to allow for multiple firms which are efficient by constructing non-empty subsets of minimal cardinality of firms with high probability of being efficient. The non-empty feature ensures that inference can always be performed. Flores-Lagunes et al.'s (2007) use of a non-empty subset keeps fidelity with the ranking research of Horrace & Schmidt (2000).

2.5. **Do Distributional Assumptions Matter?** A key question with the benchmark stochastic frontier model is the importance of the distributional assumptions on v and u. The distribution of v has almost universally been accepted as being normal in both applied and theoretical work, a recent exception being the Laplace distribution analyzed in Horrace & Parmeter (2014). While more work has been devoted to understanding the implications from alternative shapes for the distribution of u, little practical work has been undertaken to examine the extent of this assumption. Most applied papers do not rigorously check differences in estimates and inference across different distributional assumptions and few papers that engage in Monte Carlo analysis check for the impact of misspecification in the distributional assumption imposed on u. Greene (1990) is an oft mentioned

[9]In Horrace (2005) the focus is exclusively on a truncated normal distribution, but the argument holds more generally

analysis that compared average inefficiency levels across the four main distributional specifications for u (half normal, truncated normal, exponential, and gamma) and found almost no difference in average inefficiency for 123 U. S. electric utility providers. Kumbhakar & Lovell (2000) calculated the rank correlations amongst the JLMS scores from these four models. Their analysis produced rank correlations as low as 0.75 and as high as 0.98. In a small Monte Carlo analysis, Ruggiero (1999) compared rank correlations of stochastic frontier estimates assuming that inefficiency was either half normal (which was the true distribution) or exponential (a misspecified distribution) and found very little evidence that misspecification impacted the rank correlations in any meaningful fashion.

The intuition underlying these findings is that for nearly all of the proposed distributions which dominate applied work (half normal, truncated normal, exponential, gamma), the JLMS efficiency scores are monotonic in ε (Ondrich & Ruggiero 2001, p. 438) provided that the distribution of v is log-concave (which the normal distribution is). The implication here is that unless one is confident in the distributional assumptions on v and u and firm level estimates of inefficiency are required, firm rankings can be obtained via the OLS residuals without resorting to maximum likelihood analysis (Bera & Sharma 1999). Depending on how much variability there is in the estimates of the production function across different distributional assumptions it is likely that the firm rankings will be highly correlated. Thus, if interest hinges on features of the frontier, then so long as inefficiency does not depend on conditional variables, one can effectively ignore the distribution as this only affects the level of the estimated technology, but not its shape which is what influences measures such as returns to scale and elasticities of substitution.

Regarding rankings of firms, the Laplace exponential specification of Horrace & Parmeter (2014) produces an interesting result that use of a normal exponential specification cannot produce: for those observations with positive residuals, the JLMS scores are identical. That is, any observation with a positive residual, regardless of magnitude will have the same JLMS score under the assumption of a Laplace exponential convolution.

All told, if there is specific interest in firm level inefficiency then distributional assumptions are an important component of the estimation of a stochastic frontier model. However, if interest hinges on features of the production technology or on simple rankings of firms, then it is likely that OLS will be sufficient for these purposes. Further, if one does wish to deploy distributional assumptions, it may prove wise to follow the advice of Ritter & Simar (1997) and use simple one-parameter families for the distribution of u.

Our discussion so far has focused on research specifically looking at the econometric issues surrounding different distributional assumptions. Several empirical papers have looked at the sensitivity of predicted technical efficiency across a range of distributions. For example, in their study of Tunisian manufacturing firms Baccouche & Kouki (2003) explore differences across the half normal, truncated normal and exponential distributions[10] and find that their estimates of technical efficiency depend heavily on the assumed distribution. No rank correlations are provided however,

[10]They also deploy the generalized half normal distribution.

so it is not entirely surprising that their direct estimates differ. They favor using more flexible distributions for u, such as the truncated normal, given that this distribution does not directly imposes that the level of inefficiency is monotonically decreasing.

2.6. **The Importance of Skewness.** With the basic stochastic frontier model in tow, along with an estimator for firm level inefficiency, one is ready to confront data. Yet, a vexing empirical (and theoretical) issue that often arises is that the maximum likelihood estimator will return an estimate of σ_u^2 of 0, essentially indicating the lack of inefficiency in the data. While this may stand in contrast to the original intent of the efficiency analysis, there is in fact a quite logical explanation. For the basic stochastic frontier model, let the parameter vector be $\theta = (\boldsymbol{\beta}, \lambda, \sigma^2)$. Then Waldman (1982) established the following results. First, the log likelihood always has a stationary point at $\theta^* = (\hat{\boldsymbol{\beta}}_{OLS}, 0, \widehat{\sigma}^2)$, where $\hat{\boldsymbol{\beta}}_{OLS}$ is the parameter vector one obtains using OLS, i.e., ignoring the presence of inefficiency, and $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$, where $\widehat{\varepsilon}_i = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{OLS}$ are the OLS residuals. Note that these parameter values correspond to $\sigma_u^2 = 0$, that is, to full efficiency of each firm. Second, the Hessian matrix is singular at this point. It is negative semi-definite with one zero eigenvalue. Third, these parameter values are a local maximizer of the log likelihood if the OLS residuals are positively skewed. This is the so-called "wrong skew problem" (for a lucid discussion see Almanidis & Sickles 2011). When the OLS residuals have skewness of the wrong sign relative to the stochastic frontier model, maximum likelihood estimation will almost always be equivalent to the OLS estimates.

To illustrate the wrong skew consider Figure 3. Here we plot maximum likelihood estimates of σ_u against the skewness of the OLS residuals. We set $n = 100$, $\sigma_v = 1$ and select $\sigma_u = 0.5$ so that $\lambda = 0.5$, ensuring a healthy proportion of our samples will have the wrong skew. We use a single covariate (generated as standard normal) along with an intercept, with parameters $\beta_0 = \beta_1 = 1$. Using 1000 Monte Carlo simulations we see that there is almost an exact relationship between the OLS residuals' skewness and the maximum likelihood estimator of σ_u, at least for small, but negative, skewness.

It is important to note that having the skewness of the OLS residuals be of the wrong sign, and consequently a maximum likelihood estimate of σ_u^2 of 0, does not entail that anything is wrong with the stochastic frontier model. This is simply a fact that the normal half normal maximum likelihood function's identification of σ_u^2 is based entirely on the skewness of the composed residuals. Why would one expect σ_u^2 to be positive when v is assumed to be symmetric and we have positive skewness?

Further explication on this point is found in Simar & Wilson (2010). There, a series of simulation exercises are run for the sole purpose of determining how often the skewness of a draw of a given sample size from a known normal half normal convolution is of the wrong sign. Given the one-sidedness of u, it is known that the skewness of ε is negative under the assumption that v is symmetric. However, sampling variation can lead to estimated skewness of the wrong sign. For example, Simar & Wilson's (2010, Table 1) simulations revealed that for $\lambda = 1$ and $n = 200$ the

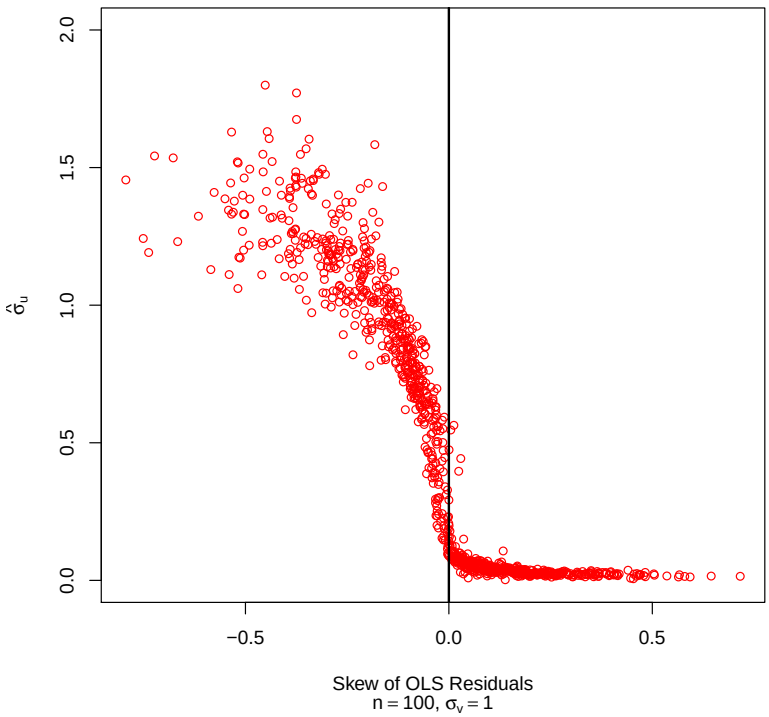

FIGURE 3. Maximum Likelihood Estimator of σ_u^2 in Normal Half Normal Stochastic Frontier Model versus Skewness of OLS Residuals. 1000 Monte Carlo Simulations, $\lambda = 0.5$.

wrong skewness occurs in almost 23% of the simulations. That is, in almost one out of four trials, a draw of 200 resulted in the wrong skew and, following Waldman (1982), would have led to an estimate of σ_u^2 of 0.

Figure 4 presents the proportion of draws with the wrong skew from the normal half normal convolution for four different plausible values of λ in empirical work across a range of plausible empirical sample sizes. Two things are immediate. First, as n increases, regardless of the value of λ, the proportion of samples with the wrong skew decreases. Second, the rate of decrease depends on λ. For $\lambda = 1.4$ we have the appearance of improperly signed skewness samples decays to zero rapidly, effectively disappear once $n > 500$. However, for $\lambda = 0.3$ at $n = 500$ we still have almost 49% of the samples with the wrong skew.

However, even with the fact that a correctly specified stochastic frontier model can produce wrongly signed skewness, the earlier literature recommended some form of potential model misspecification might be at issue. Kumbhakar & Lovell (2000, p. 92) mention that using the skewness of the OLS residuals as a useful diagnostic for model misspecification. However, we mention here that respecification based purely on the sign of the OLS residuals is improper. More formally, one

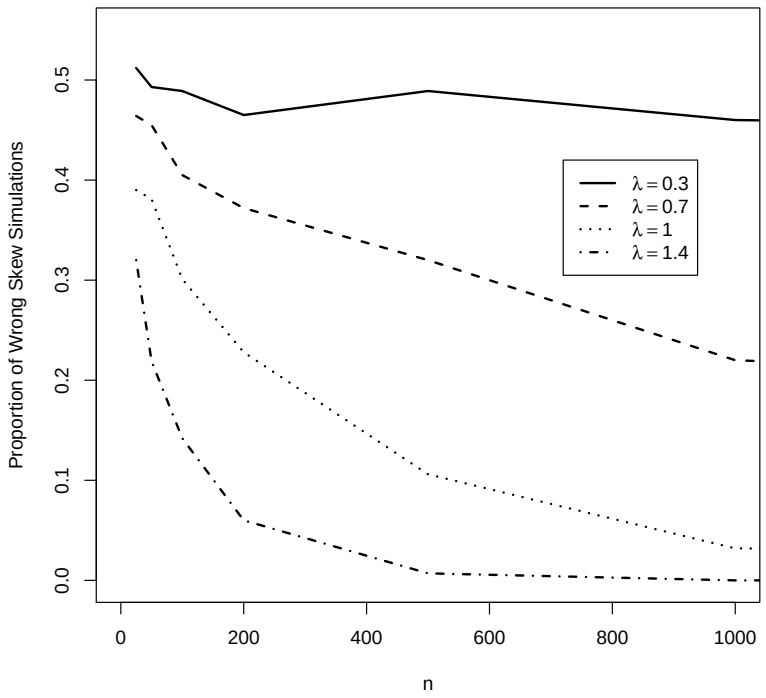

FIGURE 4. Proportion of Normal Half Normal Convolutions with Wrong Skewness

should test the OLS residuals for the sign of the skewness as the results of Simar & Wilson (2010) show that there is nothing inconsistent with the presence of inefficiency **and** OLS residuals of the wrong skew.

Given the dependence of the maximum likelihood estimator for σ_u^2 on the skewness of the OLS residuals, various alternatives have been proposed to construct estimators which deliver non zero estimates of σ_u^2 in the presence of wrong skew. For example, Li (1996), Carree (2002), and Almanidis et al. (2014) all provide a set of distributional assumptions that allow for inefficiency and a convoluted error term that can be positively skewed (all three approaches assume the inefficiency distribution is bounded from above), effectively decoupling skewness and identification of σ_u^2 (here our use of σ_u^2 is as a generic place holder for the variance of inefficiency without specifically assuming what the distribution of u is). Alternatively, Feng, Horrace & Wu (2013) have suggested using constrained optimization methods to enforce the restriction that $\sigma_u^2 > 0$ in the normal half normal stochastic frontier model. Unfortunately, this method currently requires *ad hoc* selection of a tolerance parameter which directly impacts the size of σ_u^2 when the skewness of the OLS residuals is negative, which is exactly when the restriction binds.

A more recent generalization of the stochastic frontier model appears in Hafner, Manner & Simar (2013). They suggest replacing $v - u$ with $v - u + 2E[u]1\{E[u] < 0\}$. This mean correction to the

standard convoluted error effectively decouples the sign and magnitude of $E[\varepsilon]$ from $E[u]$. That is, regardless if $E[u] \lessgtr 0$, $E[\varepsilon] = -\Big|E[u]\Big| < 0$. The impact of this simple correction is that the skewness of ε can be either positive or negative without placing bounds on the distribution of inefficiency. Thus, estimation of the parameter governing the presence of inefficiency and skewness are no longer directly linked, and one can have, even asymptotically, inefficiency and positive skewness, which is not possible in the classical normal half normal stochastic frontier model. Moreover, in Hafner et al.'s (2013) framework, the stochastic frontier model is identical to the standard stochastic frontier model when $E[u] > 0$ and so their model is a generalization. By allowing the distribution of u to have a positive mean enough flexibility is added in that identification issues related to the skewness of the OLS residuals has been dispensed with. Naturally, the cost of using this mean correction is that in small samples a bias will be introduced. A similar decoupling appears in Horrace & Parmeter (2014) who show that when the two-sided error distribution is changed from normal to Laplace that identification of the inefficiency parameter is no longer based on the skewness of the residuals.

We again stress that wrong skewness does not necessarily imply that some form of model misspecification exists. If this was a concern then proper specification testing could be undertaken, on the skewness of the OLS residuals[11] for example. Alternatively, as noted by Simar & Wilson (2010), one can still engage in proper inference on both the inefficiency parameter, σ_u as well as the JLMS scores even when the wrong skewness appears in one's empirical analysis. They propose a bootstrap aggregating (known as bagging) algorithm to construct valid confidence intervals for the JLMS scores.

[11]See Kuosmanen & Fosgerau (2009) for a test on the sign of the skewness of the OLS residuals.

3. Accounting for Multiple Outputs in the Stochastic Frontier Model

Our discussion so far has centered around the stochastic production frontier in which a single output is produced with multiple inputs. However, many production processes produce more than one output which are often aggregated into one. This may not be a good idea. In this section we consider SF models that can handle multiple outputs in a primal framework in which price information is not required. Typically a distance function formulation is used for this. Here we use a transformation function formulation because it is easier to explain without going through a whole lot of technicalities. Furthermore, under appropriate restrictions one can recover the distance function formulation starting from the transformation function.

To make the presentation more general, we start from a transformation function formulation and extend it to accommodate both input and output oriented technical inefficiency, viz., $AT(\theta\boldsymbol{x}, \lambda\boldsymbol{y}) = 1$ where $\boldsymbol{x}$ is a vector of J inputs, $\boldsymbol{y}$ is a vector of M outputs, and the A term captures the impact of observed and unobserved factors that affect the transformation function neutrally. Input-oriented (IO) technical inefficiency is indicated by $\theta \leq 1$ and output-oriented (OO) technical inefficiency is captured by $\lambda \geq 1$ (both are scalars). Thus, $\theta\boldsymbol{x} \leq \boldsymbol{x}$ is the input vector in efficiency (effective) units so that, if $\theta = 0.9$, inputs are 90% efficient (i.e., the use of each input could be reduced by 10% without reducing outputs, if inefficiency is eliminated). Similarly, if $\lambda = 1.05$, each output could be increased by 5% without increasing any input, when inefficiency is eliminated. If $\theta = 1$ and $\lambda > 1$, then we have OO technical inefficiency. Similarly, if $\lambda = 1$ and $\theta < 1$, then we have IO technical inefficiency. Finally, if $\lambda \cdot \theta = 1$, technical inefficiency is said to be hyperbolic, which means that if the inputs are contracted by a constant proportion, outputs are expanded by the same proportion. That is, instead of moving to the frontier by either expanding outputs (keeping the inputs unchanged) or contracting inputs (holding outputs unchanged), the hyperbolic measure chooses a path to the frontier that leads to a simultaneous increase in outputs and a decrease in inputs by the same rate.

3.1. The Cobb-Douglas Multiple Output Transformation Function.

We start from the case where the transformation function is separable (i.e., the output function is separable from the input function) so that $AT(\theta\boldsymbol{x}, \lambda\boldsymbol{y}) = 1$ can be rewritten as $AT_y(\lambda\boldsymbol{y}) \cdot T_x(\theta\boldsymbol{x}) = 1$. If we assume that both $T_y(\cdot)$ and $T_x(\cdot)$ are of Cobb-Douglas (to be relaxed later), the transformation function can be expressed as

$$A \prod_m \{\lambda y_m\}^{\alpha_m} \prod_j \{\theta x_j\}^{\beta_j} = 1. \tag{3.1}$$

The α_m and β_j parameters are of opposite signs. That is, either $\alpha_m < 0\ \forall m$ or $\beta_j > 0\ \forall j$ and vice versa. Note that there is an identification issue stemming from (3.1). A, α_m and β_j cannot be separately identified without further restrictions. The reason for this is that we can always rescale $\boldsymbol{y}$ or $\boldsymbol{x}$ along with A and still obtain 1. We can select one parameter to fix, our normalization, to circumvent this issue. Different normalizations provide different interpretations for (3.1). For

example, if we normalize $\alpha_1 = -1$ and $\theta = 1$, then we get a production function type specification:

$$y_1 = A \prod_{m=2} y_m^{\alpha_m} \prod_j {x_j}^{\beta_j} \ \lambda^{\sum_m \alpha_m}. \tag{3.2}$$

Output-oriented technical efficiency in this model is $TE = \lambda^{\sum_m \alpha_m}$ and output-oriented technical inefficiency is $u = \ln TE = \{\sum_m \alpha_m\} \ln \lambda < 0$ since, in (3.2), $\ln \lambda > 0$ and $\alpha_m < 0 \ \forall \ m \Rightarrow \sum \alpha_m < 0$.

If we rewrite (3.1) as

$$A\, y_1^{\sum_m \alpha_m} \prod_{m=2} \{y_m/y_1\}^{\alpha_m} \prod_j x_j^{\beta_j} \ \theta^{\sum_j \beta_j} \ \lambda^{\sum_m \alpha_m} = 1, \tag{3.3}$$

and use the normalization $\sum_m \alpha_m = -1$ and $\theta = 1$, then we get the output distance function (ODF) formulation (Shephard 1953), viz.,

$$y_1 = A \prod_{m=2} \{y_m/y_1\}^{\alpha_m} \prod_j x_j^{\beta_j} \ \lambda^{-1}, \tag{3.4}$$

where output-oriented technical inefficiency $u = -\ln \lambda < 0$. Technical inefficiency in models (3.2) and (3.4) are different because the output variables (as regressors) appear differently as different normalizations are used.

Similarly, if we rewrite (3.1) as

$$A\, x_1^{\sum_j \beta_j} \prod_m y_m^{\alpha_m} \prod_{j=2} \{x_j/x_1\}^{\beta_j} \ \theta^{\sum_j \beta_j} \ \lambda^{\sum_m \alpha_m} = 1, \tag{3.5}$$

and use the normalization $\sum_j \beta_j = -1$ (note that now we are assuming $\beta_j < 0 \, \forall \, j$ and therefore $\alpha_m > 0 \, \forall \, m$) and $\lambda = 1$, to get the input distance function (IDF) formulation (Shephard 1953), viz.,

$$x_1 = A \prod_m y_m^{\alpha_m} \prod_{j=2} \{x_j/x_1\}^{\beta_j} \ \theta^{-1}, \tag{3.6}$$

where input-oriented technical inefficiency is $u = -\ln \theta > 0$, which is the percentage over-use of inputs due to inefficiency.

Although IO and OO efficiency measures are popular, sometimes a hyperbolic measure of efficiency is used. In this measure the product of λ and θ is unity meaning that the approach to the frontier from an inefficient point takes the path of a parabola (all the inputs are decreased by k percent and the outputs are increased by $1/k$ percent). To get the hyperbolic measure from the above IDF all we need to do is to use the normalization $\sum_j \beta_j = -1$ and $\lambda = \theta^{-1}$ in (3.5) which gives the **hyperbolic input distance function** (Färe, Grosskopf, Noh & Weber 2005, Cuesta & Zofio 2005), viz.,

$$x_1 = A \prod_m y_m^{\alpha_m} \prod_{j=2} \{x_j/x_1\}^{\beta_j} \ \lambda^{\{1+\sum_m \alpha_m\}}. \tag{3.7}$$

Since (3.6) and (3.7) are identical algebraically, $-\ln\theta$ in (3.6) is the same as $(1+\sum_m \alpha_m)\ln\lambda$ in (3.7), and one can get $\ln\lambda$ after estimating inefficiency from either of these two equations.

Note that all these specifications are algebraically the same in the sense that, if the technology is known, inefficiency can be computed from any one of these specifications. It should also be noted that, although we use α_m and β_j notations in all the specifications, these are not the same because of different normalizations. However, once a particular model is chosen, the estimated parameters from that model can be uniquely linked to those in the transformation function in (3.1). Another warning: our results show algebraic relations not econometric ones. Econometric estimation will not give the same results in all formulations simply because of the fact the dependent (endogenous) variable is not the same in each formulations. This is something that is often ignored. Researchers estimating IDF (ODF) assume that all the covariates in the right-hand-side are exogenous (not correlated with inefficiency and the noise term). Since the right-hand-side variables in the IDF and ODF are different it is not possible to have a situation in which the covariates will be uncorrelated with the noise and inefficient terms no matter whether one uses an IDF or ODF.

3.2. **The Translog Multiple Output Transformation Function.** We write the transformation function as $AT(\boldsymbol{y}^*,\boldsymbol{x}^*)=1$, where $\boldsymbol{y}^*=y\lambda$, $\boldsymbol{x}^*=x\theta$, and $T(\boldsymbol{y}^*,\boldsymbol{x}^*)$ is assumed to be translog, i.e.,

$$\begin{aligned}\ln T(\boldsymbol{y}^*,\boldsymbol{x}^*) =& \sum_m \alpha_m \ln y_m^* + \frac{1}{2}\sum_m\sum_n \alpha_{mn}\ln y_m^*\ln y_n^* + \sum_j \beta_j \ln x_j^* + \\ & \frac{1}{2}\sum_j\sum_k \beta_{jk}\ln x_j^*\ln x_k^* + \sum_m\sum_j \delta_{mj}\ln y_m^*\ln x_j^*. \end{aligned} \tag{3.8}$$

The above function is assumed to satisfy the symmetry restrictions $\beta_{jk}=\beta_{kj}$ and $\alpha_{mn}=\alpha_{nm}$. As with the Cobb-Douglas specification, not all of the parameters in (3.8) are simultaneously identified. One can use the following normalizations ($\alpha_1=-1, \alpha_{1n}=0, \forall\, n, \delta_{1j}=0, \forall\, j, \theta=1$) to obtain a pseudo production function, viz.,

$$\begin{aligned}\ln y_1 =& \alpha_0 + \sum_j \beta_j \ln x_j + \frac{1}{2}\sum_j\sum_k \beta_{jk}\ln x_j\ln x_k + \sum_{m=2} \alpha_m \ln y_m + \\ & \frac{1}{2}\sum_{m=2}\sum_{n=2} \alpha_{mn}\ln y_m\ln y_n + \sum_{m=2}\sum_j \delta_{mj}\ln y_m\ln x_j + u,\end{aligned}$$

where

$$u=\ln\lambda(-1+\sum_{m=2}\alpha_m+\sum_{m=2}\sum_{n=2}\alpha_{mn}\ln y_n+\sum_{m=2}\sum_j\delta_{mj}\ln x_j)+\frac{1}{2}\sum_{m=2}\sum_{n=2}\alpha_{mn}(\ln\lambda)^2. \tag{3.9}$$

If we rewrite (3.8) as

$$\ln T(\boldsymbol{y}^*, \boldsymbol{x}^*) = \sum_{m=2} \alpha_m \ln(y_m/y_1) + \frac{1}{2}\sum_{m=2}\sum_{n=2} \alpha_{mn} \ln(y_m/y_1)\ln(y_n/y_1) + \sum_j \beta_j \ln x_j^* + \frac{1}{2}\sum_j \sum_k \beta_{jk} \ln x_j^* \ln x_k^* + \sum_{m=2}\sum_j \delta_{mj} \ln x_j^* \ln(y_m/y_1) + \left[\sum_m \alpha_m\right] \ln y_1^* + \sum_m \left[\sum_n \alpha_{mn}\right] \ln y_m \ln y_1^* + \sum_j \left[\sum_m \delta_{mj}\right] \ln x_j^* \ln y_1^*, \tag{3.10}$$

and use a different set of normalizations, viz., $\sum_m \alpha_m = -1$, $\sum_n \alpha_{mn} = 0,\ \forall\, m$, $\sum_m \delta_{mj} = 0, \forall\, j$, $\theta = 1$, we obtain the output distance function representation,[12] viz.,

$$\ln y_1 = \alpha_0 + \sum_j \beta_j \ln x_j + \frac{1}{2}\sum_j \sum_k \beta_{jk} \ln x_j \ln x_k + \sum_{m=2} \alpha_m \ln \hat{y}_m + \frac{1}{2}\sum_{m=2}\sum_{n=2} \alpha_{mn} \ln \hat{y}_m \ln \hat{y}_n + \sum_j \sum_{m=2} \delta_{mj} \ln x_j \ln \hat{y}_m + u, \tag{3.11}$$

where $u = -\ln \lambda < 0, \hat{y}_m = y_m/y_1, m = 2, \cdots, M$.

Furthermore, if we rewrite (3.8) as

$$\ln T(\boldsymbol{y}^*, \boldsymbol{x}^*) = \sum_m \alpha_m \ln y_m^* + \frac{1}{2}\sum_m \sum_n \alpha_{mn} \ln y_m^* \ln y_n^* + \sum_{j=2} \beta_j \ln(x_j/x_1) + \frac{1}{2}\sum_{j=2}\sum_{k=2} \beta_{jk} \ln(x_j/x_1)\ln(x_k/x_1) + \sum_m \sum_{j=2} \delta_{mj} \ln(x_j/x_1) \ln y_m^* + \left[\sum_j \beta_j\right] \ln x_1^* + \sum_j \left[\sum_k \beta_{jk}\right] \ln x_j \ln x_1^* + \sum_m \left[\sum_j \delta_{mj}\right] \ln y_m^* \ln x_1^*, \tag{3.12}$$

and use a different set of normalizations, viz., $\sum_j \beta_j = -1$, $\sum_k \beta_{jk} = 0, \forall\, j$, $\sum_j \delta_{mj} = 0, \forall\, m$, $\lambda = 1$, we get the input distance function representation,[13] viz.,

$$\ln x_1 = \alpha_0 + \sum_{j=2} \beta_j \ln \hat{x}_j + \frac{1}{2}\sum_{j=2}\sum_{k=2} \beta_{jk} \ln \hat{x}_j \ln \hat{x}_k + \sum_m \alpha_m \ln y_m + \frac{1}{2}\sum_m \sum_n \alpha_{mn} \ln y_m \ln y_n + \sum_m \sum_{j=2} \delta_{mj} \ln \hat{x}_j \ln y_m + u, \tag{3.13}$$

where $u = -\ln \theta > 0$, $\hat{x}_j = x_j/x_1, j = 2, \cdots, J$.

[12] Note that these normalizing constraints make the transformation function homogeneous of degree one in outputs. In the efficiency literature one starts from a distance function (which is the transformation function with inefficiency built in) and imposes linear homogeneity (in outputs) constraints to get the ODF. Here we get the same end-result without using the notion of a distance function to start with.

[13] Note that these normalizing constraints make the transformation function homogeneous of degree one in inputs. In the efficiency literature one defines the IDF as the distance (transformation) function which is homogeneous of degree one in inputs. Here we view the homogeneity property as identifying restrictions on the transformation function without using the notion of a distance function.

To get to the hyperbolic specification in the above IDF we start from (3.12) and use the normalization $\ln\lambda = -\ln\theta$ in addition to $\sum_j \beta_j = -1$, $\sum_k \beta_{jk} = 0$, $\forall\, j$, $\sum_j \delta_{mj} = 0$, $\forall\, m$. This gives the **hyperbolic** IDF, viz.,

$$\ln x_1 = \alpha_0 + \sum_m \alpha_m \ln y_m + \frac{1}{2}\sum_m \sum_n \alpha_{mn} \ln y_m \ln y_n + \sum_{j=2} \beta_j \ln \hat{x}_j + \frac{1}{2}\sum_{j=2}\sum_{k=2} \beta_{jk} \ln \hat{x}_j \ln \hat{x}_k + \sum_m \sum_{j=2} \delta_{mj} \ln \hat{x}_j \ln y_m + u_h, \tag{3.14}$$

where

$$u_h = \ln\lambda \left\{ 1 + \left[\sum_m \alpha_m\right] + \sum_m \left[\sum_n \alpha_{mn}\right] \ln y_m + \left[\sum_j \delta_{mj}\right] \ln \hat{x}_j \right\} + \frac{1}{2}\sum_m \sum_n \alpha_{mn} \{\ln\lambda\}^2. \tag{3.15}$$

It is clear from the above that u_h is related to $\ln\lambda$ in a highly nonlinear fashion. It is quite complicated to estimate $\ln\lambda$ from (3.14) starting from distributional assumption on $\ln\lambda$ unless RTS is unity.[14] However, since (3.13) and (3.14) are identical, their inefficiencies are also the same. That is, $u = -\ln\theta$ in (3.13) is the same as u_h in (3.14). Thus, the estimated values of input-oriented inefficiency $\ln\theta$ from (3.13) can be used to estimate hyperbolic inefficiency $\ln\lambda$ by solving the quadratic equation $-\ln\theta = \ln\lambda\{1 + [\sum_m \alpha_m] + \sum_m [\sum_n \alpha_{mn}] \ln y_m + [\sum_j \delta_{mj}] \ln \hat{x}_j\} + \frac{1}{2}\sum_m \sum_n \alpha_{mn}\{\ln\lambda\}^2$.

It is clear from the above that starting from the translog transformation function specification in (3.8) one can derive the production function, the output and input distance functions simply by using different normalizations. Furthermore, these formulations show how technical inefficiency transmits from one specification into another. As before we warn the readers that the notations α, β and δ are not the same across different specifications. However, starting from any one of them, it is possible to express the parameters in terms of those in the transformation function. Note that other than the input and output distance functions, technical inefficiency appears in a very complicated form. So although all these specifications are algebraically the same, the question that naturally arises is which formulation is easier to estimate.

To sum up, the lesson from this section is that one can start from a flexible parametric transformation function and use appropriate parameters normalization to get the IDF and ODF. It is not necessary to start from a distance function. Note that the transformation function can be used with or without inefficiency. In the latter case one can estimate returns to scale, technical change, input substitutability/complementarity, and other metrics of interest. The distance function is primarily designed to address inefficiency.

[14] This relationship is similar to the relationship between input- and output-oriented technical inefficiency, estimation of which is discussed in detail in Kumbhakar & Tsionas (2006).

A natural empirical question is whether one should use the IDF or the ODF. In the IDF (which is dual to a cost function) the implicit assumption is that inputs are endogenous and outputs are exogenous. The opposite is the case for ODF. It can be shown that if outputs are exogenous and firms minimize cost, input ratios are exogenous (if input prices are exogenous). This is the logic behind using IDF in applications where outputs are believed to be exogenous (service industries). Similarly, if inputs are exogenous and firms maximize revenue, output ratios can be treated as exogenous, and therefore one can use ODF to estimate the technology parameters consistently. There are not many situations where exogeneity of inputs can be justified. If both inputs and outputs are endogenous, both IDF and ODF will give inconsistent parameter estimates.

One possible solution to the endogeneity problem is to use the IDF (ODF) and append the first-order conditions (FOCs) of cost minimization (revenue maximization) and use a system approach. Tsionas, Kumbhakar & Malikov (2014) used such a system to estimate the technology represented by an IDF. Note that in this approach one needs input price data which appear in the FOCs. The alternative is to use a cost function (either a single equation or a system that includes the cost shares also). Note that estimation of a cost function relies on variability of input prices. Finally, if both inputs and outputs are endogenous one can use either the IDF or the ODF together with the FOCs of profit maximization and use a system approach similar to the IDF system. Note that for such a system, one needs output prices also. However, since profit is not directly used in this system, estimation works even if actual profit is negative for some firms. On the other hand, if profit is negative one cannot use a translog profit function.

4. Cost and Profit Stochastic Frontier Models

In this section, we discuss cost and profit frontier models by showing how technical inefficiency is transmitted from the production frontier to the cost/profit frontier. This allows to examine the extent to which cost (profit) is increased (decreased) if the production plan is inefficient. Note that here we are explicitly using economic behavior, i.e., cost minimization, while estimating the technology. Here our focus is on the examination of cost frontier models using cross-sectional data. We also restrict our attention to examining only technical inefficiency and we assume that producers are fully efficient from an allocative perspective.

In modeling and estimating the impact of technical inefficiency on production it is assumed, at least implicitly, that inputs are exogenously given and the scalar output is a response to the inputs. On the other hand, in modeling and estimating the impact of technical inefficiency on costs, it is assumed that output is given and inputs are the choice variables (i.e., the goal is to minimize cost for a given level of output). However, if the objective of producers is to maximize profit, *both* inputs and output are choice variables. That is, inputs and outputs are chosen by the producers in such a way that profit is maximized.

It is perhaps worth noting that profit inefficiency can be modeled in two ways. One way is to make the intuitive and common sense argument that, if a producer is inefficient, his/her profit will be lower, everything else being the same. Thus, one can specify a model in which actual (observed) profit is a function of some observed covariates (profit drivers) and unobserved inefficiency. In this sense the model is similar to a production function model. The error term in such a model is $v - u$ where v is noise and u is inefficiency. The other approach is to use the duality result: That is, derive a profit function allowing production inefficiency. Since profit maximization behavior is widely used in neoclassical production theory, we provide a framework to analyze inefficiency in this setting.

In what follows, we start with input-oriented (IO) technical inefficiency (Farrell 1957), since this specification is the most common within the cost frontier literature. IO inefficiency is natural because in a cost minimization case the focus is on input use, *given* outputs. That is, it is assumed that output is given and inputs are the choice variables (i.e., the goal is to minimize cost for a given level of output). The discussion on output-oriented (OO) technical inefficiency will be discussed later.

4.1. Input-oriented Technical Inefficiency for the Cost Frontier.

Here we focus on firms for whom the objective is to produce a given level of output with the minimum possible cost. We also assume that the firm is technically inefficient; that is, it either produces less than the maximum possible output or uses more inputs than is necessary to produce a given level of output. In the context of cost minimization, the input-oriented measure which focuses on input over-use is intuitive and appropriate. For an inefficient firm in this setup, the additional cost must be due to the over-use of inputs, and cost savings (from eliminating inefficiency) will come from eliminating the excess usage of inputs. Farrell (1957) used a radial measure of inefficiency, thereby assuming

that the technically efficient point in the production iso-quant can be obtained by reducing usage of all the variable inputs by the same proportion.

The cost minimization problem for producer i under an input-oriented technical inefficiency specification is (the producer/observation subscript i is omitted)

$$\min \ \boldsymbol{w}'\boldsymbol{x} \quad \text{s.t.} \quad y = m(\boldsymbol{x}e^{-\eta};\boldsymbol{\beta}), \tag{4.1}$$

$$\text{first order conditions:} \qquad \frac{m_j(\boldsymbol{x}e^{-\eta};\boldsymbol{\beta})}{m_1(\boldsymbol{x}e^{-\eta};\boldsymbol{\beta})} = \frac{w_j}{w_1}, \quad j = 2,\ldots,J, \tag{4.2}$$

where $\eta \geq 0$ is the input-oriented technical inefficiency which measures the percentage by which all the inputs are over-used in producing output level y. We use η instead of u to distinguish between input and output oriented inefficiency.

Alternatively, one can interpret $\eta \geq 0$ as the percentage by which the usage of all the inputs can be reduced without reducing the output level y. It is also possible to view $e^{-\eta} \leq 1$ as the efficiency factor. Thus, although an inefficient firm uses x_j amount of input, effectively it is worth only $x_j^e \equiv x_j e^{-\eta} \leq x_j$. The marginal product of $x_j e^{-\eta}$ is $m_j(\cdot)$, which is the partial derivative of $m(\cdot)$ with respect to the input $x_j e^{-\eta}$, and this also depends on how effectively the input is used. The second set of $J-1$ equations represent the first-order conditions of the cost minimization problem.

The $J-1$ FOCs in (4.2) along with the production function in (4.1) can be used to solve for the J input demand functions. In fact, since $x_j e^{-\eta}$ appears everywhere in (4.2), it is easier to solve for x_j, $j = 1,\ldots,J$, in their effective units which are simply x_j adjusted for technical inefficiency $(x_j e^{-\eta})$. These input demand functions can be expressed as $x_j e^{-\eta} = \psi_j(\boldsymbol{w}, y)$, $j = 1,\ldots,J$. We use them to define the cost function C^* as

$$C^*(\boldsymbol{w}, y) = \sum_j w_j x_j e^{-\eta}, \tag{4.3}$$

which can be viewed as the minimum cost function for the following problem:

$$\min_{\{x_j e^{-\eta}\}} \boldsymbol{w}'\boldsymbol{x}e^{-\eta} \quad \text{s.t.} \quad y = m(\boldsymbol{x}e^{-\eta};\boldsymbol{\beta}).$$

The $C^*(\cdot)$ function is the *frontier* cost function, which gives the minimum cost given the vector of input prices $\boldsymbol{w}$ and the observed level of output y. Note that this cost function measures the cost of producing y when inputs are adjusted for their efficiency (i.e., the cost of effective units of inputs). Thus, the minimum cost $\boldsymbol{w}'\boldsymbol{x}e^{-\eta}$ would be less than the actual cost $\boldsymbol{w}'\boldsymbol{x}$. Although $C^*(\cdot)$ is not observed, it can be used to derive the input demand functions and we can also relate it to actual (observed) cost.

To relate actual cost C^a with the unobserved minimum cost C^*, first, we make use of Shephard's lemma to (4.3) which is

$$\frac{\partial C^*}{\partial w_j} = x_j e^{-\eta} \Longrightarrow \qquad \frac{\partial \ln C^*}{\partial \ln w_j} = \frac{w_j x_j e^{-\eta}}{C^*} = \frac{w_j x_j}{w'x} \equiv S_j.$$

Therefore, $w_j x_j e^{-\eta} = C^* \cdot S_j$ or $x_j e^{-\eta} = \frac{C^* \cdot S_j}{w_j}$. We write actual cost as

$$C^a = \sum_j w_j x_j = C^* e^{\eta} \Rightarrow \ln C^a = \ln C^*(\boldsymbol{w}, y) + \eta. \tag{4.4}$$

The relationship in (4.4) shows that log actual cost is increased by η because all the inputs are over-used by η.

For the efficiency index of a producer, we take the ratio of the minimum to actual cost, which from (4.4) is $e^{-\eta} = \frac{C^*}{C^a}$. By definition, the ratio is bounded between 0 and 1, and in the estimation it is numerically guaranteed by imposing $\eta \geq 0$ so that $e^{-\eta}$ is between 0 and 1. Although this efficiency index has an intuitive interpretation, viz., the higher values indicating higher level of efficiency, one may also be interested in knowing the percentage increase in cost due to inefficiency, which may be obtained based on the approximation

$$\frac{C^a}{C^*} - 1 = e^{\eta} - 1 \approx \eta.$$

Alternatively, $\eta = \ln C^a - \ln C^*(\boldsymbol{w}, y) = \ln\left(\frac{C^a}{C^*(\boldsymbol{w},y)}\right)$. Thus $100 \times \eta$ (when η is small) is the percentage by which actual cost exceeds the minimum cost due to technical inefficiency. Note that this interpretation of η is consistent with input-oriented technical inefficiency. Since the inputs are over-used by $100 \times \eta$ percent and we are assuming no allocative inefficiency here, cost is increased by the same percentage. This is true irrespective of the functional form chosen to represent the underlying production technology.

When estimating the model in (4.4), a noise term, v, is usually appended to the equation to capture modeling errors. Unlike the production function, the v term does not have a natural interpretation. It is quite ad hoc and added to the cost function to make it stochastic.[15]

In the following, we add a subscript to represent producer i, and use the translog specification on $\ln C^*(\boldsymbol{w}_i, y_i)$, viz.,

$$\begin{aligned} \ln C_i^a =& \ln C^*(\boldsymbol{w}_i, y_i) + v_i + \eta_i \\ =& \beta_0 + \sum_j \beta_j \ln w_{j,i} + \beta_y \ln y_i + \frac{1}{2}\sum_j \sum_k \beta_{jk} \ln w_{j,i} \ln w_{k,i} + \frac{1}{2}\beta_{yy} \ln y_i \ln y_i \\ &+ \sum_j \beta_{jy} \ln w_{j,i} \ln y_i + v_i + \eta_i. \end{aligned} \tag{4.5}$$

4.1.1. *Price Homogeneity.* Symmetric restrictions require $\beta_{jk} = \beta_{kj}$. Since the cost function is homogeneous of degree 1 in the input prices (i.e., $w_{1,i}, \ldots, w_{J,i}$), it has to satisfy the following additional parameter restrictions:

$$\sum_j \beta_j = 1, \qquad \sum_j \beta_{jk} = 0 \; \forall \, k, \qquad \sum_j \beta_{jy} = 0. \tag{4.6}$$

[15]This is not a problem specific to stochastic frontier analysis. It applies to the neoclassical cost function as well where a noise term is appended before estimation.

An easier way to impose the price homogeneity condition is to use $w_{j,i}$ for an arbitrary choice of j and normalize C_i^a and other input prices by it. To see how this works, consider a simple model with $J = 2$.

$$\begin{aligned}\ln C_i^a =&\beta_0 + \beta_y \ln y_i + \beta_1 \ln w_{1,i} + \beta_2 \ln w_{2,i} + \frac{1}{2}\beta_{yy}(\ln y_i)^2 + \frac{1}{2}\beta_{11}(\ln w_{1,i})^2 \\ &+ \frac{1}{2}\beta_{22}(\ln w_{2,i})^2 + \beta_{12} \ln w_{1,i} \ln w_{2,i} + \beta_{1y} \ln w_{1,i} \ln y_i + \beta_{2y} \ln w_{2,i} \ln y_i + v_i + \eta_i.\end{aligned}$$

Price homogeneity requires that $\beta_1 + \beta_2 = 1$; $\beta_{11} + \beta_{12} = 0$, $\beta_{22} + \beta_{12} = 0$; and $\beta_{1y} + \beta_{2y} = 0$. Equivalently, the constraints are $\beta_1 = 1 - \beta_2$, $\beta_{12} = -\beta_{22}$, $\beta_{11} = -\beta_{12} = \beta_{22}$, and $\beta_{1y} = -\beta_{2y}$. If we substitute these constraints into the cost function, the price homogeneity restrictions will be built into the model. After the substitutions and straightforward manipulation, we get

$$\begin{aligned}\ln\left(\frac{C_i^a}{w_{1,i}}\right) =&\beta_0 + \beta_y \ln y_i + \beta_2 \ln\left(\frac{w_{2,i}}{w_{1,i}}\right) + \frac{1}{2}\beta_{yy}(\ln y_i)^2 + \frac{1}{2}\beta_{22} \ln\left(\frac{w_{2,i}}{w_{1,i}}\right)^2 \\ &+ \beta_{2y} \ln\left(\frac{w_{2,i}}{w_{1,i}}\right) \ln y_i + v_i + \eta_i.\end{aligned}$$

The above equation is equivalent to the one obtained by dividing C_i^a and other input prices ($w_{2,i}$ in this case) by $w_{1,i}$. We may also choose to express β_2, β_{12}, and β_{22} as functions of β_1 and β_{11} based on the price homogeneity conditions, and derive a similar model that has $w_{2,i}$ appearing as the normalizing price. That is, price homogeneity can be built into the model by an arbitrary choice of $w_{1,i}$ and $w_{2,i}$ as the normalizing price.

4.1.2. *Monotonicity and Concavity.* Production theory requires a cost function to be monotonic and concave in input prices and output. The monotonicity condition requires cost to be non-decreasing in input prices and output: $C_i^*(\boldsymbol{w}_i^1, y_i) \geq C_i^*(\boldsymbol{w}_i^0, y_i)$ if $\boldsymbol{w}_i^1 \geq \boldsymbol{w}_i^0$ and $C_i^*(\boldsymbol{w}_i, y_i^1) \geq C_i^*(\boldsymbol{w}_i, y_i^0)$ if $y_i^1 \geq y_i^0$. Given that

$$\frac{\partial C_i^*}{\partial w_{j,i}} = \frac{\partial \ln C_i^*}{\partial \ln w_{j,i}} \times \frac{C_i^*}{w_{j,i}},$$

and both C_i^* and $w_{j,i}$ are positive, then,

$$\operatorname{sign}\left(\frac{\partial C_i^*}{\partial w_{j,i}}\right) = \operatorname{sign}\left(\frac{\partial \ln C_i^*}{\partial \ln w_{j,i}}\right). \tag{4.7}$$

Note that the partial derivative on the right-hand-side of (4.7) is simply input j's cost share. Thus, the monotonicity condition on input prices can be checked from the positivity of the estimated cost shares. Similarly, we can also check the sign of $\partial \ln C_i^* / \partial \ln y_i$ for the monotonicity condition of output. Returning to (4.5), the partial derivatives (i.e., the input shares) are the following:

$$\begin{aligned}\frac{\partial \ln C_i}{\partial \ln w_{s,i}} &= \beta_s + \sum_j \beta_{sj} \ln w_{j,i} + \beta_{sy} \ln y_i, \qquad s = 1, \ldots, J, \\ \frac{\partial \ln C_i}{\partial \ln y_i} &= \beta_y + \beta_{yy} \ln y_i + \sum_j \beta_{jy} \ln w_{j,i}.\end{aligned}$$

As the shares are functions of $\ln y_i$ and $\ln w_{j,i}$, $j = 1, \ldots, J$, the partial derivatives are observation-specific.

The concavity condition requires that the following Hessian matrix with respect to input prices is negative semidefinite (Diewert & Wales 1987):

$$\frac{\partial^2 C_i^*}{\partial \boldsymbol{w}_i \partial \boldsymbol{w}_i'} = \frac{\partial^2 \ln C_i^*}{\partial \ln \boldsymbol{w}_i \partial \ln \boldsymbol{w}_i'} - \operatorname{diag}(\boldsymbol{S}_i) + \boldsymbol{S}_i \boldsymbol{S}_i', \tag{4.8}$$

where $\boldsymbol{S}_i$ is the vector of input shares defined as

$$\boldsymbol{S}_i = \frac{\partial \ln C_i^*}{\partial \ln \boldsymbol{w}_i}.$$

A matrix is negative semidefinite if all the eigenvalues are less than or equal to zero. Notice that for a translog model, the first matrix on the right-hand-side of (4.8) contains only the coefficients (but not data) of the model and hence is observation invariant. However, each of the share equations in the $\boldsymbol{S}_i$ vector is a function of the data. Thus, the Hessian matrix of (4.8) is observation-specific. Therefore, like monotonicity, concavity conditions cannot be imposed by restricting the parameters alone. Ideally, monotonicity and the concavity conditions should be satisfied for each observation.[16]

4.1.3. *Maximum Likelihood Estimation.* We now go back to the cost model with IO technical inefficiency and noise, viz.,

$$\ln C^a = \ln C^*(\boldsymbol{w}, y) + \eta + v. \tag{4.9}$$

To estimate such a model, we impose distributional assumptions on v and η, based on which the likelihood function can be derived and the parameters estimated for any parametric cost frontier.

The ML approach to estimate the cost frontier estimation is very similar to the ML approach that we used to estimate the production frontier in Section 2. The only difference is that the variable $-u$ in Section 2 is replaced by η. Thus, the same modeling strategies can be applied to the cost model. In fact, if v is normally distributed (so that theoretically one can replace $-v$ with $+v$ without altering anything) one can multiply both sides of (4.9) by -1 to get the same structure of the composed error used in Section 2. In doing so one can use the same likelihood function, the same codes, etc., to estimate the cost frontier.

There is, nevertheless, a direct way of handling the problem. With the same distribution assumption on u and η, the log-likelihood functions are very similar for the production and the cost functions, with only one important difference in the sign in front of the inefficiency term. Namely, ϵ is now $v + \eta$ whereas it was $v - u$ before, and so in practice we need only replace ϵ by $-\epsilon$ to get the likelihood function for the cost frontier model. All the statistical/economic properties of the

[16]Ryan & Wales (2000) suggested normalizing the data at a point in such a way that the number of concavity violations is minimum. That is, they were in favor of imposing monotonicity and concavity conditions locally. On the other hand, Parmeter & Racine (2012) and Parmeter, Sun, Henderson & Kumbhakar (2014) suggested a procedure to impose monotonicity and concavity conditions globally. These procedures are often quite demanding to apply in practice.

models discussed in Section 2 are applicable to the cost frontier models once the sign difference is taken into account.

Given the close relationship with the production frontier models, we omit specific details and refer back to Section 2.

4.2. **Output-Oriented Technical Inefficiency for the Cost Frontier.** In this section, we discuss the stochastic cost frontier model with output-oriented technical inefficiency. Recall that the input-oriented measure of inefficiency starts from the fact that, if a producer is not efficient, then he/she is not able to use the inputs effectively. That is, there are slacks in the inputs and it is possible to reduce input-usage without reducing output. Consequently, the input-oriented measure is practical and intuitive when output is exogenously given (demand determined) and the objective is to minimize cost (or maximize the proportional reduction in input usage) without reducing output. On the other hand, output-oriented technical inefficiency measures the potential increase in output without increasing the input quantities. Alternatively, it can be viewed as a measure of output loss resulting from failure to produce the maximum possible output permitted by the technology. Thus, the output-oriented measure is intuitive when the inputs are exogenously given to the manager and the objective is to produce as much output as possible.

Although inefficient production can be viewed from either input- or output-oriented angles, it is shown below that without additional restrictions on the underlying production function, a stochastic cost frontier model with output-oriented technical inefficiency is difficult to estimate (Kumbhakar & Wang 2006). Imposing the assumption of a homogeneous production function makes the estimation easier.

The cost minimization problem with output-oriented technical inefficiency is as follows (observation subscripts omitted):

$$\min \quad \boldsymbol{w}'\boldsymbol{x} \quad \text{s.t.} \quad y = m(\boldsymbol{x};\boldsymbol{\beta})e^{-u}, \tag{4.10}$$

where $u \geq 0$ is the output-oriented technical inefficiency (as in the case with the production function in Section 2). Multiplying e^{-u} on both sides of the production function, we rewrite the minimization problem and the associated FOCs as:

$$\begin{aligned} &\min \quad \boldsymbol{w}'\boldsymbol{x} \quad \text{s.t.} \quad ye^{u} = m(\boldsymbol{x};\boldsymbol{\beta}), \\ \text{FOCs:} \quad &\frac{m_j(\boldsymbol{x})}{m_1(\boldsymbol{x})} = \frac{w_j}{w_1}, \quad j = 2,\ldots,J, \end{aligned} \tag{4.11}$$

where as before $m_j(\cdot)$ is the marginal product of x_j (the partial derivative of $m(\cdot)$ with respect to x_j). The $J-1$ FOCs and the production function in (4.11) can be used to solve for the input demand functions of the J inputs. The solution of x_j is $x_j = x_j(\boldsymbol{w}, ye^u)$. Therefore, the minimum cost of producing ye^u given the input prices, $\boldsymbol{w}$, is $C^* = C^*(\boldsymbol{w}, ye^u) = \sum_j \boldsymbol{w}'\boldsymbol{x}$. Since $\boldsymbol{w}'\boldsymbol{x}$ is also the *actual* cost (C^a) of producing output level y when the production is technically inefficient, we

have

$$C^a = C^*(\boldsymbol{w}, ye^u).$$

To estimate the above relationship, we add a producer (observation) subscript i and assume that $C_i^*(\cdot)$ has a flexible functional form such as the translog. We also add a random error v_i for estimation purposes. With these we write the translog model as

$$\begin{aligned}\ln C_i^a =&\beta_0 + \sum_j \beta_j \ln w_{j,i} + \beta_y \ln(y_i e^u) + \frac{1}{2}\sum_k \sum_j \beta_{jk} \ln w_{j,i} \ln w_{k,i} + \frac{1}{2}\beta_{yy} \ln(y_i e^u)^2 \\ &+ \sum_j \beta_{jy} \ln w_{j,i} \ln(y_i e^u) + v_i.\end{aligned}$$

It is straightforward to show that, upon expanding $\ln(y_i e^u)$ into $\ln y_i + u_i$, the model will have three stochastic components: u_i, u_i^2, and v_i. The presence of the u_i^2 term makes the derivation of the likelihood function in closed form impossible, and so the standard maximum likelihood approach of this model is not feasible.[17]

Imposing homogeneity on the production technology simplifies the problem. If the production technology is homogenous of degree r, then our translog model will have the following parametric restrictions (Christensen & Greene 1976): $\beta_{yy} = 0$ and $\beta_{jy} = 0 \; \forall \; j$, and that $\beta_y = 1/r$ where r is the degree of homogeneity (which is the same as returns to scale). The simplification leads to the following estimation equation:

$$\begin{aligned}\ln C_i^a =&\beta_0 + \sum_j \beta_j \ln w_{j,i} + \beta_y \ln(y_i e^u) + \frac{1}{2}\sum_k \sum_j \beta_{jk} \ln w_{j,i} \ln w_{k,i} + v_i \\ =&\beta_0 + \sum_j \beta_j \ln w_{j,i} + \beta_y \ln y_i + \frac{1}{2}\sum_k \sum_j \beta_{jk} \ln w_{j,i} \ln w_{k,i} + u_i/r + v_i. \qquad (4.12)\end{aligned}$$

If one reparameterizes $\tilde{u}_i = u_i/r$, then the model (4.12) looks exactly like the cost frontier model with input-oriented inefficiency and it can be estimated using the standard maximum likelihood method.

The above example shows how the assumption of homogenous technology helps to simplify the model when the cost function has a translog specification. In fact, the simplification applies to other specifications as well (not just the translog function). In general, with a homogenous of degree r production technology, we have

$$C_i^*(\boldsymbol{w}_i, y_i e^{-u_i}) = (y_i e^{-u_i})^{\frac{1}{r}} \cdot C_i^*(\boldsymbol{w}_i) \Rightarrow \ln C_i^*(\boldsymbol{w}_i, y_i e^{-u_i}) = \frac{1}{r}\ln y_i + \ln C_i^*(\boldsymbol{w}_i) + \frac{u_i}{r}. \quad (4.13)$$

Alternative specifications of $\ln C_i^*(\boldsymbol{w}_i)$ do not make the model more difficult to estimate than (4.12).

Equation (4.12) can be derived in an alternative way. If the production function $m(\cdot)$ is homogeneous of degree r, then the production function in (4.10) is $y_i = m(\boldsymbol{x}_i; \boldsymbol{\beta})e^{-u_i} = m(\boldsymbol{x}_i \cdot e^{-u_i/r}; \boldsymbol{\beta})$.

[17] Simulated maximum likelihood proposed in Kumbhakar & Tsionas (2006) and Greene (2003) can be used to estimate these types of models for which the log-likelihood function cannot be expressed in a closed form.

With reference to (4.1), we have $\eta_i = u_i/r$. That is, *under the assumption of a homogeneous production function*, the output-oriented technical inefficiency (u_i) equals the input-oriented technical inefficiency (η_i) multiplied by a constant which is the returns to scale parameter (r). Because of the similarity, we can use the result from (4.4) and write the current model as

$$\ln C_i^a = \ln C_i^*(\boldsymbol{w}_i, y_i) + u_i/r + v_i. \tag{4.14}$$

Aside from the apparent u_i/r versus η_i, there is an important difference between (4.14) and (4.4) concerning the specification of $\ln C_i^*(\boldsymbol{w}_i, y_i)$. For (4.4), a full, unrestrictive translog specification on $\ln C_i^*(\boldsymbol{w}_i, y_i)$ may be used for empirical estimations, which includes all the interaction terms between $x_{j,i}$ and y_i and the square term of y_i. In deriving (4.14), the technology is assumed to be homogeneous. In the translog specification of $\ln C_i^*(\boldsymbol{w}_i, y_i)$, this assumption imposes the following coefficient restrictions: $\beta_{yy} = 0$ and $\beta_{jy} = 0 \; \forall \; j$. Indeed, imposing the restrictions on the translog specification of $\ln C_i^*(\boldsymbol{w}_i, y_i)$ in (4.14) leads to the identical estimation model of (4.12).

4.2.1. *Estimation of Output-Oriented Inefficiency.* This section discusses the estimation method for the OO efficiency model in which the homogeneity assumption on the technology is imposed. Our focus is on the ML estimation. By defining

$$\tilde{u}_i = \frac{1}{r} u_i,$$

where r measures returns to scale, we write the model as

$$\ln C_i^a(\boldsymbol{w}_i, y_i) = \ln C_i^*(\boldsymbol{w}_i, y_i) + u_i/r + v_i = \beta_y \ln y_i + \ln C_i^*(\boldsymbol{w}_i) + \tilde{u}_i + v_i. \tag{4.15}$$

Note that $\beta_y = 1/r$ under the homogenous technology assumption. $\ln C_i^*(\boldsymbol{w}_i)$ can be assumed to take any functional form. The likelihood function can be derived after imposing distributional assumptions on $\tilde{u}_i$ and v_i, such as the normal, half normal specification. Again, maximum likelihood estimation is equally applicable here.

Care should be taken, however, in interpreting the efficiency estimate of the model. For example, by directly applying the Jondrow et al. (1982) inefficiency formula to the model, we get $E(\tilde{u}_i|\epsilon_i)$ (ϵ_i is the composed error of the model) instead of $E(u_i|\epsilon_i)$. The interpretation of $100^* E(\tilde{u}_i|\epsilon_i)$ is the percentage increase in cost due to input overuse. This is an interpretation of inefficiency.

If one is interested in *output-oriented* measure of inefficiency, the parameters of the distribution of u_i can be easily recovered after the model is estimated. Take, for example, a truncated-normal distribution of $u_i \sim N^+(\mu, \sigma_u^2)$. Then $\tilde{u}_i = \frac{1}{r} \cdot u_i \sim N^+(\mu/r, \sigma_u^2/r^2) \equiv N^+(\tilde{\mu}, \tilde{\sigma}_u^2)$ and $\hat{\mu} = \hat{r} \cdot \hat{\tilde{\mu}}$ and $\hat{\sigma}_u^2 = \hat{r}^2 \cdot \hat{\tilde{\sigma}}_u^2$. The standard errors of $\hat{\mu}$ and $\hat{\sigma}_u^2$ can be obtained using the Delta method together with the variances and covariance of $\hat{\tilde{\mu}}$ and $\hat{\tilde{\sigma}}_u^2$. The above computation requires an estimate of the returns to scale (r). In the simple case of a single output homogeneous function, r is simply the inverse of the output coefficient: $r = 1/\beta_y$. These estimated parameters can then be used to obtain point estimates of u as well as the confidence intervals.

4.3. **Output-Oriented Technical Inefficiency for the Profit Frontier.** The profit maximization problem with output-oriented technical inefficiency is

$$\max_{y,\boldsymbol{x}} py - \boldsymbol{w}'\boldsymbol{x} \qquad \text{s.t. } y = m(\boldsymbol{x},\boldsymbol{q})e^{-u}.$$

The first-order conditions are

$$p\, m_j(\boldsymbol{x},\boldsymbol{q})e^{-u} = w_j \Rightarrow m_j(\boldsymbol{x},\boldsymbol{q}) = \frac{w_j}{p\cdot e^{-u}}, \qquad j = 1,2,\ldots,J,$$

where $\boldsymbol{x}$ and $\boldsymbol{q}$ are vectors of variable and quasi-fixed inputs, respectively; $u \geq 0$ is the output-oriented technical inefficiency, and $m_j(\boldsymbol{x},\boldsymbol{q}) = \partial m(\boldsymbol{x},\boldsymbol{q})/\partial x_j$ is the marginal product of variable input j.

Note that if we substitute $p\cdot e^{-u}$ by $\tilde{p}$ and $y\cdot e^{u}$ by $\tilde{y}$, then we can write the above profit maximization problem using $\tilde{p}$ and $\tilde{y}$, and the standard neoclassical analytics can be applied. Consequently, the profit function for our problem can simply be written as $\pi(\boldsymbol{w},\boldsymbol{q},\tilde{p})$, which comes from the following standard neoclassical maximization problem:

$$\max_{\tilde{y},\boldsymbol{x}} \tilde{p}\tilde{y} - \boldsymbol{w}'\boldsymbol{x}, \qquad \text{s.t.} \qquad \tilde{y} = m(\boldsymbol{x},\boldsymbol{q}).$$

Solutions of the input demand and the output supply functions of the above optimization problem will be functions of $\boldsymbol{w}$, $\boldsymbol{q}$, and $\tilde{p}$. Substituting these input demand and output supply functions in to the objective function, we obtain the profit function, $\pi(\boldsymbol{w},\boldsymbol{q},pe^{-u})$ which is

$$\pi(\boldsymbol{w},\boldsymbol{q},pe^{-u}) = pe^{-u}m(\boldsymbol{x}(\cdot),\boldsymbol{q}) - \boldsymbol{w}'\boldsymbol{x}(\cdot),$$

where $\boldsymbol{x}(\cdot) = \boldsymbol{x}(\boldsymbol{w},\boldsymbol{q},pe^{-u})$ is the input demand function. Note that actual profit π^a is

$$\pi^a = py(\boldsymbol{w},\boldsymbol{q},pe^{-u}) - \boldsymbol{w}'\boldsymbol{x}(\boldsymbol{w},\boldsymbol{q},pe^{-u}) = pe^{-u}m(\boldsymbol{x}(\cdot),\boldsymbol{q}) - \boldsymbol{w}'\boldsymbol{x}(\cdot) = \pi(\boldsymbol{w},\boldsymbol{q},pe^{-u}).$$

If we define the profit frontier as

$$\pi(\boldsymbol{w},p,\boldsymbol{q}) = \pi(\boldsymbol{w},\boldsymbol{q},pe^{-u})\mid_{u=0},$$

then the following relationship can be established

$$\pi^a = \pi(\boldsymbol{w},\boldsymbol{q},pe^{-u}) = \pi(\boldsymbol{w},\boldsymbol{q},p)\cdot h(\boldsymbol{w},\boldsymbol{q},p,u), \tag{4.16}$$

$$\text{or,} \qquad \ln\pi^a = \ln\pi(\boldsymbol{w},\boldsymbol{q},pe^{-u}) = \ln\pi(\boldsymbol{w},\boldsymbol{q},p) + \ln u(\boldsymbol{w},\boldsymbol{q},p,u), \tag{4.17}$$

where $u(\cdot) = \pi(\boldsymbol{w},\boldsymbol{q},pe^{-u})/\pi(\boldsymbol{w},\boldsymbol{q},p) \leq 1$ and therefore $\ln u(\cdot) \leq 0$. This result follows from the monotonicity property of profit function, i.e., since $pe^{-u} \leq p$, $\pi(\boldsymbol{w},\boldsymbol{q},pe^{-u}) \leq \pi(\boldsymbol{w},\boldsymbol{q},p)$.

The above equation shows that the log of actual profit, $\ln\pi^a$, can be decomposed into a profit frontier component, $\ln\pi(\boldsymbol{w},\boldsymbol{q},p)$, and an inefficiency component, $\ln u(\boldsymbol{w},\boldsymbol{q},p,u) \leq 0$. In the following, we show that the equation can be simplified by: (i) making the assumption of a homogeneous production technology; and (ii) utilizing the property of price homogeneity of a profit function.[18]

[18]Note the distinction between the profit function being homogeneous of degree one – a theoretical property that follows from the definition – and the production function being homogeneous of degree one – which is a restriction on the technology and, therefore, not necessary to impose. The homogeneity of the production function can be

Following Lau (1978, p. 151), it can be shown that if the underlying production function is homogeneous of degree r ($r < 1$), then the corresponding profit function is

$$\begin{aligned}\ln \pi^a &= \ln \pi(\boldsymbol{w}, \boldsymbol{q}, pe^{-u}) = \ln \pi(\boldsymbol{w}, \boldsymbol{q}, p) + \ln u(\boldsymbol{w}, \boldsymbol{q}, p, u) \\ &= \frac{1}{1-r} \ln p + \ln G(\boldsymbol{w}, \boldsymbol{q}) - \frac{1}{1-r} u, \end{aligned} \tag{4.18}$$

where $G(\boldsymbol{w}, \boldsymbol{q})$ is a homogeneous function of degree $-r/(1-r)$ in $\boldsymbol{w}$. Now, let us consider the property that a profit function is homogeneous of degree 1 in prices (i.e., in $\boldsymbol{w}$ and pe^{-u}). The price homogeneity property can be built into the profit function by normalizing the profit and the input/output prices by one of the prices. Using p to normalize the equation, we have

$$\ln(\pi^a/p) = \ln G(\boldsymbol{w}/p, \boldsymbol{q}) - \frac{1}{1-r} u, \tag{4.19}$$

which imposes the linear homogeneity restrictions automatically.

Now, let us further simplify the equation by exploring the property that $G(\boldsymbol{w}, \boldsymbol{q})$ is homogeneous of degree $-r/(1-r)$. Recall that $f(x)$ is a homogeneous function of degree μ if $f(x) = \lambda^{-\mu} f(x\lambda)$ where $\lambda \geq 1$. In the present case $\mu = -r/(1-r)$. If we choose $\lambda = 1/(w_1/p)$ and define $\tilde{w}_j = (w_j/p)/(w_1/p)$, then the profit function in (4.19) can be expressed as

$$\begin{aligned}\ln(\pi^a/p) &= -\frac{-r}{1-r} \ln \lambda + \ln G((\boldsymbol{w}/p)\lambda, \boldsymbol{q}) - \frac{1}{1-r} u \\ &= \frac{-r}{1-r} \ln(w_1/p) + \ln G(\tilde{\boldsymbol{w}}, \boldsymbol{q}) - \frac{1}{1-r} u. \end{aligned} \tag{4.20}$$

We assume a translog form of $\pi(\boldsymbol{w}, \boldsymbol{q}, pe^{-u})$, viz.,

$$\begin{aligned}\ln \pi^a =& \ln \pi(\boldsymbol{w}, \boldsymbol{q}, pe^{-u}) \\ =& \beta_0 + \sum \beta_j \ln w_j + \sum \gamma_q \ln q_q + \beta_p \ln(pe^{-u}) \\ &+ \frac{1}{2}\Big[\sum_j \sum_k \beta_{jk} \ln w_j \ln w_k + \sum_q \sum_s \gamma_{qs} \ln q_q \ln q_s + \beta_{pp} \ln(pe^{-u}) \ln(pe^{-u})\Big] \\ &+ \sum_j \beta_{jp} \ln w_j \ln(pe^{-u}) + \sum_j \sum_q \delta_{jq} \ln w_j \ln q_q + \sum_q \gamma_{qp} \ln q_q \ln(pe^{-u}). \end{aligned} \tag{4.21}$$

The symmetry restrictions are $\beta_{jk} = \beta_{kj}$ and $\gamma_{qs} = \gamma_{sq}$. Expressing (4.21) in the normalized form (to impose the price homogeneity property of a profit function) and after a few algebraic

empirically tested, whereas the homogeneity (in prices) of profit function is definitional and, therefore, not something to be tested.

manipulations, we have

$$\begin{aligned}\ln\left(\frac{\pi^a}{p}\right) &= \ln\pi(\boldsymbol{w}/pe^{-u},\boldsymbol{q}) = \beta_0 + \sum_j \beta_j \ln\left(\frac{w_j}{p}\right) + \sum_q \beta_q \ln q_q \\ &\quad + \frac{1}{2}\sum_j\sum_k \beta_{jk}\ln\left(\frac{w_j}{p}\right)\ln\left(\frac{w_k}{p}\right) \\ &\quad + \frac{1}{2}\sum_q\sum_s \gamma_{qs}\ln q_q \ln q_s + \sum_j\sum_q \delta_{jq}\ln\left(\frac{w_j}{p}\right)\ln q_q \\ &\quad + \left[-1 + \sum_j \beta_j + \sum_j\sum_q \delta_{jq}\ln q_q + \sum_j\sum_k \beta_{jk}\ln\left(\frac{w_j}{p}\right)\right] u + \left[\frac{1}{2}\sum_j\sum_k \beta_{jk}\right] u^2 \\ &\equiv \ln\pi(\boldsymbol{w}/p,\boldsymbol{q}) + \ln u(\boldsymbol{w}/p,\boldsymbol{q},u).\end{aligned} \tag{4.22}$$

Note that the penultimate line in (4.22), involving u and u^2, is the $\ln u(\cdot)$ function in (4.17) that represents profit inefficiency.[19] Thus, profit inefficiency depends not only on u but also on prices and quasi-fixed inputs. Consequently, profit inefficiency cannot be assumed to have a constant mean and variance (irrespective of its distribution).

Without further assumptions, this model is difficult to estimate, because of the presence of the u^2 term. If we make the assumption that the underlying production function is homogenous, then additional parameter restrictions apply, viz.,

$$\sum_k \beta_{jk} = 0, \quad \forall\, j,$$
$$\sum_j \delta_{jq} = 0, \quad \forall\, q.$$

These restrictions simplify the profit function of (4.22) substantially. The last line becomes $\ln u(\boldsymbol{w}/p,\boldsymbol{q},u) = -\left[1-\sum_j \beta_j\right] u$. In addition, the restrictions also simplify the deterministic

[19]See Kumbhakar (2001) for further details on the properties of profit functions with both technical and allocative inefficiency.

part of the profit function of (4.22) (i.e., the first three lines). The end result is

$$\begin{aligned}
\ln\left(\frac{\pi^a}{p}\right) &= \beta_0 + \frac{-r}{1-r}\ln(w_1/p) + \sum_j \beta_j \ln\tilde{w}_j + \sum_q \beta_q \ln q_q + \frac{1}{2}\sum_j\sum_k \beta_{jk}\ln\tilde{w}_j\ln\tilde{w}_k \\
&\quad + \frac{1}{2}\sum_q\sum_s \gamma_{qs}\ln q_q \ln q_s + \sum_j\sum_q \delta_{jq}\ln\tilde{w}_j \ln q_q - \tilde{u}, \\
&= \frac{-r}{1-r}\ln(w_1/p) + \ln G(\tilde{\boldsymbol{w}}, \boldsymbol{q}) - \tilde{u} \\
&\equiv \beta_0 + \alpha_1 \ln(w_1/p) + \sum_j \beta_j \ln(w_j/p) + \sum_q \beta_q \ln q_q + \frac{1}{2}\sum_j\sum_k \beta_{jk}\ln\tilde{w}_j\ln\tilde{w}_k \\
&\quad + \frac{1}{2}\sum_q\sum_s \gamma_{qs}\ln q_q \ln q_s + \sum_j\sum_q \delta_{jq}\ln\tilde{w}_j \ln q_q - \tilde{u}, \qquad\qquad j,k = 2,\ldots J.
\end{aligned} \tag{4.23}$$

where $\alpha_1 = -r/(1-r) - \sum_j \beta_j$, $\ln\tilde{w}_j = \ln(w_j/p) - \ln(w_1/p),\ j,k = 2,\ldots J$, and

$$\tilde{u} = u\left[1 - \sum_j \beta_j\right] \geq 0. \tag{4.24}$$

As expected, we get Lau's (1978) result starting from the translog cost function and imposing homogeneity restrictions on the underlying production function.

For profit maximizing firms operating in a competitive market, returns to scale (r) is less than unity. From the homogeneity of $G(w)$ we get $\sum_j \beta_j = -r/(1-r)$. This implies that $1 - \sum_j \beta_j = (1+r/(1-r)) = 1/(1-r) > 0$ (because $r < 1$). Since $\tilde{u}$ measures the difference of the log of maximum profit and the log of the actual profit, $100\times\tilde{u}$ is the percentage by which the profit is foregone due to technical inefficiency. It is worth noting that since $\tilde{u} \neq u$, a one percent increase in output-oriented technical efficiency does not translate into a one percent increase profit (i.e., $\partial\ln\pi/\partial\ln\{e^{-u}\} \neq 1$). For a marginal change in u, the profit is increased by $\left[1 - \sum_j \beta_j\right] = -\partial\ln\pi/\partial u = 1/(1-r) > 1$ percent. Thus, the higher the value of r, the greater is the potential for profit to be increased from increased efficiency.

One can also measure profit efficiency directly from

$$e^{-\tilde{u}} = \frac{\pi^a}{\pi(\boldsymbol{w}/p, \boldsymbol{q})}, \qquad \tilde{u} \geq 0,$$

which measures the ratio of actual profit to maximum profit, and has a value between 0 and 1. Assuming π^a is positive, the exact percentage loss of profit can be computed as $\{1 - e^{-\tilde{u}}\}\times 100$. Furthermore, once u is computed from $\tilde{u}$, which is easily done for a homogeneous function, one can compute both profit inefficiency (profit loss) and technical inefficiency (output loss). That is, one can switch from profit loss to output loss and vice-versa. However, this may not be that easy once we deal with nonhomogeneous production technologies.

4.4. **Input-Oriented Technical Inefficiency for the Profit Frontier.** Now we consider profit maximization behavior with input-oriented representation of technical inefficiency, viz.,

$$\max_{y,\boldsymbol{x}} \quad py - \boldsymbol{w}'\boldsymbol{x} \qquad \text{s.t.} \qquad y = m(\boldsymbol{x}e^{-\eta}, \boldsymbol{q}),$$

for which the FOCs are

$$p \cdot m_j(\boldsymbol{x}e^{-\eta}, \boldsymbol{q})e^{-\eta} = w_j \Rightarrow m_j(\boldsymbol{x}e^{-\eta}, \boldsymbol{q}) = \frac{w_j e^{\eta}}{p},$$

where $\eta \geq 0$ is the input-oriented technical inefficiency, $\boldsymbol{q}$ is a vector of quasi-fixed inputs, and $m_j(\cdot)$ is the marginal product of x_j (partial derivative of $m(\cdot)$ with respect to x_j).

Note that if we replace $xe^{-\eta}$ by $\tilde{x}$ and we^{η} by $\tilde{w}$, then we are back to the familiar neo-classical framework, i.e.,

$$\max_{y,\tilde{x}} \quad py - \tilde{\boldsymbol{w}}'\tilde{\boldsymbol{x}} \qquad \text{s.t.} \quad y = m(\tilde{\boldsymbol{x}}, \boldsymbol{q}).$$

Solving the system, the input demand of $\tilde{x}_j$ and the output supply of y will both be functions of $\tilde{\boldsymbol{w}}$, p, and $\boldsymbol{q}$. Therefore, the profit function is $\pi(\boldsymbol{w}e^{\eta}, \boldsymbol{q}, p)$, which is also the actual profit, π^a. That is,

$$\pi^a = \pi(\boldsymbol{w}e^{\eta}, \boldsymbol{q}, p) = \pi(\boldsymbol{w}, \boldsymbol{q}, p) \cdot g(\boldsymbol{w}, \boldsymbol{q}, p, \eta) \Longrightarrow \ln \pi^a = \ln \pi(\boldsymbol{w}, \boldsymbol{q}, p) + \ln g(\boldsymbol{w}, \boldsymbol{q}, p, \eta),$$

where $\pi(\boldsymbol{w}, \boldsymbol{q}, p) \equiv \pi(\boldsymbol{w}e^{\eta}, \boldsymbol{q}, p) \mid_{\eta=0}$ is the profit frontier in the absence of technical inefficiency, and $g(\cdot) = \pi(\boldsymbol{w}e^{\eta}, \boldsymbol{q}, p)/\pi(\boldsymbol{w}, \boldsymbol{q}, p) \leq 1$ is profit efficiency.

It can be shown that if the production function is homogeneous, the corresponding profit function can be expressed as

$$\ln \pi^a = \ln \pi(\boldsymbol{w}, \boldsymbol{q}, p) + \ln G(\boldsymbol{w}, \boldsymbol{q}, p, \eta) = \frac{1}{1-r} \ln p + \ln G(\boldsymbol{w}) - \frac{r}{1-r}\eta, \tag{4.25}$$

where $G(\boldsymbol{w}, \boldsymbol{q}, p, \eta)$ is homogeneous of degree $-r/(1-r)$. A comparison of (4.18) and (4.25) shows that $u = r \cdot \eta$. That is, under the assumption of a homogeneous production function, the output-oriented technical inefficiency (u) is equal to the input-oriented technical inefficiency (η) multiplied by returns to scale (r).

Now we consider a translog form on $\pi(\boldsymbol{w}e^{\eta}, \boldsymbol{q}, p)$, and we impose the price homogeneity assumption by normalizing the profit and the price variables by p. The result is

$$\begin{aligned} \ln\left(\frac{\pi^a}{p}\right) =& \beta_0 + \sum_j \beta_j \ln\left(\frac{w_j}{p}\right) + \sum_q \beta_q \ln q_q + \frac{1}{2}\sum_j \sum_k \beta_{jk} \ln\left(\frac{w_j}{p}\right) \ln\left(\frac{w_k}{p}\right) \\ &+ \frac{1}{2}\sum_q \sum_s \gamma_{qs} \ln q_q \ln q_s + \sum_j \sum_q \delta_{jq} \ln\left(\frac{w_j}{p}\right) \ln q_q \\ &+ \left[\sum_j \beta_j + \sum_j \sum_q \delta_{jq} \ln q_q + \sum_j \sum_k \beta_{jk} \ln\left(\frac{w_j}{p}\right)\right] \eta + \left[\frac{1}{2}\sum_j \sum_k \beta_{jk}\right] \eta^2. \end{aligned} \tag{4.26}$$

Without further assumptions, this model is difficult to estimate because both η and η^2 are in the model. If we assume that the underlying production function is homogenous, then additional parameter restrictions apply, viz., $\sum_k \beta_{jk} = 0 \; \forall \, j$, and $\sum_q \delta_{jq} = 0 \; \forall \, j$. These restrictions simplify the profit function in (4.26) substantially. The last two lines reduce to $\left[\sum_j \beta_j\right] \eta$. The assumption of homogeneous production function thus simplifies the profit frontier model as follows:

$$\ln\left(\frac{\pi^a}{p}\right) = \ln \pi(\boldsymbol{w}/p, \boldsymbol{q}) - \tilde{\eta}, \tag{4.27}$$

where $\tilde{\eta} = -\eta \; \sum_j \beta_j = \frac{r}{1-r}\eta \geq 0$ and

$$\begin{aligned}\ln \pi(\boldsymbol{w}/p, \boldsymbol{q}) =& \beta_0 + \sum_j \beta_j \ln\left(\frac{w_j}{p}\right) + \sum_q \beta_q \ln q_q \\ &+ \frac{1}{2}\sum_j \sum_k \beta_{jk} \ln(\omega_j) \ln(\omega_k) + \frac{1}{2}\sum_q \sum_s \gamma_{qs} \ln q_q \ln q_s.\end{aligned}$$

In the above expression $\omega_j = \frac{w_j/p}{w_J/p} = \frac{w_j}{w_J}$ and $\omega_k = \frac{w_k/p}{w_J/p} = \frac{w_k}{w_J}$.

As shown in (4.19) and (4.27), the profit functions associated with OO and IO technical inefficiency are observationally equivalent. Note that $\ln G(\boldsymbol{w}/p, \boldsymbol{q})$ in (4.19) is the same as $\ln \pi(\boldsymbol{w}/p, \boldsymbol{q})$ in (4.27).

The effect of IO technical inefficiency on profit can be measured using $\tilde{\eta}$ and $e^{-\tilde{\eta}}$, as we have seen in other cases. That is, $100 \times \tilde{\eta}$ gives the percentage of profit loss due to the technical inefficiency. Again, since $\tilde{\eta} \neq \eta$, a one percent decrease in input-oriented technical inefficiency does not translate into one percent increase in profit (i.e., $\partial \ln \pi / \partial \ln\{(\eta)\} \neq 1$). The percentage increase in profit is $r/(1-r) \gtreqless 1$ depending on the value of the returns to scale parameter, r. Instead of measuring inefficiency, one can measure of profit efficiency from $e^{-\tilde{\eta}}$, which is the ratio of actual to maximum profit.

We have noted the difference (in interpretation) between input and output-oriented measures of technical inefficiency. Similar differences are observed if one examines their impact on profit. However, it is important to remind the reader that it does not matter whether one uses the input or the output-oriented measure. The estimation is exactly the same regardless of the inefficiency orientation. Furthermore, one can switch from u to η and vice versa. This is not as simple if we dispense with the homogeneity assumption on the production technology.

4.5. **Estimation of Technical Efficiency Using the Full System.** Here we consider cases where outputs are exogenously given to a firm. A cost-minimizing firm chooses the levels of inputs in order to produce the given level of output with the lowest possible cost. When output and input prices are exogenously given to a firm, cost minimization is equivalent to profit maximization. Because inputs are endogenous in the cost minimization framework, input-oriented (as opposed to output-oriented) technical inefficiency is usually chosen as the preferred approach to model technical inefficiency.

In the standard neoclassical context, Christensen & Greene (1976) proposed using a system approach, consisting of the cost function and the cost share equations, to estimate the cost function parameters. Here we consider a similar system, but we allow producers to be technically inefficient. That is, compared to the cost function which consists of only the cost function adjusted for technical inefficiency, here we include the cost share equations and form a system to estimate the cost function parameters. The use of these share equations does not require any additional assumptions, and the share equations do not contain any new parameters that are not already in the cost function (other than the parameters associated with the errors in the cost share equations). Thus, the additional information provided by these share equations helps in estimating the parameters more precisely.

In addition to the efficiency gain for the parameter estimators via inclusion of the share equations there is another important advantage in the current context: Residuals of the share equations may be interpreted as allocative inefficiency (or functions of them). This issue will be discussed later.

4.5.1. *Single Output, Input-Oriented Technical Inefficiency.* A producer's cost minimization problem with input-oriented technical inefficiency is

$$\min_{\boldsymbol{x}} \ \boldsymbol{w}'\boldsymbol{x} \quad \text{s.t.} \quad y = m(\boldsymbol{x}e^{-\eta}), \tag{4.28}$$

which gives the following first order conditions (FOCs):

$$\text{FOCs:} \qquad \frac{m_j(\boldsymbol{x}e^{-\eta})}{m_1(\boldsymbol{x}e^{-\eta})} = \frac{w_j}{w_1}, \quad j = 2, \ldots, J, \tag{4.29}$$

where $\eta \geq 0$ is the input-oriented technical inefficiency, and $m_j(\cdot)$ is the partial derivative of $m(\cdot)$ with respect to x_j.

The $J-1$ FOCs in (4.29) with the production function in (4.28) can be used to solve for J input demand functions in the form $x_j e^{-\eta}$, $j = 1, \ldots, J$. That is, $x_j e^{-\eta}$ becomes a function of $\boldsymbol{w}$ and y, viz., $x_j e^{-\eta} = \psi_j(\boldsymbol{w}, y)$. We can use these to define the pseudo cost function given $\boldsymbol{w}$ and y as

$$C^*(\boldsymbol{w}, y) = \sum_j w_j x_j e^{-\eta}, \tag{4.30}$$

which can be viewed as the minimum cost function for the following problem:

$$\min_{\boldsymbol{x}e^{-\eta}} \ \boldsymbol{w}'\boldsymbol{x}e^{-\eta} \quad \text{s.t.} \quad y = m(\boldsymbol{x}e^{-\eta}).$$

The $C^*(\cdot)$ function is also the *frontier cost function*, which gives the minimum cost to produce the observed level of output, y, given the input prices, $\boldsymbol{w}$. Due to technical inefficiency, the observed input use ($\boldsymbol{x}$) is *too high*, in the sense that a fully efficient producer can produce the same level of output with less inputs. The efficient input use would be $\boldsymbol{x}e^{-\eta}$, which is less than the observed input $\boldsymbol{x}$ (since $0 \leq e^{-\eta} \leq 1$), and the efficient minimum cost would be $\boldsymbol{w}'\boldsymbol{x}e^{-\eta}$, which is less than the actual cost $\boldsymbol{w}'\boldsymbol{x}$.

We apply Shephard's lemma to (4.30), i.e.,

$$\frac{\partial C^*}{\partial w_j} = x_j e^{-\eta} \Longrightarrow \frac{\partial \ln C^*}{\partial \ln w_j} = \frac{w_j x_j e^{-\eta}}{C^*}.$$

Since the actual cost is $C^a = \sum_j w_j x_j = e^{\eta} \sum_j w_j x_j e^{-\eta} = C^* e^{\eta}$, we have

$$\ln C^a(\boldsymbol{w}, y, \eta) = \ln C^*(\boldsymbol{w}, y) + \eta, \tag{4.31}$$

and the actual cost share S_j of input j is

$$S_j = \frac{w_j x_j}{C^a} = \frac{w_j x_j}{C^* e^{\eta}} = \frac{\partial \ln C^*}{\partial \ln w_j}, \quad j = 1, \ldots, J. \tag{4.32}$$

The cost frontier model is based on estimation of (4.31) alone. For the purpose of estimating model parameters, a system of equations is formed consisting of (4.31) and the $J-1$ share equations from (4.32). We drop one of the share equations (which share equation is dropped is inconsequential for the analysis) and include only $J-1$ of them because $\sum_j S_j = 1$ by construction. A random statistical error, ζ_j, is also added to the jth share equation $j = 2, \ldots, J$ for estimation purposes. The system of equations is thus

$$\ln C^a(\boldsymbol{w}, y, \eta) = \ln C^*(\boldsymbol{w}, y) + \eta, \tag{4.33}$$

$$S_j = \frac{\partial \ln C^*}{\partial w_j} + \zeta_j, \qquad j = 2, \ldots, J. \tag{4.34}$$

It is worth pointing out here that we do not give ζ_j a structural interpretation except that it is a statistical error. Later we show that ζ_j can be interpreted as a function of allocative inefficiency. For now, we assume that every producer is allocatively efficient.

Because the cost function is homogeneous of degree 1 in input prices, the parameters in (4.33) and (4.34) need to satisfy the homogeneity property. In addition, since the derivative of $\ln C^*(\boldsymbol{w}, y)$ appears on the right-hand-side of (4.34) and thus the parameters of $\ln C^*(\boldsymbol{w}, y)$ also appear in (4.34). Consequently, cross equation constraints on the parameters of (4.33) and (4.34) have to be imposed in the estimation process.

Notice that the share equations in (4.34) introduce no new parameter to the model except those associated with ζ_i's which are *usually* not of interest. If only technical inefficiency is considered, the advantage of estimating the system of equations in (4.33) and (4.34), as opposed to the single equation in (4.33), is the gain in efficiency (i.e., more precise parameter estimates). The gain, however, comes at a cost which involves the estimation of a complicated system. On the other hand, if both technical and allocative inefficiency are considered in the model and one wants to decompose technical and allocative inefficiency, it is necessary to use the system equation approach. We will discuss the cost frontier model with both technical and allocative inefficiency later.

4.5.1.1. **Maximum Likelihood Estimation**

In estimating the above cost function using OLS, we treated technical inefficiency (η) as the noise term. Now we separate it from the noise terms, v, which is added to the (log) cost frontier function. The two error components are identified by imposing distributional assumptions on them. The likelihood function can be derived based on the distributional assumptions, and model parameters are then estimated numerically by maximizing the log-likelihood function. Following this approach, the system of equations is

$$\ln C^a = \ln C^*(\boldsymbol{w}, y) + \eta + v, \tag{4.35}$$

$$S_j = \frac{\partial \ln C^*}{\partial \ln w_j} + \zeta_j, \qquad j = 2, \ldots, J. \tag{4.36}$$

If the translog specification is adopted, then the system of equations become

$$\begin{aligned}
\ln\left(\frac{C^a}{w_1}\right) =& \ln C^*(\boldsymbol{w}/w_1, \boldsymbol{q}, y) + \eta + v \\
=& \beta_0 + \sum_{j=2} \beta_j \ln\left(\frac{w_j}{w_1}\right) + \sum_r \gamma_r \ln q_r + \beta_y \ln y \\
&+ \frac{1}{2}\left[\sum_{j=2}\sum_{k=2} \beta_{jk} \ln\left(\frac{w_j}{w_1}\right)\ln\left(\frac{w_k}{w_1}\right) + \sum_r\sum_s \gamma_{rs}\ln q_r \ln q_s + \beta_{yy}\ln y \ln y\right] \\
&+ \sum_{j=2}\sum_r \delta_{jr} \ln\left(\frac{w_j}{w_1}\right)\ln q_r + \sum_{j=2}\beta_{jy}\ln\left(\frac{w_j}{w_1}\right)\ln y + \sum_r \gamma_{ry}\ln q_r \ln y + \eta + v,
\end{aligned} \tag{4.37}$$

$$S_j = \beta_j + \sum_{k=2} \beta_{jk} \ln\left(\frac{w_k}{w_1}\right) + \sum_r \delta_{jr}\ln q_r + \beta_{jy}\ln y + \zeta_j, \qquad j = 2, \ldots, J. \tag{4.38}$$

The above system is similar to (4.33) and (4.34). The difference being the v term in the cost frontier.

Estimation proceeds once distributional assumptions on the noise and inefficiency terms are in place. In is common to assume that η is half normal while $\boldsymbol{\xi} \sim N(\mathbf{0}, \Omega)$ where Ω is the $J \times J$ covariance matrix of $\boldsymbol{\xi} = (v, \zeta_2, \ldots, \zeta_J)'$. The elements of ξ are assumed to be independent of η.

Denote $Z = \boldsymbol{d}\eta + \boldsymbol{\xi}$ where $\boldsymbol{d} = (1, 0, \ldots, 0)'$, which is a column vector of $J \times 1$. Based on the above distributional assumptions on η and $\boldsymbol{\xi}$, derivation of the above log-likelihood function follows the usual procedure. Since both η and $\boldsymbol{\xi}$ are *i.i.d.* across firms, we drop the firm subscript in the following derivation. The pdf of Z, $f(Z)$, can be expressed as

$$f(Z) = \int_0^\infty f(Z, \eta) d\eta = \int_0^\infty f(Z|\eta) h(\eta) d\eta, \tag{4.39}$$

where $f(Z,\eta)$ is the joint pdf of Z and η, and $h(\eta)$ is the pdf of η. Using the distributional assumptions on η and $\boldsymbol{\xi}$, the above integral can be expressed as

$$f(Z) = \frac{2}{(2\pi)^{(J+1)/2}|\Omega|^{1/2}\sigma_u}\int_0^\infty \exp\Big\{-\frac{1}{2}[(Z-\boldsymbol{d}\,\eta)'\Omega^{-1}(Z-\boldsymbol{d}\,\eta)+\eta^2/\sigma_u^2]\Big\}d\eta$$
$$= \frac{2\,\sigma\,e^{-a/2}}{(2\pi)^{(J/2)}|\Omega|^{1/2}\sigma_u}\Phi(Z'\Omega^{-1}\boldsymbol{d}\,\sigma). \tag{4.40}$$

The log-likelihood function for a sample of N firms can then be written as (Kumbhakar 2001)

$$\mathcal{L} = -N/2\,\ln|\Omega| + N\ln\sigma + \sum_{i=1}^{N}\Phi(Z_i'\Omega^{-1}\boldsymbol{d}\,\sigma) - N\ln\sigma_u(1/2)\sum_{i=1}^{N}a_i, \tag{4.41}$$

where $\sigma^2 = (1/\sigma_u^2 + \boldsymbol{d}'\Omega^{-1}\boldsymbol{d})^{-1}$ and $a_i = Z_i'\Omega^{-1}Z_i - \sigma^2(Z_i'\Omega^{-1}\boldsymbol{d})^2$.

The likelihood function is derived allowing arbitrary correlations among all the error terms, i.e., v and ξ_j are freely correlated. We consider some special cases of this by imposing constrains on Ω. For example, if

$$\Omega = \begin{pmatrix}\sigma_v^2 & \mathbf{0}\\ \mathbf{0}' & \Sigma\end{pmatrix}, \tag{4.42}$$

then the error in the cost function is uncorrelated with the errors in the cost share equations $\boldsymbol{\zeta}$ for which the variance covariance matrix is Σ. Further, if Σ is diagonal then all the errors are independent of each other.

Maximization of the log-likelihood in (4.41) gives consistent estimates of the parameters in the cost function as well as those in Ω and σ_u^2. The estimated parameters can be used to estimate η. Since the conditional mean of η given Z is $N(Z'\Omega^{-1}d\sigma^2, \sigma^2)$ truncated at zero from below ($\eta \geq 0$), we can use the Jondrow et al. (1982) type formula to estimate η, which is

$$\hat{\eta} = E(\eta|Z) = \tilde{\mu} + \sigma\frac{\phi(\tilde{\mu}/\sigma)}{\Phi(\tilde{\mu}/\sigma)}, \tag{4.43}$$

where $\tilde{\mu} = Z'\Omega^{-1}d\sigma^2$.

Alternatively, one can estimate technical efficiency from

$$\widehat{TE} = E(e^{-\eta}|Z)) = \frac{1-\Phi[\sigma-(\tilde{\mu}/\sigma)]}{\Phi(\tilde{\mu}/\sigma)}\,e^{-\tilde{\mu}+(1/2)\sigma^2}. \tag{4.44}$$

4.5.2. *Multiple Outputs With Input-Oriented Technical Inefficiency.* The model in the previous section assumes that there is a single output in the production. In reality, firms often produce multiple outputs. If these multiple outputs are not taken into account then clearly the estimated inefficiency may be biased. Thus, in this section we examine how to empirically model multiple outputs.

The cost minimization problem of a firm with multiple outputs is

$$\min_{\boldsymbol{x}} : \boldsymbol{w}'\boldsymbol{x} \qquad \text{s.t.} \qquad M(\boldsymbol{y}, \boldsymbol{x}e^{-\eta}) = 0. \tag{4.45}$$

Note that we define the production possibility function as $M(\boldsymbol{y}, \boldsymbol{x}e^{-\eta}) = 0$. In the single output case this representation is simply $y - m(\boldsymbol{x}e^{-\eta}) = 0$. Thus, the technology specification is different from the single output model. The rest of the derivation is straightforward and similar to the derivation of a single output model.

The first order condition of the above minimization problem is

$$FOCs : \quad \frac{w_j}{w_1} = \frac{\frac{\partial M}{\partial x_j e^{-\eta}}}{\frac{\partial M}{\partial x_1 e^{-\eta}}}, \qquad j = 2, \ldots, J.$$

The (inefficiency adjusted) input demand functions are $x_j^* = x_j e^{-\eta} = \phi(\boldsymbol{w}, \boldsymbol{y})$. If we define the pseudo cost function as

$$C^*(\boldsymbol{w}, \boldsymbol{y}) = \sum_j w_j x_j e^{-\eta},$$

then it can be viewed as the minimum cost of the following problem:

$$\min_{\boldsymbol{x}e^{-\eta}} \ \boldsymbol{w}'\boldsymbol{x}e^{-\eta} \quad \text{s.t.} \quad M(\boldsymbol{y}, \boldsymbol{x}e^{-\eta}) = 0.$$

Also note that $C^*(\boldsymbol{w}, \boldsymbol{y})$ is the cost frontier because it gives the minimum cost of producing $\boldsymbol{y}$ with input price $\boldsymbol{w}$.

Applying Shephard's lemma, and after some manipulation, we have

$$S_j = \frac{w_j x_j}{C^a} = \frac{w_j x_j}{C^* e^{\eta}} = \frac{\partial \ln C^*}{\partial \ln w_j}, \quad j = 1, \ldots, J, \tag{4.46}$$

which is exactly the same as in the single output case.

Since the actual cost is $C^a = \sum_j w_j x_j = C^* e^{\eta}$, we have

$$\ln C^a = \ln C^*(\boldsymbol{w}, \boldsymbol{y}) + \eta. \tag{4.47}$$

After imposing a functional form on $\ln C^*(\boldsymbol{w}, \boldsymbol{y})$, the above equation can be estimated to obtain parameters of the cost function and the inefficiency index. This is similar to the single equation cost frontier estimation introduced earlier. We can also use the share equation in estimation. To do so, we append random errors to the share equations in (4.46) and obtain

$$S_j = \frac{\partial \ln C^*}{\partial \ln w_j} + \zeta_j, \qquad j = 2, \ldots, J. \tag{4.48}$$

We may use the system of equations consisting (4.47) and (4.48) to estimate the parameters in the model.

Now we consider an example in which $\ln C^*(\boldsymbol{w}, \boldsymbol{y})$ is translog and there are quasi-fixed inputs, $\boldsymbol{q}$, in the production process. The model is

$$\begin{aligned}\ln C^a =& \ln C^*(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{y}) + \eta \\ =& \beta_0 + \sum_j \beta_j \ln w_j + \sum_r \gamma_r \ln q_r + \sum_m \theta_m \ln y_m \\ &+ \frac{1}{2}\left[\sum_j \sum_k \beta_{jk} \ln w_j \ln w_k + \sum_r \sum_s \gamma_{rs} \ln q_r \ln q_s + \sum_m \sum_n \theta_{mn} \ln y_m \ln y_n\right] \\ &+ \sum_j \sum_r \delta_{jr} \ln w_j \ln q_r + \sum_j \sum_m \phi_{jm} \ln w_j \ln y_m + \sum_r \sum_m \gamma_{rm} \ln q_r \ln y_m + \eta,\end{aligned} \tag{4.49}$$

$$S_j = \beta_j + \sum_{k=1} \beta_{jk} \ln w_k + \sum_r \delta_{jr} \ln q_r + \sum_m \phi_{jm} \ln y_m + \zeta_j, \qquad j = 2, \ldots, J. \tag{4.50}$$

When it comes to estimation, the multiple output model is no different from the single output model. Note that a stochastic noise component, v, has to be added to (4.49).

4.5.3. *Cost Minimization: Multiple Outputs With Output-Oriented Technical Inefficiency.* Most of the models in the efficiency literatures assume a single output. However, in reality most of the firms produce more than one output. One can justify the use of a single output model by aggregating all the outputs. For example, in the case of airlines, passenger and cargo outputs can be aggregated into a single output, such as revenue, which is not a physical measure of output. However, this might cause other problems associated with aggregation and prices. Furthermore, a single output model cannot capture substitutability between outputs. In some other cases production of an intended output (such as electricity) also produces an unintended output (pollution) and it does not make much sense to aggregate intended and unintended outputs because they have different properties. Intended output can be freely disposed but the unintended outputs cannot. Ignoring unintended outputs from the model is likely to give incorrect estimates of inefficiency. Similarly, in the case of airlines, if one models only passenger output, efficiency estimates are likely to be incorrect because the omitted output cargo will cause omitted variable bias. In this section we examine how to estimate a model with multiple outputs.

For the cost minimization problem the first order conditions are

$$\begin{aligned}&\min_{\boldsymbol{x}} : \boldsymbol{w}'\boldsymbol{x} \qquad \text{s.t.} \qquad M(\boldsymbol{y}e^u, \boldsymbol{x}) = 0, \\ &\text{FOCs :} \qquad \frac{w_j}{w_1} = \frac{\partial M}{\partial x_j} \Big/ \frac{\partial M}{\partial x_1}, \qquad j = 2, \ldots, J.\end{aligned}$$

The input demands are $x_j = x_j(\boldsymbol{w}, \boldsymbol{y}e^u)$. The minimum cost is

$$C^* = \sum_j w_j x_j(\boldsymbol{w}, \boldsymbol{y}e^u) = C^*(\boldsymbol{w}, \boldsymbol{y}e^u),$$

which can be viewed as the minimum cost of the following problem:

$$\min_{\boldsymbol{x}} \ \boldsymbol{w}'\boldsymbol{x} \quad \text{s.t.} \quad M(\boldsymbol{y}e^u, \boldsymbol{x}) = 0. \tag{4.51}$$

Applying Shephard's lemma, and after some manipulations, we have

$$S_j = \frac{w_j x_j}{C^* e^\eta} = \frac{w_j x_j}{C^a} = \frac{\partial \ln C^*}{\partial \ln w_j}, \quad j = 1, \ldots, J,$$

which is exactly the same as in the single output case.

The system of equations is represented by

$$\ln C^a(\boldsymbol{w}, \boldsymbol{y}, u) = \ln C^*(\boldsymbol{w}, \boldsymbol{y}e^u), \tag{4.52}$$

$$S_j = \frac{\partial \ln C^*}{\partial \ln w_j} + \zeta_j, \qquad j = 2, \ldots, J, \tag{4.53}$$

where ζ_j is the random variable of statistical error appended to the jth share equation.

Consider an example where $\ln C^*(\boldsymbol{w}, \boldsymbol{y}e^u)$ is translog. The system of equations for it is

$$\begin{aligned}
\ln C^a =& \ln C^*(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{y}e^u) \\
=& \beta_0 + \sum_j \beta_j \ln w_j + \sum_r \gamma_r \ln q_r + \sum_m \theta_m (\ln y_m + u) \\
&+ \frac{1}{2}\left[\sum_j \sum_k \beta_{jk} \ln w_j \ln w_k + \sum_r \sum_s \gamma_{rs} \ln q_r \ln q_s + \sum_m \sum_n \theta_{mn} (\ln y_m + u)(\ln y_n + u)\right] \\
&+ \sum_j \sum_r \delta_{jr} \ln w_j \ln q_r + \sum_j \sum_m \phi_{jm} \ln w_j (\ln y_m + u) + \sum_r \sum_m \gamma_{rm} \ln q_r (\ln y_m + u),
\end{aligned} \tag{4.54}$$

$$S_j = \beta_j + \sum_{k=1} \beta_{jk} \ln w_k + \sum_r \delta_{jr} \ln q_r + \sum_m \phi_{jm} (\ln y_m + u) + \zeta_j, \qquad j = 2, \ldots, J. \tag{4.55}$$

It is clear that the above system is quite complicated. First, the inefficiency u appears nonlinearly (it involves u^2). Second, inefficiency also appears in the cost share equations. Derivation of the likelihood function for this model is quite complex. In fact, it is not possible to get a closed form expression for the likelihood function. Use of maximum simulated likelihood is an option though we know of no empirical work deploying this method.

5. Determinants of Inefficiency

5.1. **The Impact of Exogenous Influences on the Stochastic Frontier Model.** After the introduction of the stochastic frontier model and the conditional mean estimator of firm level inefficiency, the natural evolution of the model was to use it to understand inefficiency. Is firm level inefficiency dependent upon observable characteristics and if so how should this relationship be modeled in the context of a stochastic frontier? Is the distribution of inefficiency heteroscedastic and what are the implications of this feature? The first models that looked at the behavior (features) of firm level inefficiency pertain to the panel data specifications of Kumbhakar (1987) and Battese & Coelli (1992) (see Section 7). A related model that investigates the components/features of the distribution of inefficiency (in a cross section of firms) is Deprins (1989) and Deprins & Simar (1989*a*, 1989*b*), however, this model does not allow for noise.

To consider the implications of an attempt to determine what influences firm level inefficiency consider the original normal, half-normal model of Aigner et al. (1977), which assumes that both v_i and u_i are homoscedastic, i.e., both σ_v^2 and σ_u^2 are constant. In a traditional linear regression framework heteroscedasticity has no impact on the bias/consistency of the corresponding parameter estimators. However, if we allow σ_u^2 to depend on a set of variables, which we will refer to as $\boldsymbol{z}$, then ignoring this relationship will lead to biased and inconsistent parameter estimators of all model parameters, except in special settings. Kumbhakar & Lovell (2000, Section 3.4) (see also Wang & Schmidt 2002) provide a detailed discussion on the consequences of ignoring heteroscedasticity in the stochastic frontier. A concise description of their discussion, assuming that v_i and u_i are heteroscedastic is:

- Ignoring the heteroscedasticity of v_i still gives consistent estimates of the frontier function parameters ($\boldsymbol{\beta}$) except for the intercept, which is downward biased. Estimates of technical efficiency are biased.
- Ignoring the heteroscedasticity of u_i causes biased estimates of the frontier function parameters as well as the estimates of technical efficiency.

To understand the second point more clearly note that from Section 2, $E[u] = \sqrt{2/\pi}\sigma_u$. If we eschew distributional assumptions and estimate the production frontier via OLS, then we have that only the OLS intercept is biased. However, if $\sigma_u^2 = \sigma_u^2(\boldsymbol{z})$, then omission of this leads to biased parameter estimates of all model parameters via OLS given that the assumed model is

$$y_i = m(\boldsymbol{x}_i;\boldsymbol{\beta}) + \varepsilon_i^*,$$

whereas the true model is

$$y_i = m(\boldsymbol{x}_i;\boldsymbol{\beta}) + \sqrt{2/\pi}\sigma_u(\boldsymbol{z}_i) + \varepsilon_i^* \equiv \tilde{m}(\boldsymbol{x}_i,\boldsymbol{z}_i;\boldsymbol{\beta},\boldsymbol{\delta}) + \varepsilon_i^*.$$

In essence, the estimates of $m(\boldsymbol{x}_i;\boldsymbol{\beta})$ are conflated with $\sigma_u(\boldsymbol{z}_i)$, unless $\boldsymbol{x}$ and $\boldsymbol{z}$ are uncorrelated, which is the standard omitted variable misspecification. If some elements of $\boldsymbol{z}$ are actual inputs of production then there is no possibility to have zero correlation. The reason for this phenomena

is simply the fact that the mean and variance of v depend on different parameters, but given the truncation of u, it must be the case that the mean and variance depend on the same parameters, albeit in different fashions. Thus, it is not possible to allow u to be heteroscedastic without also allowing the mean of u to vary. Thus, care beyond the traditional stochastic frontier setup is required when the distribution of inefficiency is dependent upon a set of covariates.

Caudill & Ford (1993), Caudill, Ford & Gropper (1995), and Hadri (1999) contain the main (initial) proposals to specifically model heteroscedasticity via parameterization through a vector of observable variables, $\sigma_u^2(\boldsymbol{z};\boldsymbol{\delta}^u)$. For instance, $\sigma_{u,i}^2 = e^{g(\boldsymbol{z}_{u,i};\boldsymbol{\delta}^u)}$, where $\boldsymbol{z}_{u,i}$ is a $q \times 1$ vector of variables including a constant of 1, and $\boldsymbol{\delta}^u$ is the $q \times 1$ corresponding parameter vector. The exponential function is used to ensure a positive estimate of the variance parameter for all $\boldsymbol{z}$ and $\boldsymbol{\delta}^u$. A similar parameterization can be deployed to allow for potential heteroscedasticity in the noise term. Formally, their parameterizations are

$$\sigma_{u,i}^2 = e^{\boldsymbol{z}'_{u,i}\boldsymbol{\delta}^u}, \tag{5.1}$$

$$\sigma_{v,i}^2 = e^{\boldsymbol{z}'_{v,i}\boldsymbol{\delta}^v}. \tag{5.2}$$

The vectors $\boldsymbol{z}_{u,i}$ and $\boldsymbol{z}_{v,i}$ may or may not be equivalent, and they may also contain all or part of the $\boldsymbol{x}_i$ vector.

The log-likelihood function of the heteroscedastic model is the same as in (2.6), except that we replace σ_u^2 and σ_v^2 with (5.1) and (5.2), respectively, in the log-likelihood function.[20] In this framework all of the model parameters are estimated in a single step. With the parameter estimates in tow, technical inefficiency can be computed using (2.20) or (2.21) with the appropriate forms of σ_u^2 and σ_v^2 substituted into the expressions.

5.2. **Proper Modeling of the Determinants of Inefficiency.** Although the models of Caudill & Ford (1993), Caudill et al. (1995), and Hadri (1999) are motivated by the (possible) presence of heteroscedasticity in u_i, this model can also be recast as an attempt to discern the direct impact of exogenous determinants of the specific level of inefficiency for a given firm. In the classic, cross-sectional stochastic frontier model, all inefficiency is random and there is no explanation regarding its presence concomitant to observable features of the firm. In this section we detail how this extreme position of inefficiency can be transformed into a more realistic discussion whereby inefficiency (on average) is determined through observables.

For a stochastic frontier analysis, a researcher may not only want to know the level of inefficiency for each producer, but also which, if any, factors impact the level of inefficiency. For example, in studying efficiency within the banking industry, a researcher may want to know whether the inefficiency of a bank is affected by the use of information technology, the amount of assets the bank has access to, the type of bank, or the type of ownership structure in place. Similarly, the government might be interested in whether its regulations (such as allowing banks to merge)

[20]Actually, given the reparameterization of the log-likelihood function, the specifications for σ_u and σ_v imply specific specifications for λ and σ.

improves banks' performance. To answer these questions, we may want to estimate the relationship between inefficiency and possible determinants. The first research to tackle this issue is found in Kumbhakar, Ghosh & McGuckin (1991), Reifschneider & Stevenson (1991), Huang & Liu (1994), and Battese & Coelli (1995). Reifschneider & Stevenson (1991), in their discussion of allowing the variance of the inefficiency term to be a function of a set of variables, $\boldsymbol{z}$, use the term inefficiency explanatory variables. More contemporary parlance refers to $\boldsymbol{z}$ as z-variables or determinants of inefficiency.

At various points in time practitioners have deployed a simpler, two step analysis to model the influence of specific covariates on firm level inefficiency. This approach constructs estimates of observation-specific inefficiency via the Jondrow et al. (1982) conditional mean in the first step, and then regresses these inefficiency estimates on a vector of exogenous variables $\boldsymbol{z}_i$ in a second step. A negative coefficient of the exogenous variable in the regression is taken as indication that firms with larger values of the variables tend to have a lower level of inefficiency (i.e., they are more efficient). Pitt & Lee (1981) were the first to implement this type of approach, albeit in a slightly different form. Ali & Flinn (1989), Kalirajan (1990, appendix), and Bravo-Ureta & Rieger (1991), and many others followed this approach two-step. An even more recent example using the two-step approach is Wollni & Brümmer (2012, Tables 6 and 7). This method has no statistical merit and duplication of this approach should be avoided.

Criticisms against this two-step procedure have long pointed toward the biases that arise at various stages of the process; most prominently, the first stage model is misspecified (Battese & Coelli 1995). As explained in Wang & Schmidt (2002), if $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are correlated then the first step of the two-step procedure suffers from omitted variable bias (as we detailed above). Even when $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are uncorrelated, ignoring the dependence of the inefficiency on $\boldsymbol{z}_i$ will cause the estimated first-step technical efficiency index to be underdispersed, so that the results of the second-step regression are likely to be biased downward. The intuition underlying this result is easier to grasp if we begin by assuming that $\boldsymbol{x}$ and $\boldsymbol{z}$ are independent so that the first stage residual is an unbiased estimator for ε which is used to construct the JLMS estimate. The JLMS efficiency score, as discussed in Section 2, is a shrinkage estimator, shrinking towards the mean.[21] If we ignore the dependence of σ_u on $\boldsymbol{z}$ then we shrink the estimated level of inefficiency too much for firms with large u (and too little for firms with small u), resulting in less dependence of the JLMS scores on $\boldsymbol{z}$ than is otherwise present. This last feature implies that a second stage regression should produce downward biased estimates of the effects of $\boldsymbol{z}$ on u. Caudill & Ford (1993) provide Monte Carlo evidence of the bias of the model parameters from ignoring the impact of $\boldsymbol{z}$ on u while Wang & Schmidt (2002) provide Monte Carlo evidence of the bias on the second stage parameters.

Given the undesirable statistical properties of the two-step procedure, the preferred approach to studying the exogenous influences on efficiency is the single-step procedure. This procedure estimates the parameters of the relationship between inefficiency and $\boldsymbol{z}_i$ together with all the other

[21]Note that the use of the JLMS score already implies that there is less variance in the estimated inefficiency scores than is present in the actual inefficiency levels.

parameters of the model via maximum likelihood. While the initial discussion on the inclusion of determinants of inefficiency centered their discussion on the normal truncated-normal model, the point is valid regarding any distributional assumption on inefficiency; specifically, application to the half-normal model is straightforward.

5.3. **Marginal Effects of the Exogenous Determinants.** If u_i follows a half-normal distribution, i.e., $u_i \sim N^+(0, \sigma_u^2)$, and v_i is assumed to be homoscedastic, then σ_u^2 is the (only) parameter to be parameterized by the $\boldsymbol{z}_i$ vector. The parameterization function of (5.1) is well suited for this purpose. The mean of u_i is

$$E(u_i) = \sqrt{2/\pi} e^{\boldsymbol{z}_i'\boldsymbol{\delta}^u} = e^{\frac{1}{2}\ln(2/\pi) + \boldsymbol{z}_i'\boldsymbol{\delta}^u}. \tag{5.3}$$

Note that the $\frac{1}{2}\ln(2/\pi)$ term can be absorbed by the constant term in $\boldsymbol{z}_i'\boldsymbol{\delta}^u$. Therefore, by parameterizing σ_u^2, we allow the $\boldsymbol{z}_i$ variables to affect the expected value of inefficiency. More importantly, however, is that the parameterization (5.1) produces maximum likelihood estimates of $\boldsymbol{\delta}^u$ which may not be very informative. This is because the relationship between $E(u_i)$ and $\boldsymbol{z}_u$ is nonlinear, and so the slope coefficients $\boldsymbol{\delta}^u$ are *not* the marginal effects of $\boldsymbol{z}_u$. For instance, assume the kth variable in $\boldsymbol{z}_u$ has an estimated coefficient that is 0.5. This number itself tells us nothing about the magnitude of the kth variable's (marginal) effect on the inefficiency.

As such, the computation of the marginal effect of the z variables may be useful for empirical purposes. Given the half-normal assumption of u_i and the parameterization function of (5.1), the marginal effect of the kth variable of $\boldsymbol{z}_{u,i}$ on $E(u_i)$ can be computed as[22]

$$\frac{\partial E\left(u_i\right)}{\partial \boldsymbol{z}_u\left[k\right]} = \delta_k^u \sqrt{2/\pi}\sigma_{u,i} \tag{5.4}$$

where $\sqrt{2/\pi}$ is approximately 0.80.

Note that (5.4) also implies

$$\operatorname{sign}\left(\frac{\partial E(u_i)}{\partial \boldsymbol{z}_u[k]}\right) = \operatorname{sign}(\delta_k^u). \tag{5.5}$$

Therefore, the sign of the coefficient reveals the direction of impact of $\boldsymbol{z}_{u,i}[k]$ on $E(u_i)$. So if we do not compute the exact marginal effect, we may still say something about the direction of the impact by the sign of the coefficient. This is a convenient property, but as we will see later, the property does not always hold in models with a more complicated setup. Given that we will have n marginal effects for each variable, a concise statistic to present is the average partial effect (APE)

[22]Here marginal effects are based on the unconditional mean of u_i, although the JLMS formula uses the conditional mean, viz., $E(u_i|\varepsilon_i)$ as a point estimate of u_i. Kumbhakar & Sun (2013) derive the formulae for computing marginal effects using the JLMS formula.

on inefficiency or the partial effect of the average (PEA):

$$APE(\boldsymbol{z}_u[k]) =(\delta_k^u\sqrt{2/\pi})\left(n^{-1}\sum_{i=1}^{n}\sigma_{u,i}\right) \tag{5.6}$$

$$PEA(\boldsymbol{z}_u[k]) =\delta_k^u\sqrt{2/\pi}e^{\bar{\boldsymbol{z}}_u'\boldsymbol{\delta}^u}. \tag{5.7}$$

Either of these measures can be used to provide an overall sense for the impact of a given variable on the level of inefficiency. However, these statistics should also be interpreted with care. Neither necessarily reflects the impact of a given covariate for a given firm.

5.4. **How to Incorporate Exogenous Determinants of Efficiency.** In Section 5.2, we discussed how the quest for understanding the attributes of inefficiency evolved from a two-step estimation procedure to a theoretically preferred one-step estimation method. This section explores the issue further.

The one-step estimation method of investigating exogenous effects on inefficiency was first introduced in the truncated-normal model by Kumbhakar et al. (1991) and Reifschneider & Stevenson (1991). The same modeling strategy was later deployed by Huang & Liu (1994) and Battese & Coelli (1995), each using slightly different algebraic forms for the pre-truncated parameterization of the mean function of u_i. The above studies, which we label KGMHLBC, assume that the mean of the distribution of the pre-truncated u_i is a linear function of the exogenous variables under investigation. That is, they abandon the constant-mean assumption on μ, and assume, instead, that the mean is a linear function of some exogenous variables, viz.,

$$\mu_i = \boldsymbol{z}_{u,i}'\boldsymbol{\rho}^u, \tag{5.8}$$

where $\boldsymbol{z}_{u,i}$ is the vector of exogenous variables of observation i, and $\boldsymbol{\rho}^u$ is the corresponding coefficient vector (this is the same notation as the earlier setup, just with a different parameterization, for a different parameter). The log-likelihood function is the same as (2.10), except that (5.8) is used in place of μ. As before maximum likelihood estimation can be carried out to obtain estimates of $\boldsymbol{\rho}^u$ along with all other model parameters.

In addition to being a sensible approach to investigating the exogenous influences on efficiency, another appeal of the KGMHLBC model is that it makes the distributional shape of u_i even more flexible. The added flexibility stems from the allowance for an observation-specific mean of the pre-truncated distribution, with the mean determined by observation-specific variables. This is in contrast to the normal truncated-normal model of Stevenson (1980), where the mean of the pre-truncated distribution is identical for all observations. In a literature where the distributional assumption of u_i is essential and yet potentially open to criticism, anything that introduces greater flexibility is always regarded as beneficial.

Recall that in Section 5.2, we showed that the half-normal heteroscedastic model proposed by Caudill & Ford (1993), Caudill et al. (1995), and Hadri (1999) (CFCFGH hereafter), which parameterizes σ_u^2 by a function of $\boldsymbol{z}$, can also be used to address the issue of exogenous determinants of

inefficiency. The same conclusion applies here as well. A natural question to ask then is: which of the parameterization approaches, KGMHLBC or CFCFGH, is better in investigating exogenous influences on efficiency of a truncated-normal model? Wang (2002) argues that neither approach can be easily justified, and a more appropriate approach may come from combining the parameterizations of both models.

Consider that in the normal truncated-normal stochastic frontier model that the mean and variance of inefficiency are, respectively,

$$E(u_i) = \sigma_u \left[\frac{\mu}{\sigma_u} + \frac{\phi\left(\frac{-\mu}{\sigma_u}\right)}{\Phi\left(\frac{\mu}{\sigma_u}\right)} \right], \tag{5.9}$$

$$V(u_i) = \sigma_u^2 \left[1 - \frac{\mu}{\sigma_u} \left[\frac{\phi\left(\frac{-\mu}{\sigma_u}\right)}{\Phi\left(\frac{\mu}{\sigma_u}\right)} \right] - \left[\frac{\phi\left(\frac{-\mu}{\sigma_u}\right)}{\Phi\left(\frac{\mu}{\sigma_u}\right)} \right]^2 \right]. \tag{5.10}$$

The implication here is that both the mean and the variance of u_i are functions of μ *and* σ_u, and there is no justification of choosing to parameterize one over the other.

By considering the first two moments we see how the distinction between the modeling proposals of KGMHLBC and CFCFGH is blurred: regardless of whether μ or σ_u^2 is parameterized, all moments of u_i become observation-specific, and exogenous influences on mean efficiency are introduced. If the goal is to study how exogenous variables affect inefficiency, there is no particular reason why $\boldsymbol{z}_i$ should be assumed to exert influence through μ but not σ_u, or through σ_u but not μ. Without further information or assumptions, the decision of parameterizing only μ or σ_u would appear arbitrary.

Wang's (2002) model calls for parameterizing μ and σ_u^2 by the *same* vector of exogenous variables. The double parameterization is not only less ad hoc, but it also accommodates non-monotonic relationships between firm level inefficiency and its determinants. The latter can be of great importance to empirical research. The downside of this approach is that the model is more complex and, as a result, convergence problems may arise in the use of nonlinear optimization routines to maximize the log likelihood function. The double parameterization uses both (5.1) and (5.8). The log-likelihood function is the same as in (2.10), except that (5.1) and (5.8) are substituted in place of σ_u^2 and μ, respectively. Lai & Huang (2010) present a hierarchy of testing surrounding the parameterization of both μ and σ_v^2 in the truncated normal setup of Wang (2002). This can be useful to assist in reducing the complexity of the model, though issues of pre-test bias may arise as well.

5.4.1. *Marginal Effects.* The generality of the model of Wang (2002) is such that it contains both the KGMHLBC and CFCFGH models subject to specific parameter restrictions. This makes computation of marginal effects across all of the models simpler. Here we detail the marginal effects for Wang's (2002) model with the parameterizations (5.8) and (5.1) reproduced here for ease of

reference.

$$\mu_i = \boldsymbol{z}'_{u,i}\boldsymbol{\rho}^u,$$
$$\sigma^2_{u,i} = e^{\boldsymbol{z}'_{u,i}\boldsymbol{\delta}^u}.$$

The marginal effect of the kth element of $\boldsymbol{z}_{u,i}$ on $E(u_i)$ and $V(u_i)$ are as follows:[23]

$$\frac{\partial \mathrm{E}(u_i)}{\partial \boldsymbol{z}_u[k]} = \rho^u_k \left[1 - \Lambda_i\xi_i - \xi_i^2\right] + \delta^u_k\sigma_{u,i}\left[\left(1+\Lambda_i^2\right)\xi_i + \Lambda_i\xi_i^2\right]/2, \tag{5.11}$$

$$\frac{\partial V(u_i)}{\partial \boldsymbol{z}_u[k]} = \frac{\rho^u_k}{\sigma_{u,i}}\xi_i(E(u_i)^2 - V(u_i)) + \delta^u_k\sigma^2_{u,i}\left\{1 - \frac{1}{2}\xi_i\left(\Lambda_i + \Lambda_i^3 + (2+3\Lambda_i^2)\xi_i + 2\Lambda_i\xi_i^2\right)\right\}, \tag{5.12}$$

where $\Lambda_i = \mu_i/\sigma_{u,i}$, $\xi_i = \frac{\phi(\Lambda_i)}{\Phi(\Lambda_i)}$ and ρ^u_k and δ^u_k are the corresponding coefficients in (5.8) and (5.1), respectively, with $E(u_i)$ and $V(u_i)$ provided in (5.9) and (5.10). For the KGMHLBC model, σ^2_u is not parameterized, so the above formulas would set $\delta^u_k = 0$ and $\sigma_{u,i} = \sigma_u$. For the CFCFG model, the marginal effects are calculated setting $\rho^u_k = 0$ and $\mu_i = 0$. We mention here that considering the variance of inefficiency (and the associated impact of determinants on it) is useful if one thinks about production risk.

Wang (2002) demonstrates that the marginal effect of $\boldsymbol{z}_u[k]$ on either the mean or the variance of firm level inefficiency in the KGMHLBC model is monotonic, implying that an exogenous variable would *either* increase the mean and/or the variance of inefficiency, *or* decrease the mean and/or the variance of the inefficiency. The direction of the impact is monotonic across the sample, and is completely determined by the sign of ρ^u_k. Alternatively, the implied marginal effects on the mean and variance from Wang's (2002) model is non-monotonic, implying that, depending on the values of exogenous variables, the impact on inefficiency can change directions in the sample.

Allowing for a non-monotonic relationship between inefficiency and its determinants has important implications for practitioners. For instance, we may expect a younger farmer's (expected) technical efficiency to increase with age due to the accumulation of experience, while age may be a detrimental factor for an aged farmer due to deteriorated physical and mental capability (this is the example consider in Wang 2002). In this setting, age and inefficiency has a non-monotonic relationship. As this example illustrates, the accommodation of non-monotonicity can be important in empirical work to capture intuitive explanations regarding inefficiency. Wang's (2002) study found that ignoring the potential non-monotonic relationship between age and expected inefficiency (through application of KGMHLBC) resulted in smaller estimated marginal effects.

Although it may be reasonable to expect that adding squared terms of (or interactions between) variables in the KGMHLBC model may also account for inherent non-monotonicity in the mean and variance, the empirical example in Wang (2002) provided an indication that the addition of further nonlinearities into the parameterization of μ_i might not perform well in practical settings,

[23] See Kumbhakar & Sun (2013) for the equivalent formulas for $E(u_i|\varepsilon_i)$ and $V(u_i|\varepsilon_i)$.

in the sense that it fails to capture the non-linearity adequately given the extreme flexibility and overspecification. Non-monotonicity arises naturally in the model of Wang (2002) without resorting to a more flexible specifications of μ and σ_u^2.

5.5. **The Scaling Property.** The models we have discussed pertaining to incorporation of determinants of inefficiency take the approach of parameterizing one or all of the parameters as functions of $\boldsymbol{z}_i$ in the distribution of u or v. Wang & Schmidt (2002) (see also Simar, Lovell & van den Eeckaut 1994) proposed a different modeling strategy in which the random variable representing inefficiency has the following form:

$$u_i \sim g(\boldsymbol{z}_{u,i};\, \boldsymbol{\delta}^u)\, u_i^*, \tag{5.13}$$

where $g(\cdot) \geq 0$ is a function of the exogenous variables while $u_i^* \geq 0$ is a random variable. Distributional assumptions (such as half-normal or truncated-normal) can be imposed on u_i^*. Importantly, u_i^* does not depend on $\boldsymbol{z}_{u,i}$.

The model specified in (5.13) implies that the random variable u_i $(i = 1, 2, \ldots, n)$ follows a common distribution given by u^*, but each is weighted by a different, observation-specific *scale* of $g(\boldsymbol{z}_{u,i}, \boldsymbol{\delta}^u)$. Wang & Schmidt (2002) labeled $g(\cdot)$ the scaling function, and u^* the basic distribution. When u_i follows the formulation in (5.13) it is then said to exhibit the scaling property. The fully specified normal, truncated-normal stochastic frontier model with the scaling property is $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i - u_i$ where $u_i \sim g(\boldsymbol{z}_{u,i},\, \boldsymbol{\delta}^u) \cdot N^+(\tau,\, \sigma_u^2)$ and v is normally distributed with variance σ_v^2. Here τ and σ_u^2 are unconstrained constant parameters, and $\boldsymbol{z}_i$ is a variable vector which does *not* contain a constant. In this setup, the distribution of u_i is based on the basic distribution $N^+(\tau,\, \sigma_u^2)$ and the scale is stretched by the non-negative scaling function $g(\boldsymbol{z}_{u,i},\, \boldsymbol{\delta}^u)$, which can be specified as $e^{\boldsymbol{z}_i'\boldsymbol{\delta}^u}$ to ensure positivity and to mimic earlier parameterizations.

An attractive statistical feature of the model with a scaling property is that it captures the idea that the *shape* of the distribution of u_i is the same for all firms. The scaling function $g(\cdot)$ essentially stretches or shrinks the horizontal axis, so that the scale of the distribution of u_i changes but its underlying shape does not. In comparison, the KGMHLBC and Wang (2002) models allow different scalings for each u_i, so that for some u_i the distribution is close to a normal, while for some u_i the distribution is the extreme right tail of a normal with a mode of zero (the amount of truncation changes which impacts the shape). In comparison, for a model with the scaling property the mean and the standard deviation of u change with $\boldsymbol{z}_i$, but the shape of the distribution is fixed.

Another advantage of the scaling property specification is the ease of interpretation on the $\boldsymbol{\delta}$ coefficients when $g(\boldsymbol{z}_i,\, \boldsymbol{\delta}) = e^{\boldsymbol{z}_i'\boldsymbol{\delta}^u}$,

$$\frac{\partial \ln E(u_i)}{\partial \boldsymbol{z}_u[k]} = \delta_k^u. \tag{5.14}$$

That is, δ_k^u is the semi-elasticity of expected inefficiency with respect to $\boldsymbol{z}_u[k]$ and more importantly, this interpretation is distinct from any distributional assumption placed on u^*. This type of interpretation is usually unavailable in other model specifications.

TABLE 1. Models with Scaling Properties

	$g(\boldsymbol{z}_i;\, \boldsymbol{\delta})$	u^*
Aigner et al. (1977)	1	$N^+(0,\, \sigma_u^2)$
Meeusen & van den Broeck (1977)	1	$Exp(\sigma_u^2)$
CFCFGH	$e^{\boldsymbol{z}'_{u,i}\boldsymbol{\delta}^u}$	$N^+(0,\, 1)$
Stevenson (1980)	1	$N^+(\mu,\, \sigma_u^2)$

Perhaps most importantly, the scaling property provides an attractive economic interpretation, as detailed in Alvarez, Amsler, Orea & Schmidt (2006). u^* can be viewed as the base inefficiency level of the firm, capturing natural talent (random). Then, the scaling function allows a firm to exploit (or fail to exploit) these talents through other variables, $\boldsymbol{z}_u$, which might include experience of the plant manager, the operating environment of the firm, regulatory restrictions and the like. As Alvarez et al. (2006) make clear, the scaling property corresponds to a multiplicative decomposition of inefficiency into two independent parts, one random and one deterministic.

Some of the models introduced earlier can be seen as a special case of the scaling-property model; see Table 1. There are also models, such as KGMHLBC and Wang (2002), that do not have this scaling property. Wang's (2002) model does have the ability to reflect the scaling property. This requires that both the mean and the variance of the truncated normal are parameterized identically (both with exponential functions, say) and with the same parameters in each parameterization. We also mention that any distributional assumption involving a single parameter family (such as half-normal or Exponential) will automatically possess the scaling property.

5.5.1. *Marginal Effects.* The calculation of marginal effects for this model can also be obtained. From (5.9), we have

$$E(u_i) = e^{\boldsymbol{z}'_{u,i}\boldsymbol{\delta}^u} \cdot E(u^*), \tag{5.15}$$

where,

$$E(u^*) = \sigma_u \left[\frac{\tau}{\sigma_u} + \frac{\phi\left(\frac{\tau}{\sigma_u}\right)}{\Phi\left(\frac{\tau}{\sigma_u}\right)} \right]. \tag{5.16}$$

Thus, the marginal impact of the kth determinant of inefficiency is

$$\frac{\partial E(u_i)}{\partial \boldsymbol{z}_u[k]} = \delta_k^u e^{\boldsymbol{z}'_i\boldsymbol{\delta}^u} \cdot E(u^*). \tag{5.17}$$

$E(u^*)$ depends on the unknown parameters of the basic distribution and will need to be estimated. To obtain the estimated value, one can replace τ and σ_u^2 in the above equation by $\hat{\tau}$ and $\hat{\sigma}_u^2$, respectively, from the maximization of the log-likelihood in (2.10) with the scaling property imposed.

Similarly, for the marginal effect on the variance, we have

$$V(u_i) = e^{2\boldsymbol{z}'_{u,i}\boldsymbol{\delta}^u} \cdot V(u^*), \tag{5.18}$$

so that

$$\frac{\partial V(u_i)}{\partial \boldsymbol{z}_u[k]} = 2\delta_k^u \cdot e^{2\boldsymbol{z}'_{u,i}\boldsymbol{\delta}^u} \cdot V(u^*), \tag{5.19}$$

where

$$V(u^*) = \sigma_u^2 \left[1 - \frac{\tau}{\sigma_u} \left[\frac{\phi\left(\frac{\tau}{\sigma_u}\right)}{\Phi\left(\frac{\tau}{\sigma_u}\right)} \right] - \left[\frac{\phi\left(\frac{\tau}{\sigma_u}\right)}{\Phi\left(\frac{\tau}{\sigma_u}\right)} \right]^2 \right], \tag{5.20}$$

and will need to be constructed from the maximum likelihood estimates.

5.5.2. *A Test of the Scaling Property.* The scaling property is not a fundamental feature, it is, as with the choice of distribution on u, an assumption on the features of inefficiency. As such it can be tested against models that do not possess this property for the inefficiency distribution.

A test for the presence of the scaling property can be constructed within the framework of Wang's (2002) model. To begin, assume that $u_i \sim N^+(\mu_i, \sigma_{u,i}^2)$ where $\mu_i = \mu e^{\boldsymbol{z}'_{u,i}\boldsymbol{\rho}^u}$ and $\sigma_{u,i} = \sigma_u e^{\boldsymbol{z}'_{u,i}\boldsymbol{\delta}^u}$. This is almost identical to our earlier discussion of the Wang (2002) model except that we are modeling the mean of inefficiency in an exponential fashion as opposed to linearly. A test of the scaling property then corresponds to testing the null hypothesis $H_0 : \boldsymbol{\rho}^u = \boldsymbol{\delta}^u$. This can be done using any of the trinity of classical test statistics: likelihood ratio, Lagrange multiplier or Wald. Deploying any of these three test statistics will produce (asymptotically) a χ_q^2 random variable which can be used to generate p-values. Alvarez et al. (2006) provide further details on the construction of all three of these test statistics.

Several other variants of the scaling property exist within this framework, but face more demanding analysis given that without further restrictions or assumptions the tests become nonstandard. For example, if we assume *ex ante* that $\boldsymbol{\rho}^u = 0$, then an alternative test of the scaling property is $H_0 : \mu = 0$; i.e., we are testing the truncated normal specification against a half normal specification).

As it currently stands all tests of the scaling property hinge on a given distributional assumption. An important avenue for future research is the development of a test (or tests) that do not require specific distributional assumptions. This should be possible in the context of the distribution free approach that stems from the scaling property (which we turn to next), but we leave this for future development.

5.6. **Estimation Without Distributional Assumptions.** When the distribution of inefficiency possesses the scaling property, it is possible to estimate the stochastic frontier model without imposing distributional assumptions, provided a single determinant of inefficiency exists. This is perhaps the most fundamental benefit of the scaling property, the stochastic frontier *and* the deterministic component of inefficiency can be recovered without requiring a specific distributional assumption. The reason for this is that under the scaling property, the conditional mean of u on $\boldsymbol{z}_u$ only depends on the basic distribution up to scale; any basic distribution will produce the same (scaled) conditional mean. More importantly, this scale can be estimated. The elegance of this

result is that the model can be estimated using nonlinear least squares. The model we discuss here was first proposed by Simar et al. (1994) (but without use of the 'scaling property' terminology) and later expounded on in Wang & Schmidt (2002) and Alvarez et al. (2006).

Typically, if one were to maintain the common parameterization that has appeared in the literature discussing the scaling property, then the regression model would be

$$y = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i - e^{\boldsymbol{z}_{u,i}'\boldsymbol{\delta}^u}u_i^* \tag{5.21}$$

and the conditional mean of y given $\boldsymbol{x}$ and $\boldsymbol{z}_u$ is

$$E[y|\boldsymbol{x},\boldsymbol{z}_u] = \boldsymbol{x}'\boldsymbol{\beta} - e^{\boldsymbol{z}_u'\boldsymbol{\delta}^u}\mu^* \tag{5.22}$$

where $\mu^* = E(u^*)$. Our regression model is

$$y = \boldsymbol{x}_i'\boldsymbol{\beta} - e^{\boldsymbol{z}_{u,i}'\boldsymbol{\delta}^u}\mu^* + v_i - e^{\boldsymbol{z}_{u,i}'\boldsymbol{\delta}^u}(u-\mu^*) = \boldsymbol{x}_i'\boldsymbol{\beta} - e^{\boldsymbol{z}_{u,i}'\boldsymbol{\delta}^u}\mu^* + \varepsilon_i^*, \tag{5.23}$$

which can be estimated using nonlinear least squares as

$$\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\delta}}^u,\widehat{\mu}^*\right) = \min_{\boldsymbol{\beta},\boldsymbol{\delta}^u,\mu^*} n^{-1}\sum_{i=1}^{n}\left[y_i - \boldsymbol{x}_i'\boldsymbol{\beta} + \mu^* e^{\boldsymbol{z}_{u,i}'\boldsymbol{\delta}^u}\right]^2. \tag{5.24}$$

The need for nonlinear least squares stems from the fact that the scaling function must be positive and so if it was specified as linear this would be inconsistent with theoretical requirements on the variance of the distribution. The presence of μ^* implies that one cannot include a constant in $\boldsymbol{z}_u$, as this leads to identification issues. Given that the error term ε_i^* is heteroscedastic,

$$Var(\varepsilon^*|\boldsymbol{x},\boldsymbol{z}_{u,i}) = \sigma_v^2 + \sigma_u^{2*}e^{2\boldsymbol{z}_u'\boldsymbol{\delta}^u},$$

where $\sigma_v^2 = Var(v_i)$ and $\sigma_u^{2*} = Var(u^*)$, either a generalized nonlinear least squares algorithm would be required (though this requires distributional assumptions to disentangle σ_v^2 and σ_u^{2*}), or, more readily, heteroscedastic robust standards (White 1980) can be constructed to ensure that valid inference is conducted.

Given that nonlinear least squares algorithms are readily available across a wide range of statistical software, it is surprising to us that this approach is not as common as full blown maximum likelihood estimation, which, while more efficient, requires distributional assumptions. The nonlinear least squares approach is not computationally harder to implement and allows users to eschew distributional assumptions, at the expense of imposing the scaling property. Further, calculation of expected firm efficiency can be done without requiring distributional assumptions, leading to the potential for more robust conclusions regarding observation specific inefficiency. Note that here the implicit assumption is that the z_u variables are different from the x variables, although because of nonlinearity if some x variables can be used as z_u and their effects can be identified. In Section 8 we will again return to this model but show how to relax functional form assumptions on the scaling function.

5.7. **Testing for Determinants of Inefficiency.** Perhaps a more interesting test is not the presence of the scaling property, but if inefficiency truly depends on a set of exogenous determinants. Kim & Schmidt (2008) propose a suite of tests for exactly this purpose.

They propose both one-step and two-step tests. The two-step tests are similar in spirit to the original two-step procedure used to determine *how* inefficiency depended upon $\boldsymbol{z}_u$. Recall our main conclusion, the two-step procedure should not be used for this purpose. However, when the null hypothesis is that $\boldsymbol{z}_u$ does not impact u, then the two-step procedure is a valid approach because both the misspecification bias in the first step and the under dispersion bias introduced to the second step no longer appear.

Unfortunately, Kim & Schmidt (2008) show that the F-test of the joint significance of the coefficients on the z-variables is not asymptotically valid. The test requires a correction for the variance matrix. This correction however, is distribution specific and so the test is anchored to the distributional assumption used in the first stage to produce the conditional mean inefficiency estimates. An interesting technical finding of Kim & Schmidt (2008) is that the Lagrange multiplier test, under the distributional assumption of u being Exponential (which implies the scaling property), is asymptotically equivalent to the two-step procedure. In general this is not the case.

As an alternative to inference within the confines of the two-step procedure, a different route is to deploy the assumption of the scaling property and use nonlinear least squares. Kim & Schmidt (2008) also discuss the implications of a test based on this framework. It turns out that in this setting one has an identification problem given that under $H_0 : \boldsymbol{\delta}^u = 0$ it follows that

$$y_i = \alpha + \boldsymbol{x}_i'\boldsymbol{\beta} - \mu^* e^{\boldsymbol{z}_{u,i}'\boldsymbol{\delta}^u} + \varepsilon_i^* = (\alpha - \mu^*) + \boldsymbol{x}_i'\boldsymbol{\beta} - \mu^*(1 - e^{\boldsymbol{z}_{u,i}'\boldsymbol{\delta}^u}) + \varepsilon_i^*.$$

One can only identify μ if at least one element of $\boldsymbol{\delta}^u$ is nonzero. This poses an issue for inference given that identification failures under the null hypothesis invalidate the asymptotic properties of Wald and likelihood ratio tests (since both tests are not scale invariant to the presence of μ^*). Kim & Schmidt's (2008) solution to this problem is to use the score test principle (following the discussion in Wooldridge 2010) which involves estimates under the imposition of the null hypothesis. The score (Lagrange multiplier) test is based on the derivative of the nonlinear least squares criterion function in (5.24) with respect to $\boldsymbol{\delta}^u$, evaluated at the restricted estimates:

$$\frac{2}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta} + \mu^* e^{\boldsymbol{z}_{u,i}'\boldsymbol{\delta}^u}\right)\left(\mu^* e^{\boldsymbol{z}_{u,i}'\boldsymbol{\delta}^u}\boldsymbol{z}_{u,i}\right) = \frac{2}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta} + \mu^*\right)\left(\mu^*\boldsymbol{z}_{u,i}\right). \tag{5.25}$$

The score test determines how close the derivative of the nonlinear least squares function (with respect to the parameters under the null hypothesis) is to 0. If the parameter restrictions are true then this should be close to 0. Note that the derivative in (5.25) (the LM statistic) is equivalent to the F-test for the significance of the coefficients, $\boldsymbol{\zeta}$ of $\mu^*\boldsymbol{z}_u$ in the regression

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + (\mu^*\boldsymbol{z}_{u,i})'\boldsymbol{\zeta} + \epsilon_i. \tag{5.26}$$

The key feature that makes this test work without distributional assumptions, as pointed out in Kim & Schmidt (2008), is that F-tests are invariant to the scale of the covariates. Thus, even though μ^* is unobserved and requires a distributional assumption to estimate it, in this specific regression its only purpose is to scale $\boldsymbol{z}_u$. This is not the case if we were to test $H:\boldsymbol{\delta}^u = 0$ using a Wald or likelihood ratio statistic. Thus, for our purposes, we can simply set $\mu^* = 1$ and regress y on $(\boldsymbol{x}, \boldsymbol{z}_u)$ and test the significance of $\boldsymbol{\zeta}$.

We conclude our discussion by mentioning that the scaling property does not have to be explicitly imposed to use the nonlinear least squares framework to test for the presence of determinants of inefficiency. The F-test for the regression in (5.26) is still a valid test for the impact of $\boldsymbol{z}_u$, it is just no longer a Lagrange multiplier test. The reason for this is that $H:\boldsymbol{\zeta} = 0$ is simply a test that $E[y|\boldsymbol{x}, \boldsymbol{z}_u]$ does not depend on $\boldsymbol{z}_u$. Thus, this test could also be used to determine if $\boldsymbol{z}_u$ influenced inefficiency in the KGMHLBC model that does not possess the scaling property. However, if one were to use this test, it is important to recognize that more powerful tests could be constructed (Kim & Schmidt 2008).

6. Accounting for Heterogeneity in the Stochastic Frontier Model

The accurate measurement of performance is important to understanding the true technology. If multiple technologies are available, a more accurate characterization of productivity potential involves accurate identification of them. The proposition that a range of production technologies exist in an industry is generally accepted, in principle, but rarely in the execution of empirical analyses. Estimation of the technology rests on the assumption that all producers in the sample use the same technology. However, farms in an industry may use several different technologies. In such a case estimating a common production technology using a production function encompassing the entire sample of observations may not be appropriate in the sense that the estimated technology can fail to represent a "true" technology, but rather multiple technologies are simultaneously used, is likely to be biased. In this context, the "single" (or aggregate) technology estimated presents misleading characterizations of scale economies, elasticities of substitutions and other measures of production structure.

To this point we have assumed a homogeneous production technology across all firms. However, firms in a particular industry may use different technologies. In such a case estimating a common frontier function encompassing every sample observation may not be appropriate in the sense that the estimated technology is not likely to represent the true technology. That is, the estimate of the underlying technology may be biased. Furthermore, if the unobserved technological differences are not taken into account during estimation, the effects of these omitted unobserved technological differences might be inappropriately labeled as inefficiency.

When heterogeneous technologies are present estimates of production technology and inefficiency can be severely misleading. Further, if there are differences in the overall mapping of inefficiency then an approach which treats all firms identically will produce poor results. A simple way to handle this is to incorporate *a priori* information to split the sample so that homogeneity of technology and efficiency is retained. And the applied efficiency field has long such recognized the importance of production technology heterogeneity. Early studies that allowed for different technologies across firms include Mester (1993), who split US savings and loans banks based on if they are mutual or stock owned (essentially a private versus public split), Altunbas, Evans & Molyneux (2001), splitting on organizational structure of banks, and Mester (1996) and Bos & Schmiedel (2007) splitting based on geography. Beyond the banking context, Bravo-Ureta (1986), Tauer (1998), Newman & Matthews (2006) specify heterogeneous technologies for dairy farms based on breed, location and production processes, respectively.

This is easily tackled when observable classification information is present. The normal, half-normal stochastic frontier likelihood function[24] to be optimized in this setting is

$$\ln \mathcal{L}^c = -n \ln \sigma_c + \sum_{i=1}^{n} \ln \Phi(-\varepsilon_{ci}\lambda_c/\sigma_c) - \frac{1}{2\sigma_c^2} \sum_{i=1}^{n} \varepsilon_{ci}^2, \tag{6.1}$$

[24]It should be clear that this framework can be extended to accommodate any of the distributional assumptions discussed in Section 2.

where $\varepsilon_{ci} = y_{ci} - \boldsymbol{x}'_{ci}\boldsymbol{\beta}^c$ and $c = 1, \ldots, C$, for the C different classifications that exist. In Mester's (1993) setup $C = 2$, but there is nothing that precludes having a binary classification of firms. For example, one could further subdivide the mutual and stock savings and loans by broad geographic regions. The likelihood function in (6.1) is just the normal, half-normal stochastic frontier likelihood function in (2.6), but with the observations partitioned according to which class they belong and where each class has its own unique parameters. In this setup it is straightforward to test for differences in the parameters across the different classes using a likelihood ratio test.

The main feature of these types of studies is that the sample observations are classified into several groups based on some a priori sample separation information such as ownership of firms, location of firms, etc. Once such classifications are made, separate analyses are carried out for each class in the second stage. This procedure is not efficient in the sense that (i) the classification might be inappropriate, and (ii) information contained in one class is not used to estimate the technology (production or cost frontier) of firms that belong to other classes. In most of the empirical applications this inter-class information may be quite important because firms belonging to different classes often come from the same industry. Although their technologies may be different, they share some common features.

However, it may not be the case that for a given application that a clear classification exists (or is observable). Further, the main disadvantage of this approach is that the *a priori* sample separation is, to some extent, arbitrary. To see this, note that Koetter & Poghosyan (2009) find in their study of banks, that even with similar organizational structure, banks operate under different technological structures. Thus, more powerful methods are needed if it is believed that latent heterogeneity exists in the underlying production or structure. Additionally, while allowing for production heterogeneity certainly adds a layer of flexibility to the stochastic frontier model, there could also be differences in the underlying stochastic components of the model, in both v and u. A model that can aptly characterize this specific setting is the latent class stochastic frontier model, which we now turn our attention to.

6.1. **Latent Class Models for Stochastic Frontiers.** The latent class model (LCM) permits us to exploit the information contained in the data more efficiently, with the advantage that the LCM can easily incorporate technological heterogeneity by estimating a mixture of production functions.[25] In the standard finite mixture model the proportion of firms (observations) in a group is assumed to be fixed (a parameter to be estimated), see, e.g., Caudill (2003). The proposed research will assume that firms in the sample use multiple technologies and the probability of a firm using a particular technology can be explained by some covariates and may vary over time. The LCM can also be extended to accommodate the simultaneous existence of multiple technologies in which firms are not necessarily fully efficient. The stochastic frontier approach is a technique that is now widely used to estimate technical efficiency of firms. Recently a few studies have combined the

[25]As an alternative to a latent class approach, Tsionas & Kumbhakar (2004) propose a stochastic frontier production function augmented with a Markov switching structure to account for different technology parameters across heterogenous countries.

stochastic frontier approach with the latent class structure in order to estimate a mixture of frontier functions. In particular, Caudill (2003) introduces an expectation-maximization (EM) algorithm to estimate a mixture of two stochastic cost frontiers in the presence of no sample separation information. Orea & Kumbhakar (2004) use the standard procedure (brute force maximization of the likelihood function) while Greene (2005*b*) deploys maximum simulated likelihood to obtain the maximum likelihood estimator in which the (prior) probability of adopting a particular technology is assumed to depend on firm characteristics. Class membership probabilities are likely to be different between the standard LCM (in which all firms are assumed to be fully efficient) and the LCM that allows firms to be technically inefficient.

Efforts to accommodate heterogeneity lead to the problem of determining the criteria to use to group firms and whether the technologies across groups differ completely or only in the intercepts. These problems can be avoided by estimating these technologies simultaneously along with the (prior) probability of using them (that might vary with firm). Based on the estimated (posterior) probability we can classify the sample firms into several groups. It is also possible to find the optimal number of groups/classes that the data support.

To properly construct the latent class stochastic frontier model (LCSFM), we assume that there is a latent, unobservable sorting of the data into C classes.[26] For an observation from class c, the conditional density can be characterized as

$$f(y_i|\boldsymbol{x}_i, c) = f(\Theta_c, y_i, \boldsymbol{x}_i), \tag{6.2}$$

where Θ_c is the collection of parameters, specific to class c for the stochastic frontier model. This is akin to the subjective sample separation described previously, but here no information is available on what these classes are and which firms belong to what classes. In the sample selection model we have information on these classes and we also know which firms belong to what classes. Presence of technological heterogeneity in the absence of such information is where one may want to use LCSFMs. Although information on the factors explaining class participation (adoption of a particular technology) can be used to estimate the probability of being in a class such classification is always probabilistic (never be 100% sure about who uses what technology). However, all the data is used to estimate the technologies as well as the probability of using a particular technology.

One assumption that we will maintain in our description of the LCSFM is that $f(\cdot)$ is the same across the classes. That is, we restrict ourselves to having all classes have normal, half-normal structure instead of, for example if $C = 2$, having one class of the normal, half-normal framework and the other class having the normal, Gamma framework.

For the normal, half-normal stochastic frontier we have $\Theta_c = (\boldsymbol{\beta}_c, \sigma_c, \lambda_c)$ and our class specific density is

$$P(i|c) = \frac{2}{\sigma_c}\phi(\varepsilon_{ci}/\sigma_c)\Phi\left(\frac{\varepsilon_{ci}\lambda_c}{\sigma_c}\right) \tag{6.3}$$

[26]While the LCSFMs of Caudill (2003), Orea & Kumbhakar (2004), and Greene (2005*b*) all are constructed around a panel, our discussion here will focus on the cross-sectional case.

and $\varepsilon_{ci} = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_c$. $P(i|c)$ represents observation i's contribution to the conditional on c likelihood. The unconditional likelihood for observation i is simply the average over all C classes:

$$P(i) = \sum_{c=1}^{C} \Pi(i,c) P(i|c), \tag{6.4}$$

where $\Pi(i,c)$ is the prior probability of observation i's membership in class c. In this framework $\Pi(i,c)$ reflects the uncertainty residing around class membership for a given observation. Earlier models used prior information such that $\Pi(i,c)$ was either 0 or 1. $\Pi(i,c)$ can be parameterized based on a set of covariates (similar to the inclusion of determinants of inefficiency), but it must be borne in mind that $0 \leq \Pi(i,c) \leq 1$ and so a logit or probit parameterization is appropriate. For $C > 2$ the multinomial logit form would parameterize $\Pi(i,c)$ as

$$\Pi(i,c) = \frac{e^{\boldsymbol{z}_{l,i}'\boldsymbol{\pi}_c}}{1 + \sum_{m=1}^{C-1} e^{\boldsymbol{z}_{l,i}'\boldsymbol{\pi}_m}}, \tag{6.5}$$

where $\boldsymbol{z}_{l,i}$ captures the fact that here $\boldsymbol{z}$ is a variable reflecting inclusion into a specific class as opposed to influencing inefficiency directly ($\boldsymbol{z}_u$ as in Section 5). Note also that $\Pi(i,c)$ does not have to be dependent upon a set of covariates, this is just an option for the analyst. In this case $\Pi(i,c) = \Pi(c)$ is a constant across observations.

The log-likelihood function for the LCSFM is then

$$\ln \mathcal{L} = \sum_{i=1}^{n} \ln P(i), \tag{6.6}$$

which, as noted by Greene (2005*b*) can be optimized in a brute force fashion, through the EM algorithm or maximum simulated likelihood. The EM has intuitive appeal. Define the posterior probability of class c given observation i as

$$w(c|i) = \frac{P(i|c)\Pi(i,c)}{\sum_{m=1}^{C} P(i|m)\Pi(i,m)}. \tag{6.7}$$

Then, maximum likelihood estimates can be found by iterating between

$$\widehat{\boldsymbol{\Theta}}_c = \max \sum_{i=1}^{n} w(c|i) \ln P(i|c), \quad c = 1, \ldots, C \tag{6.8}$$

and

$$(\widehat{\boldsymbol{\pi}}_1, \widehat{\boldsymbol{\pi}}_2, \ldots, \widehat{\boldsymbol{\pi}}_C) = \max \sum_{i=1}^{n} \sum_{c=1}^{C} w(c|i) \ln \Pi(i,c) \tag{6.9}$$

making sure to update $w(c|i)$ after each iteration. The EM algorithm essentially acts as though there is some initial guess at the class probabilities, then the class parameters are estimated through maximization of the log-likelihood function of the cth class. Finally, with these class parameters

fixed, new multinomial coefficients are estimated and the whole process is repeated. Finally, if the class membership probabilities are fixed across observations then the estimates of π_c represent posterior probabilities. Once a solution is reached, $w(c|i)$ provide estimates of the class probabilities for a given observation. Class assignment is typically based on the largest posterior probability.

To construct final estimates one can either use the class specific estimates based on the estimated largest posterior probability. For many observations the largest posterior probability may be close to one and this will be an acceptable strategy. However, it is also likely that there will exist observations for which there is a reasonable probability of belonging in more than one class and so choosing a reference class will be misleading. In this case one can compute a posterior-probability-weighted sum of the parameters, as

$$\widehat{\boldsymbol{\Theta}} = E(\widehat{\boldsymbol{\Theta}}|c) = \sum_{c=1}^{C} \widehat{w}(c|i)\widehat{\boldsymbol{\Theta}}_c. \tag{6.10}$$

Estimates of inefficiency can also be calculated in this fashion. Thus, one would calculate the conditional mean inefficiency scores for each class and then aggregate using the posterior probabilities.

Lastly, as Greene (2005*b*) mentions, one cannot treat C as a parameter to be estimated. C is assumed known at the start of the analysis. However, the setup with $C-1$ classes is nested in the C class structure and so one can select a model with C classes based on traditional model selection criteria such as the Akaike or Schwarz information measures (this is advocated in Orea & Kumbhakar 2004). Note that a formal test of C classes against $C-1$ is tricky given that if there exists C classes and one only estimates $C-1$ classes then the parameter estimates from this model are inconsistent. Further, if there are $C-1$ classes then the test is essentially $\boldsymbol{\Theta}_C = \boldsymbol{\Theta}_{C-1}$ and the model cannot identify $\boldsymbol{\pi}_C$ for any pair of classes. This is problematic because it is not clear what the appropriate degrees of freedom for the likelihood ratio test, say, is. More research on formal tests of the number of mixtures is an interesting avenue for future research.

6.2. **The Zero Inefficiency Stochastic Frontier.** Several interesting aspects from the latent class model arise in practice. Notably, in their study of cross-country productivity and growth Bos, Economidou & Koetter (2010, Group A, Table 3) and Bos, Economidou, Koetter & Kolari (2010, Group B, Table 1) report estimates of λ that are quite small and produce group specific inefficiencies that are almost 1. We say almost because the stochastic frontier model, based on the distributional assumptions place on u do not allow for the presence of fully efficient firms. As Wheat et al. (2014) note, the probability that u_i is 0 for any firm is 0. But the latent class estimates of Bos, Economidou & Koetter (2010) and Bos, Economidou, Koetter & Kolari (2010) suggest a small group of countries that are, for all intents and purposes, fully efficient.

Kumbhakar, Parmeter & Tsionas (2013) develop a model that allows some firms to be fully efficient ($u_i = 0$ for some i) while others can be inefficient $u_i > 0$ and this information is not available to the econometrician. Thus, the problem is to assess which regime (efficient or inefficient) each firm belongs to, which is identical to the latent class models just introduced. The key difference is that in the latent class models developed earlier, all producers, regardless of class were inefficient.

As noted in Kumbhakar et al. (2013), many papers which deploy the LCM framework typically recover a class where the variance parameter of inefficiency in one class is effectively 0, suggesting no inefficiency. Thus, explicitly modeling this zero inefficiency class can produced interesting insights which the traditional LCM overlooks. Given the prevalence of firms with 0 inefficiency, Kumbhakar et al. (2013) refer to their stochastic frontier model as a zero inefficiency stochastic frontier (ZISF) model. The ZISF model is formulated as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i \text{ with probability } p \text{ and } y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + (v_i - u_i) \text{ with probability } (1-p),$$

where p is the probability of a firm being fully efficient. Alternatively, p is the proportion of firms that are fully efficient and $1-p$ is the proportion of firms that are technically inefficient.

Note that the technology in both regimes is the same. Since the regimes are unobserved, the ZISF falls into the category of a latent class model. Previous studies on latent class SF models either have different probabilities on the technology, $\boldsymbol{x}_i'\boldsymbol{\beta}$, or the composed error, $\varepsilon_i = v_i - u_i$ (or both). By acknowledging the fact that a subset of firms belong to a fully efficient regime, the ZISF model boils down to an analysis of regime membership based solely on inefficiency. The composed error term in the ZISF is $v_i - u_i(1 - 1\{u_i = 0\})$ where $1\{u_i = 0\} = p$. Thus, the ZISF error term is not the same as the probability weighted composed error term $v_i - u_i$ which would be the case if one were to use a latent class SF model with identical technology.[27]

The density of our convoluted error term, assuming the conventional normal and half normal distributions for v and u is

$$(p/\sigma_v)\phi(\varepsilon/\sigma_v) + (1-p)\left[\frac{2}{\sigma}\phi\left(\frac{\varepsilon}{\sigma}\right)\Phi\left(\frac{-\varepsilon\lambda}{\sigma}\right)\right]. \tag{6.11}$$

The ZISF convoluted error distribution follows directly from the intuition afforded from our earlier discussion on regime membership.[28] This density function is a mixture between a normally distributed random variable and the common, convoluted density from a normal/half-normal SF model. Thus, any of the array of standard SF densities can be used in deriving the likelihood function for the ZISF model. A natural extension of the model discussed here that does not require additional computational effort is to make the probability of full efficiency a parametric function of a set of covariates:

$$p_i = \frac{\exp(\boldsymbol{z}_{pi}'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{z}_{pi}'\boldsymbol{\gamma})}, \text{ or } p_i = \Phi\left(\boldsymbol{z}_{pi}'\boldsymbol{\gamma}\right), \qquad \text{for } i = 1, \ldots, n,$$

where $\boldsymbol{z}_{pi}$ is an $m \times 1$ vector of exogenous variables which influence whether a firm is inefficient or not and $\boldsymbol{\gamma}$ is an $m \times 1$ vector of parameters. Notice that in this framework $\boldsymbol{z}_p$ does not directly influence the level of inefficiency of a firm, rather, it influences the likelihood that a firm is fully efficient. In this case we can think of $\boldsymbol{z}_p$ as indirectly influencing the level of inefficiency. Contrast this to the influence of $\boldsymbol{z}_p$ that we discussed in Section 5

[27] The case of a finite mixture of heterogeneous composed error distributions with homogeneous technology can be derived as a special case of Greene (2005*b*).

[28] See Kumbhakar & Lovell (2000, ch. 3) for more on these derivations.

In any latent class model a key concern is identifying the number of classes. Here we do not face as difficult a problem because we are dealing *a priori* with two classes (fully efficient and inefficient firms). The only additional parameter in our ZISF model in comparison with the standard SF model is p. In a two class mixture model this is the mixing proportion parameter in the population. Thus, statistical identification of this parameter requires non-empty cells (non-zero observations in each group) in the sample (see McLachlan & Peel 2000). The variance of the inefficiency distribution, σ_u^2, in the SF model is identified through moment restrictions on the composed density of the inefficiency and random components (skewness of the composed error distribution). When $\lambda \to 0$ identification of the model breaks down as the inefficient regime and the fully efficient regime become indistinguishable. However, this case does not pose a problem intuitively since $\lambda \to 0$ implies that firms are becoming close to fully efficient.

Grassetti (2011) proposed a panel extension of the ZISF in the spirit of Pitt & Lee (1981) and used this model to study efficiency of Italian hospitals. Rho & Schmidt (2013) discussed identification issues within the ZISF model proposed by Kumbhakar et al. (2013). Specifically, they documented the presence of the 'wrong skewness' problem of Waldman (1982) as well as identification of σ_u^2 when $p \approx 1$ and identification of p when $\sigma_u^2 \approx 0$, i.e. when all firms are fully efficient we cannot tell if it is because $p = 1$ or because $\sigma_u^2 = 0$. This has important implications for conducting inference.

6.3. **Sample Selection in the Stochastic Frontier.** The models we have assumed so far have allowed heterogeneity in the stochastic frontier model in an exogenous fashion. There are different technologies but those technologies are exogenous to the setup or there are efficient and inefficient firms operating in the same, given environment. In this section we detail a recent set of models that allow firms to chose the (pre-specified) technology that they wish to use. The incorporation of a choice amounts to a sample selection problem in the spirit of Heckman (1976). One interesting aspect of this selection is *how* selection enters the model.

Several early approaches to deal with potential sample selection within a stochastic frontier model proceeds by following the two-step Heckman (1976) correction.[29] In the first stage the probability of selection is estimated and then the inverse Mill's ratio is calculated for each observation based on this estimates. The estimated inverse Mill's ratio is then included as a regressor in the final regression. For example, see the hospital study of Bradford, Kleit, Krousel-Wood & Re (2001) and the Finnish farming study of Sipiläinen & Oude Lansink (2005), both of which follow this two-step approach.

As Greene (2010) makes clear, this limited information, two-step approach in the standard linear regression setting works precisely because of the linearity of the main regression model. However, Kumbhakar, Tsionas & Sipiläinen (2009) note that when inefficiency is present no two-step approach will work and so full information maximum likelihood estimation is required.[30]

[29]Earlier studies acknowledged the potential presence of selection within a stochastic frontier setting (Kaparakis, Miller & Noulas 1994), but did not account for it.

[30]See also Lai, Polachek & Wang (2009).

6.3.1. *Selection Based on Noise.* The sample selection stochastic frontier model of Greene (2010) consists of the selection equation

$$d_i = 1\left\{\boldsymbol{z}'_{s,i}\boldsymbol{\alpha} + w_i > 0\right\} \tag{6.12}$$

where w_i is distributed as standard normal and the $\boldsymbol{z}_s$ represent the variables which influence selection into (or out of) the sample, and the standard stochastic production frontier

$$y_i = \boldsymbol{x}'_i\boldsymbol{\beta} + v_i - u_i.$$

Selection is an issue because it is assumed that $(w_i, v_i) \sim N_2[\mathbf{0}, \boldsymbol{\Sigma}]$ where the off-diagonal elements of $\boldsymbol{\Sigma}$ are equal to $\rho\sigma_v$. Thus, there exist unobservables that are correlated with both w_i and v_i, which, if not properly accounted for will bias the coefficient estimates of the stochastic production frontier.

The conditional density of y given $\boldsymbol{x}$, $\boldsymbol{z}_s$, d_i and u_i is

$$f(y|\boldsymbol{x}, \boldsymbol{z}_s, d_i, u_i) = d_i\phi(v_i/\sigma_v)\Phi\left(\frac{\rho v_i/\sigma + \boldsymbol{z}'_{s,i}\boldsymbol{\alpha}}{\sqrt{1-\rho^2}}\right) + (1-d_i)\Phi(-\boldsymbol{z}'_{s,i}\boldsymbol{\alpha}), \tag{6.13}$$

where $v_i = y_i - \boldsymbol{x}'_i\boldsymbol{\beta} + u_i$. Aside from the presence of u_i in v_i this is identical to the full information conditional density in the sample selection model. The conditional density in (6.13) is not operational given that u_i is unobserved. As our discussion of the maximum simulated likelihood estimator made clear, we can either integrate out u_i from the conditional density, which depending upon the distributional assumption, may not yield an analytically tractable solution, or use simulation techniques. We follow Greene (2010) and detail MSL estimation of the sample selection model. The simulated log likelihood function is

$$\ln \mathcal{L}_s = \sum_{i=1}^{n} \ln\left(R^{-1}\sum_{r=1}^{R} d_i\phi(\tilde{v}_i/\sigma_v)\Phi\left(\frac{\rho\tilde{v}_i/\sigma + \boldsymbol{z}'_{s,i}\boldsymbol{\alpha}}{\sqrt{1-\rho^2}}\right) + (1-d_i)\Phi(-\boldsymbol{z}'_{s,i}\boldsymbol{\alpha})\right), \tag{6.14}$$

where $\tilde{v}_i = y_i - \boldsymbol{x}'_i\boldsymbol{\beta} + \sigma_u|U_{ir}|$ with U_{ir} being drawn from a standard normal distribution. This simulated log likelihood function can be optimized directly to obtain estimates of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, σ_u, σ_v and ρ.

Greene (2010) recommends a simplified two-step estimator noting that $\boldsymbol{\alpha}$ can be estimated in a consistent fashion separately from the rest of the parameters. See his discussion for more details on this approach. We mention here that sample selection in the stochastic frontier model of Greene (2010) arises due to correlation between v and w. This is pertinent in the context of the empirical application of Greene (2010) who looks at efficiency in health care and there is a stark difference in efficiency scores based on OECD membership. Thus, it could be that OECD membership may be acting as a selection mechanism.

6.3.2. *Sample Selection Based on Inefficiency.* A key feature of the sample selection model of Greene (2010) is that the choice of technology is influenced by correlation between random error in the selection and frontier models; however, if this choice of technology is based on some aspect

of inefficiency then a different form of sample selection arises. Kumbhakar et al. (2009) construct a model that explicitly takes account of this phenomena. The stochastic production frontier of Kumbhakar et al. (2009) is identical to that in Greene (2010). The difference is in the selection equation, which they specify as

$$d_i = 1\left\{\boldsymbol{z}'_{s,i}\boldsymbol{\alpha} + \delta u_i + w_i > 0\right\}, \tag{6.15}$$

where it is assumed that $w_i \sim N(0,1)$. Note that, conditionally on $\boldsymbol{z}$, if u_i were known then the distribution of d_i is normal. Thus, we can use a similar type of maximum simulated likelihood argument to estimate the parameters of the selection function and the production frontier.

First, Kumbhakar et al. (2009) show that the conditional joint distribution of y_i and d_i is

$$f(y_i, d_i|u_i) = \frac{1}{\sigma_v}\phi(v_i)\Phi(\boldsymbol{z}'_{s,i}\boldsymbol{\alpha} + \delta u_i)^{d_i}\left(1 - \Phi(\boldsymbol{z}'_{s,i}\boldsymbol{\alpha} + \delta u_i)\right)^{1-d_i}. \tag{6.16}$$

Here $v_i = y_i - \boldsymbol{x}'_i\boldsymbol{\beta} - u_i$. The unconditional density is obtained by integrating u_i out from (6.17) as

$$f(y_i, d_i) = \int_0^\infty f(y_i, d_i|u_i)f(u_i)du_i. \tag{6.17}$$

While this integral is not available in closed form, it can be solved using simulation. In this case the simulated log likelihood function is

$$\ln \mathcal{L}_s = \sum_{i=1}^{n} \ln\left(R^{-1}\sum_{r=1}^{R}\frac{1}{\tilde{\sigma}_v}\phi(\tilde{v}_i)\Phi(\boldsymbol{z}'_{s,i}\boldsymbol{\alpha} + \delta|U_{ir}|)^{d_i}\left(1 - \Phi(\boldsymbol{z}'_{s,i}\boldsymbol{\alpha} + \delta|U_{ir}|)\right)^{1-d_i}\right), \tag{6.18}$$

where $\tilde{v}_i = y_i - \boldsymbol{x}'_i\boldsymbol{\beta} + |U_{ir}|$ with U_{ir} being drawn from

$$N^+(0, \sigma_{u1})P(d_i = 1) + N^+(0, \sigma_{u0})P(d_i = 0). \tag{6.19}$$

Draws from this mixture distribution can easily be taken by first taking two draws from an independent standard normal, call these U^0_{ir} and U^1_{ir}. Then $U_{ir} = \sigma_{u1}|U^1_{ir}| \cdot d_i + \sigma_{u0}|U^0_{ir}| \cdot (1 - d_i)$. This simulated log likelihood function can be optimized directly to obtain estimates of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, σ_{u1}, σ_{u0}, σ_v and δ.

6.3.3. *What is Selection Based On?* The key difference between the selection model of Greene (2010) and that of Kumbhakar et al. (2009) is that the former is based on correlated noise between the selection and outcome equations (w_i and v_i) whereas in the later it is based on the common inefficiency term u_i appearing in both the selection and outcome equations. Although Kumbhakar et al. (2009) allowed both noise and inefficiency terms in the selection equation (the production function), the noise term in the selection equation (w_i) is assumed to be independent of the noise term v_i in the outcome equation, which is opposite to what Greene (2010) assumed. Thus, a generalization of the Kumbhakar et al. (2009) model would be allow correlation between w_i and v_i, as in Greene (2010). This would allow one to test the hypothesis that inefficiency does not drive selection (meaning $\delta = 0$) under which the Greene (2010) model is appropriate. On the other hand, one can also test the hypothesis that w_i and v_i are uncorrelated which will support the Kumbhakar

et al. (2009) model. In fact it is also possible to test whether v_i is present in the selection model by hypothesizing $\sigma_v = 0$.

Note that even if there is no noise in the Kumbhakar et al. (2009) model, selection cannot be ignored. In the current version of the Kumbhakar et al. (2009) model no selection hypothesis can be tested by hypothesizing $\delta = 0$ since w_i is assumed to be independent of v_i. Currently, a framework where selection is based on ε does not exist. This would be interesting as it may lead to the development of a test for how selection arises.

7. The Stochastic Frontier Model with Panel Data

Currently our discussion of inefficiency has been in the context of a cross-section of firms, observed at a single point in time. While this allows one to assess inefficiency, it provides a rather limited overview of how firms are operating. When these firms can be observed over time then more diverse insights can be gleaned. Further, having access to panel data enables the modeler to take into account some of the latent heterogeneity that may exist in the dataset beyond what is possible using a cross-sectional approach. That is, with a panel, some of the unobserved heterogeneity could be inefficiency or individual specific heterogeneity. *A priori* this is unknown.

Having information on units over time also enables one to examine whether inefficiency has been persistent over time or whether the inefficiency of firms is time-varying. Indeed, there may be a component of inefficiency that has been persistent over time and another that varies over time. Related to this, and a key question that needs to be considered with regards to the time-invariant individual effects, is whether the individual effects represent (persistent) inefficiency, or whether the effects are independent of the inefficiency and capture (persistent) unobserved heterogeneity.

Motivating the use of panel data to model technical efficiency, Schmidt & Sickles (1984) mention three problems with cross-sectional models that are used to measure inefficiency. First, the ML method, used to estimate parameters and the inefficiency estimates using the JLMS formula, depends on distributional assumptions for the noise and the inefficiency components. Second, the technical inefficiency component has to be independent of the regressor(s) (at least in a single equation framework) – an assumption that is unlikely to be true if firms maximize profit and inefficiency is known to the firm (see Mundlak 1961). Third, the JLMS estimator is not consistent, in the sense that the conditional mean or mode of $u|v-u$ never approaches u as the number of firms (cross-sectional units) approaches infinity.

If panel data are available, that is, each unit is observed at several different points of time, some of these rigidities can be removed. However, to overcome some of these limitations, the panel models make other assumptions, some of which may or may not be realistic. Here we review the panel models that are used in the efficiency literature.

Our discussion of available stochastic frontier methods for panel data will attempt to classify the models in terms of the assumptions made on the temporal behavior of inefficiency. We begin with the most restrictive of the models in terms of assumed behavior of inefficiency and end with recent developments that establish the identification of both time invariant and time varying inefficiency jointly. For the sake of simplicity, we consider only the case of balanced panel data, but the results of the following sections can be easily extended to unbalanced panels.

7.1. **Time-invariant Technical Inefficiency Models.**

We first consider the case in which inefficiency is assumed to be constant over time and individual specific. In this case, the unobservable individual effects of the classic panel data model is the base from which inefficiency is measured.

The model may be written as

$$\begin{aligned} y_{it} =& m(\boldsymbol{x}_{it};\boldsymbol{\beta}) + \varepsilon_{it}, \\ \varepsilon_{it} =& v_{it} - u_i, \qquad u_i \geq 0, i = 1,\ldots,N; t = 1,\ldots,T \end{aligned} \tag{7.1}$$

where $m(\boldsymbol{x}_{it};\boldsymbol{\beta})$ is a linear in parameters function of the variables in the vector $\boldsymbol{x}_{it}$, and $u_i \geq 0$ is the time-invariant technical inefficiency of individual i. This model utilizes the panel feature of the data via u_i which is specific to an individual and does not change over time.

The stochastic panel model with time invariant inefficiency can be estimated under either the fixed effects or random effects framework (Wooldridge 2010). Which framework to select depends on the level of relationship one is willing to assume between inefficiency and the covariates of the model. Under the fixed effects framework correlation is allowed between $\boldsymbol{x}_{it}$ and u_i, whereas under the random effects framework no correlation is present between $\boldsymbol{x}_{it}$ and u_i.

Neither of these approaches require distributional assumptions on u_i and are, thus, labeled as distribution free approaches. These models are discussed in detail in Schmidt & Sickles (1984). Note that the model in (7.1) is the same as the one-way error component model widely discussed in the panel data literature,[31] except that the individual effects are assumed to be one-sided. The idea is to make a simple transformation and interpret the transformed individual effects as time-invariant inefficiency as opposed to pure firm heterogeneity.

One (current) limitation of the above time-invariant inefficiency model is that separate identification of inefficiency and individual heterogeneity is not considered. Recent developments in the efficiency literature have proposed models that allow both effects to be identified and estimated. Examples include the true fixed-effect and true random-effect models to be discussed later.

7.1.1. *Estimation under the Fixed Effects Framework.* For ease of exposition, we assume $f(\cdot)$ is linear in $\boldsymbol{x}_{it}$ (e.g., the log of input quantities in a Cobb-Douglas production function model). The time-invariant inefficiency panel data stochastic frontier panel model can then be written as

$$y_{it} = \beta_0 + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - u_i \tag{7.2}$$

$$\begin{aligned} =& (\beta_0 - u_i) + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} \\ =& \alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} \end{aligned} \tag{7.3}$$

where $\alpha_i \equiv \beta_0 - u_i$. Under the fixed effects framework, u_i and thus α_i, $i = 1,\ldots,N$ are allowed to have arbitrary correlation with $\boldsymbol{x}_{it}$.

The model in (7.3) looks similar to a standard fixed-effects (FE) panel data model. Schmidt & Sickles (1984) presented standard FE panel data estimation methods (for example, the within estimator) to estimate the model. The standard FE panel methods yield consistent estimates of $\boldsymbol{\beta}$, but $\hat{\alpha}_i$ is a biased estimator of u_i because $u_i > 0$ by construction. Nevertheless, after $\hat{\alpha}_i$ is obtained a simple transformation can be applied to recover $\hat{u}_i \geq 0$ which will be consistent provided $T \to \infty$.

[31] See, for example, Baltagi (2013) and Hsiao (2014), amongst others.

An important implication under the fixed effects framework is that u_i is allowed to be freely correlated with $\boldsymbol{x}_{it}$ in the model. This may be a desirable property for empirical applications in which inefficiency is believed to be correlated with the inputs used (Mundlak 1961). A disadvantage of the FE approach, on the other hand, is that no other time-invariant variables, such as gender, race, region, etc., can be included in $\boldsymbol{x}_{it}$ because doing so entails perfect multicollinearity between the α_i and the time-invariant regressors.

The above model may be estimated using OLS after including individual dummies as regressors for α_i. This technique is often referred to as the least square dummy variable (LSDV) method. The coefficients of the dummies are the estimates of α_i. Notice that, since there is one FE parameter for each cross-sectional unit (individual, firm, etc.), the number of dummies to be included in the model is equal to the number of cross-sectional units in the data (when no intercept model is chosen). For a panel data set with many cross-sectional units, estimation might be an issue because it requires inverting a $(N + K) \times (N + K)$ matrix where N is the number of cross-sectional units and K is the number of regressors ($\boldsymbol{x}_{it}$ variables).

This difficulty can be easily overcome by transforming the model before estimation to remove α_i. The transformation can be carried out either using a first-difference or within transformation. For example, the within transformation subtracts cross-sectional means of the data from each cross section (e.g., replacing y_{it} by $y_{it} - \bar{y}_{i\cdot}$ and x_{it} by $x_{it} - \bar{x}_{i\cdot}$, where $\bar{y}_{i\cdot} = (1/T)\sum_t y_{it}$, etc.), thereby eliminating α_i. The resulting model can then be easily estimated by OLS (which requires inversion of a $K \times K$ matrix). The values of $\hat{\alpha}_i$ are recovered from the mean of the residuals for each cross sectional unit, i.e., $\hat{\alpha}_i = \bar{y}_{i\cdot} - x'_{i\cdot}\hat{\beta}$. The transformed models yield consistent estimates of $\boldsymbol{\beta}$ for either T or $N \to \infty$. Consistency of $\hat{\alpha}_i$, however, requires $T \to \infty$.

Once the $\hat{\alpha}_i$ are available, the following transformation is used to obtain estimated value of $\hat{u}_i$ (Schmidt & Sickles 1984):

$$\hat{u}_i = \max_i\{\hat{\alpha}_i\} - \hat{\alpha}_i \geq 0, \quad i = 1, \cdots, N. \tag{7.4}$$

This formulation implicitly assumes that the most efficient unit in the sample is 100% efficient. In other words, estimated inefficiency in the fixed-effects model is relative to the best unit in the sample. If one is interested in estimating firm-specific technical efficiency, it can obtained from

$$\widehat{TE}_i = e^{-\hat{u}_i}, \quad i = 1, \cdots, N. \tag{7.5}$$

7.1.2. *Estimation Under the Random Effects Framework.* Rather than explicitly allowing for correlation between $\boldsymbol{x}_{it}$ and unobserved time invariant firm inefficiency, it might be plausible to assume that α_i is uncorrelated with $\boldsymbol{x}_{it}$. When the assumption of no correlation between the covariates and firm inefficiency is indeed correct, then estimation of the stochastic frontier panel data model under the random effects framework provides more efficient estimates than estimation under the fixed effects framework. An important empirical advantage of the random effects framework is that time-invariant variables, such as gender, race, etc., may be included in the $\boldsymbol{x}_{it}$ vector of explanatory variables without leading to collinearity with α_i.

Estimation of the stochastic frontier panel data model under the random effects framework can be estimated in several different fashions. First, one can eschew distributional assumptions and use generalized least squares (GLS) as is common for standard estimation of the linear in parameters panel data model in the random effects framework. Similar to estimation under the fixed effects framework, the GLS estimates of unobserved individual heterogeneity need to be modified and reinterpreted to recover estimates of inefficiency. An alternative to using GLS is to explicitly impose distributional assumptions on the random components of the model, and estimate the parameters of the model by maximum likelihood. This approach was originally proposed by Pitt & Lee (1981). Once the parameters are estimated via ML , JLMS type conditional mean estimators can be used to estimate firm-specific inefficiency Kumbhakar (1987).

7.1.2.1. **The Distribution Free Approach: GLS Estimation**

Assume $E(u_i) = \mu$ and denote $u_i^* = u_i - \mu$. We rewrite the model as

$$\begin{aligned} y_{it} =& \beta_0 + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - u_i \\ =& (\beta_0 - \mu) + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - u_i^* \\ =& \alpha^* + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - u_i^*, \end{aligned} \tag{7.6}$$

where $\alpha^* \equiv \beta_0 - \mu$. The model in (7.6) can be estimated using GLS to account for the heteroscedastic nature of the composed term, $v_{it} - u_i^*$. Following Baltagi (2013), the slope coefficients and intercept can be estimated by regressing $\breve{y}_{it}$ on $\breve{\boldsymbol{x}}_{it}$ where $\breve{y}_{it} = y_{it} - \theta\bar{y}_{i\cdot}$ and $\theta = 1 - \sigma_v/\sigma_1$ and $\sigma_1^2 = T\sigma_u^2 + \sigma_v^2$. A variety of approaches exist to construct estimates of the variance components to deploy a feasible GLS estimator.

If we define $\hat{\varepsilon}_{it} = y_{it} - \boldsymbol{x}_{it}'\hat{\boldsymbol{\beta}}$, then an estimate of $\alpha_i \equiv \beta_0 - u_i$ may be derived by the time average of $\hat{\varepsilon}_{it}$ for each cross-sectional, viz.,

$$\hat{\alpha}_i = \frac{1}{T}\sum_t (\hat{\varepsilon}_{it} - \hat{\alpha}^*), \quad i = 1, \cdots, N. \tag{7.7}$$

Here, we use the implicit assumption that the time average of $\hat{v}_{it}$ is zero which is true as $T \to \infty$. Finally, the estimate of firm-specific inefficiency, $\hat{u}_i$, is obtained using (7.4).

An alternative approach is to take a Bayesian approach and estimate u_i^* using the best linear unbiased predictor (BLUP), which is

$$\hat{u}_i^* = -\left\{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_v^2 + T\hat{\sigma}_u^2}\right\}\sum_t \hat{\varepsilon}_{it}, \quad i = 1, \cdots, N. \tag{7.8}$$

Then the estimate of firm-specific inefficiency, $\hat{u}_i$, is obtained from

$$\hat{u}_i = \max_i\{\hat{u}_i^*\} - \hat{u}_i^* \geq 0, \quad i = 1, \cdots, N. \tag{7.9}$$

7.1.2.2. **Distributional Assumptions Required: Maximum Likelihood Estimation**

We can estimate (7.2) using ML by imposing distributional assumptions on v_{it} and u_i, as in Pitt & Lee (1981). Their model is specified as:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - u_i, \tag{7.10}$$

where u_i has a half normal distribution with variance parameter σ_u^2 and v_{it} has a normal distribution. This is identical to the cross-sectional stochastic frontier model of Aigner et al. (1977) except all of the variables, except inefficiency, are in a panel setting.

The likelihood function for the ith observation is (see Pitt & Lee 1981)

$$\begin{aligned}\ln L_i =& \text{constant} + \ln \Phi\left(\frac{\mu_{i*}}{\sigma_*}\right) + \frac{1}{2}\ln(\sigma_*^2) - \frac{1}{2}\left\{\frac{\sum_t \varepsilon_{it}^2}{\sigma_v^2} + \left(\frac{\mu}{\sigma_u}\right)^2 - \left(\frac{\mu_{i*}}{\sigma_*}\right)^2\right\} \\ & - T\ln(\sigma_v) - \ln(\sigma_u) - \ln\Phi\left(\frac{\mu}{\sigma_u}\right),\end{aligned} \tag{7.11}$$

where

$$\mu_{i*} = \frac{\mu\sigma_v^2 - \sigma_u^2\sum_t \varepsilon_{it}}{\sigma_v^2 + T\sigma_u^2}, \tag{7.12}$$

$$\sigma_*^2 = \frac{\sigma_v^2\sigma_u^2}{\sigma_v^2 + T\sigma_u^2}. \tag{7.13}$$

The log-likelihood function of the model is obtained by summing $\ln L_i$ over i, $i = 1, \ldots, N$. The MLE of the parameters is obtained by maximizing the log-likelihood function.

After estimating the parameters of the model, inefficiency for each i can be computed from either the mean or the mode (Kumbhakar & Lovell 2000, p. 104)

$$E(u_i|\varepsilon_i) = \mu_{i*} + \sigma_*\left[\frac{\phi\big(-\mu_{i*}/\sigma_*\big)}{1 - \Phi\big(-\mu_{i*}/\sigma_*\big)}\right] \tag{7.14}$$

and

$$M(u_i|\varepsilon_i) = \left\{\begin{array}{ll} \mu_{i*} & \text{if } \mu_{\text{i}*} \geq 0, \\ 0 & \text{otherwise} \end{array}\right\} \tag{7.15}$$

which are the extended JLMS estimators of inefficiency. We may set $\mu = 0$ in the above equations for the half-normal distribution of u_i or make μ a function of exogenous variables ($\mu = \boldsymbol{z}_i'\boldsymbol{\delta}$) to accommodate determinants of inefficiency. Note that both estimators of inefficiency are consistent as $T \to \infty$ (Kumbhakar 1987).

7.2. **Time-varying Technical Inefficiency Models.**

The array of models introduced in the previous section assume technical inefficiency to be individual-specific and time-invariant. That is, the inefficiency levels may be different for different individuals, but they do not change over time; unfortunately the implication of this framework is that an inefficient unit (e.g., a firm) never learns over time. This might be the case in some situations where, for example, inefficiency is associated with managerial ability and there is no change in management for any of the firms during the

period of the study or if the time dimension of the panel is particularly short. Even this is, at times, unrealistic, particularly when market competition is taken into account. To accommodate the notion of productivity and efficiency improvement, we need to consider models that allow inefficiency to change over time.

Models in which the inefficiency effects are time-varying are more general than the time-invariant models, in the sense that the time invariant models can be viewed as special cases of the time varying models. Much is similar when moving to a time-varying inefficiency framework; for instance, we can build models under the fixed or random effects framework.

7.2.1. *Distribution Free Approaches.*

7.2.1.1. **The Cornwell, Schmidt, and Sickles (1990) Model**

Recall the Schmidt & Sickles (1984) model in (7.3):

$$y_{it} = \alpha_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + v_{it}, \tag{7.16}$$

where $\alpha_i \equiv \beta_0 - u_i$ and the inefficiency term (confounded in the firm-specific effect) is time-invariant. To allow this effect to be time-varying, Cornwell, Schmidt & Sickles (1990) suggest replacing α_i by α_{it} where

$$\alpha_{it} = \alpha_{0i} + \alpha_{1i}t + \alpha_{2i}t^2. \tag{7.17}$$

Note that the parameters α_{0i}, α_{1i} and α_{2i} are firm-specific and t is the time trend variable. Hereafter, we denote the above model as the CSS model. If the number of cross-sectional units (N) is not large, one can define N firm dummies and interact these dummies with time and time squared. These variables along with the regressors (i.e., the $\boldsymbol{x}$ variables) are then used in the OLS regression. Since all the firm-dummies are used, the intercept term in the regression has to be suppressed to avoid the exact multicollinearity problem. The coefficients associated with the firm dummies and their interactions are the estimates of α_{0i}, α_{1i} and α_{2i}. These estimated coefficients can be used to obtain estimates of α_{it}.

More generally, if we represent the model as

$$y_{it} = \alpha_{0i} + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}; \quad \varepsilon_{it} \equiv v_{it} + \alpha_{1i}t + \alpha_{2i}t^2 \tag{7.18}$$

then the form of the model looks like a standard panel data model. Similar to the model of Schmidt & Sickles (1984), we may apply the within estimator to obtain consistent estimates of $\hat{\boldsymbol{\beta}}$ in (7.18), and then the estimated residuals of the model ($\hat{\varepsilon}_{it} \equiv y_{it} - \boldsymbol{x}'_{it}\hat{\boldsymbol{\beta}}$). These residuals are then regressed on a constant, a time trend, and the square of the time trend for *each* i. The fitted values from these regressions provide estimates of α_{it} in (7.17). Finally, $\hat{u}_{it}$ is obtained from

$$\hat{u}_{it} = \hat{\alpha}_t - \hat{\alpha}_{it} \quad \text{and} \quad \hat{\alpha}_t = \max_j(\hat{\alpha}_{jt}). \tag{7.19}$$

That is, for each t we define the maximum and calculate efficiency relative to the best firm in that year. Since the maximum of $\hat{\alpha}_{jt}$ is likely to change over time, the same firm may not be

efficient in every year. In other words, efficiency in the above framework is relative to the best firm in the sample *in a given year* and this may be represented by different firms in different years. Alternatively, if one defines the maximum over all i and t, then efficiency is relative to the best firm in the sample (defined over all time periods).

The estimation procedure outlined above is easy to implement. It mainly relies on the standard panel data estimators within the fixed effects framework. It should be noted that since t appears in the inefficiency function, it cannot appear as a regressor in $\boldsymbol{x}_{it}$ to capture technical change (a shift in the production/cost function, $m(\boldsymbol{x})$). In other words, the above model cannot separate inefficiency from technical change.

7.2.1.2. **The Lee and Schmidt (1993) Model**

The advantage of the CSS model is the modeling flexibility of inefficiency. The temporal behavior of inefficiency is flexible enough to allow efficiency to increase or decrease and to differ across cross-sectional units. A problem of the CSS model is that it may be over parameterized in the specification of inefficiency. In a model with large N and small T the model will have too many parameters ($3N$ parameters in the α_{it} function alone).

A somewhat parsimonious time-varying inefficiency model was proposed by Lee & Schmidt (1993) in which u_{it} is specified as

$$u_{it} = u_i \lambda_t, \tag{7.20}$$

where λ_t, $t = 1, \ldots, T$, are time specific effects to be estimated. Although the model is quite flexible because no parametric function is assumed for the temporal behavior of inefficiency, the down-side (compared to the CSS model) is that the temporal pattern of inefficiency is exactly the same for all firms (λ_t). Under the fixed effects framework, this specification can be viewed as an interactive effects panel data model and estimation can be undertaken by introducing both firm and time dummies. Note that appropriate identification restrictions need to be imposed, since all the coefficients of firm-and time-dummies cannot be estimated due to multicollinearity.

Although Lee & Schmidt (1993) considered estimation under both the fixed and random effects framework, here we present their model slightly differently. The reason for this is to make their model comparable to several other popular time-varying inefficiency models: Kumbhakar (1990) and Battese & Coelli (1992). In this vein, we assume u_i to be random and λ_t to be a fixed parameter. Since λ_t is a parameter (i.e., the coefficient on a time dummy variable, TD_t, after an arbitrary normalization of one coefficient being set to unity for identification purposes).

The Lee-Schmidt model is much more general than either the Kumbhakar (1990), Battese & Coelli (1992) models. Both of these models can be derived as a special case of the Lee-Schmidt model by imposing appropriate restrictions on λ_t. For example, we write the Battese Coelli inefficiency as $u_i \times \exp(-\gamma(t-T)), t = 1, \ldots T$, it will be a special case of the Lee-Schmidt under under the restrictions $\lambda_T = 1, \lambda_{T-1} = \exp(-\gamma), \lambda_{T-2} = \exp(-2\gamma), \ldots, \lambda_1 = \exp(-\gamma(1-T))$. Note that the

Battese-Coelli formulation has one parameter while the Lee-Schmidt model has $(T-1)$ parameters (one is normalized to unity for identification).

Similarly, the time-invariant inefficiency model can be derived as a special case from all these model. That is, if $\lambda_t = 1 \; \forall \; t$, then the model reduces to the time-invariant inefficiency model.

Estimation of this model is similar to the models discussed below where u_i is a one-sided random variable and $\lambda_t \equiv G(t)$. Because of this similarity we skip the details. Once λ_t and u_i are estimated, technical inefficiency can be estimated from

$$\hat{u}_{it} = \max_i \{\hat{u}_i \hat{\lambda}_t\} - \{\hat{u}_i \hat{\lambda}_t\}. \tag{7.21}$$

7.3. **Time-varying Inefficiency Models with Deterministic and Stochastic Components.** The time-varying models that we have considered so far model inefficiency in a completely deterministic fashion. For example, in the Lee-Schmidt model both the components of u_i and λ_t are deterministic. In fact, the model can be estimated assuming u_i is random while λ_t is a deterministic function of time (e.g., time dummies). Although Lee and Schmidt estimated this model without any distributional assumptions on u_i, we will consider these type of models with distributional assumptions on u_i. We use the following generic formulation to discuss the various models in a unifying framework, viz.,

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + v_{it} - u_{it}, \tag{7.22}$$

where $u_{it} = G(t)u_i$ with $G(t) > 0$ being a function of time (t). This is just a generalization of the cross-sectional stochastic frontier model with the scaling property discussed in Section 5 to the panel data setting. To stay consistent with the different models that have appeared in the literature using this framework we assume that $v_{it} \sim N(0, \sigma_v^2)$ and $u_i \sim N^+(\mu, \sigma_u^2)$. Note that the scaling property here fits in with Wang & Schmidt's (2002) description of base inefficiency, albeit here, only time influences improvements in inefficiency.

In this model, inefficiency is not fixed for a given individual; instead, it changes over time and also across individuals. Inefficiency in this model is composed of two distinct components: one is a non-stochastic time component, $G(t)$, and the other is a stochastic individual component, u_i. It is the stochastic component, u_i, that utilizes the panel structure of the data in this model. The u_i component is individual-specific and the $G(t)$ component is time-varying and is common across individuals.

Given $u_i \geq 0$, $u_{it} \geq 0$ is ensured by having a non-negative $G(t)$. There are several popular forms for $G(t)$ in the literature. Kumbhakar's (1990) model assumes

$$G(t) = \left[1 + \exp(\gamma_1 t + \gamma_2 t^2)\right]^{-1}, \tag{7.23}$$

while Lee & Schmidt's (1993) model assumes

$$G(t) = \lambda_t \, TD_t, \tag{7.24}$$

where λ_t are the coefficients associated with the time dummy variables, TD_t. Note that in this model, u_{it} is not restricted to be positive because $G(t)$ is not restricted to be positive. This model has not appear much in the empirical literature.

Perhaps the most commonly applied time-varying inefficiency model[32] is that of Battese & Coelli (1992) which assumes

$$G(t) = \exp\left[\gamma(t-T)\right], \tag{7.25}$$

where T is the terminal period of the sample. The Kumbhakar (1990) and Battese & Coelli (1992) specifications can be directly comparable if one writes the Kumbhakar formulation as

$$G(t) = 2 \times \left[1 + \exp(\gamma_1(t-T) + \gamma_2(t-T)^2)\right]^{-1}. \tag{7.26}$$

Then, at the terminal point $u_{iT} = u_i$. Finally, a more recent specification appears in Kumbhakar & Wang (2005) who use

$$G(t) = \exp\left[\gamma(t-\underline{t})\right], \tag{7.27}$$

where $\underline{t}$ is the beginning period of the sample.

Analytically, (7.25) and (7.27) are the same, but they are interpreted differently. In the Battese & Coelli (1992) and the reformulated specification of Kumbhakar (1990), $u_i \sim N^+(\mu, \sigma_u^2)$ specifies the distribution of inefficiency at the terminal point, i.e., $u_{it} = u_i$ when $t = T$. With (7.27), $u_i \sim N^+(\mu, \sigma_u^2)$ specifies the initial distribution of inefficiency. Depending on the specific application, one specification may be preferred over the other. We note in passing that the Kumbhakar & Wang (2005) model also controls for firm specific heterogeneity within the fixed effects framework, a feature that the other previously mentioned models do not capture.

Note that the time-invariant random-effects model can be obtained as a special case from all these models by imposing appropriate parametric restrictions. Further, these restrictions can be tested using the LR test. For example, if $\gamma = 0$, then the Kumbhakar & Wang (2005) model and the Battese & Coelli (1992) model collapse to the time-invariant RE model in (7.10). Similarly, the Lee & Schmidt (1993) model becomes a time-invariant random-effects model in (7.10) if $\lambda_t = 1 \; \forall \; t$, i.e., the coefficients of the time dummies are all 1. Finally, the Kumbhakar (1990) model reduces to the time-invariant random-effects model in (7.10) if $\gamma_1 = \gamma_2 = 0$. Because the Kumbhakar (1990) model has two parameters in the $G(t)$ function, the temporal pattern is more flexible than what is capable in the specifications of $G(t)$ for the Battese & Coelli (1992) and the Kumbhakar & Wang (2005) models. It is worth noting that the Lee & Schmidt (1993) model is quite general and the other models discussed here (as well as other models which specify $G(t)$ parametrically) can be viewed as a special case of it.[33] In the original random-effects version of the Lee & Schmidt (1993) model, no distributional assumption was made on u_i. However, to compare it with the Kumbhakar

[32]This is more likely due to the dissemination of the freely available statistical package FRONTIER which implements this model at the push of a button rather than some realized theoretical or economic advantage.

[33]Cuesta (2000) considered a model in which the parameters of the $G(t)$ function are made firm-specific, i.e., in his model $G(t)$ is $G_i(t)$ with firm specific parameters in $G_i(t)$.

(1990), Battese & Coelli (1992), and Kumbhakar & Wang (2005) models similar assumptions on u_i are required to make proper comparisons.

Given the similarity of all these models we specify the log-likelihood function of each cross sectional observation i in generic form, i.e., the one in (7.22), which is (Kumbhakar & Lovell 2000, p. 111)

$$\begin{aligned}\ln L_i =& \text{constant} + \ln \Phi\left(\frac{\mu_{i*}}{\sigma_*}\right) + \frac{1}{2}\ln(\sigma_*^2) - \frac{1}{2}\left\{\frac{\sum_t \varepsilon_{it}^2}{\sigma_v^2} + \left(\frac{\mu}{\sigma_u}\right)^2 - \left(\frac{\mu_{i*}}{\sigma_*}\right)^2\right\} \\ & - T\ln(\sigma_v) - \ln(\sigma_u) - \ln\Phi\left(\frac{\mu}{\sigma_u}\right), \end{aligned} \tag{7.28}$$

where

$$\mu_{i*} = \frac{\mu\sigma_v^2 - \sigma_u^2 \sum_t G(t)\varepsilon_{it}}{\sigma_v^2 + \sigma_u^2 \sum_t G(t)^2}, \tag{7.29}$$

$$\sigma_*^2 = \frac{\sigma_v^2\sigma_u^2}{\sigma_v^2 + \sigma_u^2 \sum_t G(t)^2}. \tag{7.30}$$

The log-likelihood function of the model is obtained by summing $\ln L_i$ over i, $i = 1, \ldots, N$, which is maximized to get the ML estimates of the parameters.

Once the parameter estimates are obtained, inefficiency can be predicted from either the mean or the mode (Kumbhakar & Lovell 2000, p. 111)

$$E(u_i|\varepsilon_i) = \mu_{i*} + \sigma_*\left[\frac{\phi\big(-\mu_{i*}/\sigma_*\big)}{1 - \Phi\big(-\mu_{i*}/\sigma_*\big)}\right] \tag{7.31}$$

and

$$M(u_i|\varepsilon_i) = \begin{cases} \mu_{i*} & \text{if } \mu_{\text{i}*} \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{7.32}$$

7.4. **Models that Separate Firm Heterogeneity from Inefficiency.** As mentioned repeatedly, the time-invariant inefficiency stochastic frontier panel data model is a standard panel data model where α_i is the unobservable individual effect and standard panel data fixed- and random-effects estimators are applied here to estimate the model parameters including α_i. The only difference is that we transform the estimated value of $\hat{\alpha}_i$ to obtain estimates of u_i. A notable drawback of this approach is that individual heterogeneity cannot be distinguished from inefficiency: all time-invariant heterogeneity is confounded with inefficiency, and therefore $\hat{u}_i$ will capture heterogeneity in addition to, or even instead of, inefficiency (Greene 2005*b*). Another potential issue of the model is the time-invariant assumption of inefficiency. If T is large, it seems implausible that the level of inefficiency of a firm may stay constant for an extended period of time or that a firm which was persistently inefficient would survive in the market.

So the question is: Should one view the time-invariant component as persistent inefficiency (as per Kumbhakar 1991, Kumbhakar & Heshmati 1995, Kumbhakar & Hjalmarsson 1993, Kumbhakar

& Hjalmarsson 1998) or as individual heterogeneity that captures the effects of (unobserved) time-invariant covariates and has nothing to do with inefficiency? If the latter setting holds, then the results from the time-invariant inefficiency models are incorrect. A less rigid perspective is that the truth lies somewhere in the middle; inefficiency may be decomposed into a component that is persistent over time and a component that varies over time. Unless persistent inefficiency is separated from the time-invariant individual effects, one has to choose either the model in which α_i represents persistent inefficiency or the model in which α_i represents an individual-specific effect (heterogeneity). Here, we will consider both specifications. In particular, we will consider models in which inefficiency is time-varying irrespective of whether the time-invariant component is treated as inefficiency or not. Thus, the models we will focus on is

$$y_{it} = \alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - u_{it}. \tag{7.33}$$

Compared to a standard panel data model, we have the additional time-varying inefficiency term, $-u_{it}$, in (7.33). If one treats α_i, $i = 1, \cdots, N$ as a random variable that is correlated with $\boldsymbol{x}_{it}$ but does not capture inefficiency, then the above model becomes what has been termed the 'true fixed-effects' panel stochastic frontier model (Greene 2005*a*). The model is labeled as the 'true random-effects' stochastic frontier model when α_i is treated as uncorrelated with $\boldsymbol{x}_{it}$. Note that these specifications are not different from the models proposed by Kumbhakar and coauthors mentioned above. The difference is in the interpretation of the 'time-invariant term', α_i.

Estimation of the model in (7.33) is not straightforward. When α_i, $i = 1, \cdots, N$, are embedded in the fixed effects framework, the model encounters the incidental parameters problem (Neyman & Scott 1948). The incidental parameters problem arises when the number of parameters to be estimated increases with the number of cross-sectional units in the data, which is the case with the α_i in (7.33). In this situation, consistency of the parameter estimates is not guaranteed even if $N \to \infty$ because the number of α_i increases with N. Therefore, usual asymptotic results may not apply. In addition to this specific statistical problem, another technical issue in estimating (7.33) is that the number of parameters to be estimated can be prohibitively large for large N.

For a standard linear panel data model (i.e., one that does not have $-u_{it}$ in (7.33)), the literature has developed estimation methods to deal with this problem. The methods involve transforming the model so that α_i is removed before estimation. Without α_i in the transformed model, the incidental parameters problem no longer exists and the number of parameters to be estimated is not large. Methods of transformation include conditioning the model on α_i's sufficient statistic [34] to obtain the conditional MLE, and the within-transformation model or the first-difference transformation model to construct the marginal MLE (e.g., Cornwell & Schmidt 1992).

These standard methods, however, are usually not applicable to (7.33). For the conditional MLE of (7.33), Greene (2005*b*) showed that there is no sufficient statistic of α_i. For the marginal MLE, the resulting model after the within or first-difference transformation usually does not have

[34] A sufficient statistic contains all the information needed to compute any estimate of the parameter.

a closed form likelihood function, if one uses standard procedures.[35] In general this would not pose an issue as regression methods can be easily applied. However, given the precise interest in recovering estimates of the parameters of the distribution of inefficiency, maximum likelihood or specific moments of the distribution of the transformed error component are needed. This precipitates methods that can recover information regarding u_{it}.

Greene (2005*b*) proposed a tentative solution. He assumed u_{it} follows a simple i.i.d. half-normal distribution and suggested including N dummy variables in the model for α_i, $i = 1, \cdots, N$ and then estimating the model by MLE without any transformation. He found that the incidental parameters problem does not cause significant bias to the model parameters when T is large. The problem of having to estimate more than N parameters is dealt with by employing an advanced numerical algorithm.

There are some recent econometric developments on this issue. First, Chen et al. (2014) proposed a solution in the fixed-effects framework. Following a theorem in Domínguez-Molina, González-Farías & Ramos-Quiroga (2003), they showed that the likelihood function of the within transformed and the first-difference model have closed form expressions. The same theorem in Domínguez-Molina et al. (2003) is used in Colombi, Kumbhakar, Martini & Vittadini (2014) to derive the log-likelihood function in the random effects framework. We will discuss these models in more detail in Section 7.6.

Using a different approach, Wang & Ho (2010) solve the problem classified in Greene (2005*b*) by proposing a class of stochastic frontier models in which the within and first-difference transformation on the model can be carried out while also providing a closed form likelihood function. The main advantage of such a model is that because the α_is are removed from the model, the incidental parameters problem is avoided entirely. As such, consistency of the estimates is obtained for either $N \to \infty$ or $T \to \infty$, which is invaluable for applied settings. A further computational benefit is that the elimination of α_is reduces the number of parameters to be estimated to a manageable number. Formally, the Wang & Ho (2010) model is:

$$y_{it} = \alpha_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \tag{7.34}$$

with $v_{it} \sim N(0, \sigma_v^2)$, $u_{it} = g_{it}u_i^*$ and $u_i^* \sim N^+(\mu, \sigma_u^2)$, the now familiar scaling property model with a truncated normal distribution for the basic distribution of inefficiency. For the scaling function Wang & Ho (2010) set $g_{it} = g(\boldsymbol{z}'_{u,it}\boldsymbol{\delta}^u)$. The key feature that allows the model transformation to be applied is the scaling property. As u_i^* does not change with time, the within and the first-difference transformations leave this stochastic term intact. As Wang & Ho (2010) show that the within-transformed and the first-differenced models are algebraically identical we only provide discussion on the first-differenced model.

[35] Colombi, Martini & Vittadini (2011) showed that the likelihood function has a closed form expression. But this involves knowledge of a closed skew-normal distribution – something that has not been very well known in stochastic frontier models until recently. See, for example, Chen, Schmidt & Wang (2014).

Using the notation $\Delta w_{it} = w_{it} - w_{it-1}$ for variable w_{it}, and letting the stacked vector of Δw_{it}, for a given i and $t = 2, \ldots, T$, be defined as $\Delta\tilde{w}_i = (\Delta w_{i2}, \Delta w_{i3}, \ldots, \Delta w_{iT})'$ the log-likelihood function for the ith cross-sectional unit is (see Wang & Ho 2010, p. 288)

$$\begin{aligned}\ln \mathcal{L}_i^D = &-\frac{1}{2}(T-1)\ln(2\pi) - \frac{1}{2}\ln(T) - \frac{1}{2}(T-1)\ln(\sigma_v^2) - \frac{1}{2}\Delta\tilde{\varepsilon}_i'\Sigma^{-1}\Delta\tilde{\varepsilon}_i \\ &+ \frac{1}{2}\left(\frac{\mu_*^2}{\sigma_*^2} - \frac{\mu^2}{\sigma_u^2}\right) + \ln\left(\sigma_*\Phi\left(\frac{\mu_*}{\sigma_*}\right)\right) - \ln\left(\sigma_u\Phi\left(\frac{\mu}{\sigma_u}\right)\right),\end{aligned} \tag{7.35}$$

where

$$\mu_{*i} = \frac{\mu/\sigma_u^2 - \Delta\tilde{\varepsilon}_i'\Sigma^{-1}\Delta\tilde{h}_i}{\Delta\tilde{h}_i'\Sigma^{-1}\Delta\tilde{h}_i + 1/\sigma_u^2}; \quad \sigma_{*i}^2 = \frac{1}{\Delta\tilde{h}_i'\Sigma^{-1}\Delta\tilde{h}_i + 1/\sigma_u^2},$$

and $\Delta\tilde{\varepsilon}_i = \Delta\tilde{y}_i - \Delta\tilde{\boldsymbol{x}}_i\boldsymbol{\beta}$ with the $(T-1)\times(T-1)$ variance-covariance matrix Σ of $\Delta\tilde{v}_i$ is

$$\Sigma = \begin{bmatrix} 2\sigma_v^2 & -\sigma_v^2 & 0 & \ldots & 0 \\ -\sigma_v^2 & 2\sigma_v^2 & -\sigma_v^2 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -\sigma_v^2 \\ 0 & 0 & \ldots & -\sigma_v^2 & 2\sigma_v^2 \end{bmatrix}. \tag{7.36}$$

The matrix has $2\sigma_v^2$ on the diagonal and $-\sigma_v^2$ on the off-diagonals. The final log-likelihood function is

$$\ln \mathcal{L} = \sum_{i=1}^{N} \ln \mathcal{L}_i^D.$$

After the model parameters have been estimated, the observation-specific inefficiency index is computed from

$$E(u_{it}|\Delta\tilde{\varepsilon}_i) = h_{it}\left[\mu_{*i} + \sigma_{*i}\left\{\frac{\phi(\mu_{*i}/\sigma_{*i})}{\Phi(\mu_{*i}/\sigma_{*i})}\right\}\right] \tag{7.37}$$

evaluated at $\Delta\tilde{\varepsilon}_i = \Delta\hat{\tilde{\varepsilon}}_i$. The model of Wang & Ho (2010) represents another demonstration of the usefulness of the scaling property in applied settings. However, a limitation of their model is that it does not completely separate persistent and time-varying inefficiency, a subject which we now turn our attention to.

7.5. **Models that Separate Persistent and Time-varying Inefficiency.** Although several of the models discussed earlier can separate firm-heterogeneity from time-varying inefficiency (which is either modeled as the product of a time-invariant random variable and a deterministic function of covariates or distributed i.i.d. across firms and over time), none of these models consider persistent technical inefficiency. Identifying the magnitude of persistent inefficiency is important, especially in short panels, because it reflects the effects of inputs like management (Mundlak 1961) as well as other unobserved inputs which vary across firms but not over time. Thus, unless there is a change in something that affects the management practices at the level of the firm (such as changes in

ownership or new government regulations), it is unlikely that persistent inefficiency will change. Alternatively, time varying efficiency can change over time without operational changes in the firm.

This distinction between the time-varying and persistent components is important from a policy perspective as each yields different implications. Colombi et al. (2014) refer to time-varying inefficiency as short-run inefficiency and mention that it can arise due to failure in allocating resources properly in the short run. They argue that for example, a hospital with excess capacity may increase its efficiency in the short-run by reallocating the work force across different activities. Thus, some of the physicians' and nurses' daily working hours might be changed to include other hospital activities such as acute discharges. This is a short-run improvement in efficiency that may be independent of short-run inefficiency levels in the previous period, which can justify the assumption that u_{it} is i.i.d. However, this does not impact the overall management of the hospital and so is independent from time invariant inefficiency.

To help formalize this issue more clearly we consider the model

$$y_{it} = \beta_0 + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} = \beta_0 + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - (u_i + \tau_{it}) \tag{7.38}$$

The error term, ε_{it}, is decomposed as $\varepsilon_{it} = v_{it} - u_{it}$ where u_{it} is technical inefficiency and v_{it} is statistical noise. The technical inefficiency part is further decomposed as $u_{it} = u_i + \tau_{it}$ where u_i is the persistent component (for example, time-invariant ownership) and τ_{it} is the residual (time-varying) component of technical inefficiency, both of which are non-negative. The former is only firm-specific, while the latter is both firm- and time-specific.

Such a decomposition is desirable because, since u_i does not change over time, if a firm or government wants to improve efficiency, then some change in management or policy needs to occur. Alternatively, u_i also does not fully capture inefficiency because it does not account for learning over time since it is time invariant. The residual component can capture this aspect. In this model the size of overall inefficiency, as well as the components, are important to know because they convey different types of information. Thus, for example, if the residual inefficiency component for a firm is relatively large in a particular year then it may be argued that inefficiency is caused by something which is unlikely to be repeated in the next year. On the other hand, if the persistent inefficiency component is large for a firm, then it is expected to operate with a relatively high level of inefficiency over time, unless some changes in policy and/or management take place. Thus, a high value of u_i is of more concern from a long term point of view because of its persistent nature than is a high value of τ_{it}.

The advantage of the present specification is that one can test the presence of the persistent nature of technical inefficiency without imposing any parametric form of time-dependence. By including time in the $\boldsymbol{x}_{it}$ vector, we separate exogenous technical change from technical inefficiency.

To estimate the model we rewrite (7.38) as

$$y_{it} = \alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \omega_{it} = \beta_0 - u_i - E(\tau_{it}) + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - (\tau_{it} - E(\tau_{it})). \tag{7.39}$$

The error, ω_{it}, has zero mean and constant variance. Thus, the model in (7.39) fits perfectly into the standard panel data model with firm-specific effects (one-way error component model), and can be estimated either by the least-squares dummy variable approach (under the fixed effects framework) or by generalized least-squares (under the random effects framework[36]).

7.5.1. *Estimation Under the Fixed Effects Framework.* We use a multi-step procedure to estimate model (7.39) under the fixed effects framework. Given the numerous unobserved components in the model we require a four step estimation procedure to estimate both $\boldsymbol{\beta}$ and the measures of technical inefficiency.

Step 1: The standard within transformation can be performed on (7.39) to remove α_i before estimation. Since both the components of ω_{it} are zero mean and constant variance random variables, the within transformed ω_{it} will generate a random variable that has zero mean and constant variance. OLS can be used on the within transformed version of (7.39) to obtain consistent estimates of $\boldsymbol{\beta}$.

Step 2: Given the estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, from Step 1, we obtain pseudo residuals $r_{it} = y_{it} - \boldsymbol{x}'_{it}\hat{\boldsymbol{\beta}}$, which contain $\alpha_i^* + \omega_{it}$. Using these, we first estimate α_i^* from the mean of r_{it} for each i. Then, we can estimate u_i from $\max_i \hat{\alpha}_i - \hat{\alpha}_i^* = \max_i\{\bar{r}_i\} - \bar{r}_i$ where $\bar{r}_i$ is the mean (over time) of r_{it} for firm i. Note that the intercept, β_0, and ω_{it} are eliminated by taking the mean of r for a firm firm. The above formula gives an estimate of u_i relative to the best firm in the sample.

Step 3: With our estimates of $\boldsymbol{\beta}$ and u_i, we calculate residuals $e_{it} = y_{it} - \boldsymbol{x}'_{it}\hat{\boldsymbol{\beta}} + \hat{u_i}$, which contains $\beta_0 + v_{it} - \tau_{it}$. At this stage additional distributional assumptions are required to separate v_{it} from τ_{it}. Here we follow convention and assume $v_{it} \sim$ i.i.d. $N(0, \sigma_v^2)$ and $\tau_{it} \sim$ i.i.d. $N^+(0, \sigma_\tau^2)$. Maximum likelihood can be deployed here treating e_{it} as the dependent variable which is i.i.d.) to estimate β_0 and the parameters associated with v_{it} and τ_{it}. The log-likelihood for this setup is

$$\ln \mathcal{L} = -NT \ln \sigma + \sum_{i=1}^{N}\sum_{t=1}^{T} \ln \Phi(-e_{it}\lambda/\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2 \tag{7.40}$$

Step 4: Once $\hat{\beta}_0$ is removed from e_{it}, a JLMS conditional mean or median technique can be used to estimate τ_{it} for each observation.

To summarize estimation under the fixed effects framework, we estimate (7.39) using standard fixed-effects panel data tools to obtain consistent estimates of $\boldsymbol{\beta}$ in Step 1. Next in step 2, we estimate persistent technical inefficiency, u_i. In Step 3, we estimate β_0 and the parameters associated with the random components, v_{it} and τ_{it}. Finally, in Step 4, the time-varying (residual) component of inefficiency, τ_{it}, is estimated.

[36]A further transformation is needed to make α_i a zero mean random variable

Note that no distributional assumptions are used in the first two steps. Unfortunately, distributional assumptions are used to estimate residual inefficiency given that it cannot be separately identified without these.

7.5.2. *Estimation Under the Random-Effects Framework.* If we are willing to assume that u_i is uncorrelated with $\boldsymbol{x}_{it}$, then we can estimate model (7.39) under the random effects framework. To begin we rewrite (7.39) as

$$y_{it} = \beta_0 - E(u_i) - E(\tau_{it})\beta_0^* + \boldsymbol{x}_{it}'\boldsymbol{\beta} - u_i^* + v_{it} - (\tau_{it} - E(\tau_{it})) = \beta_0^* + \boldsymbol{x}_{it}'\boldsymbol{\beta} - u_i^* + \omega_{it}, \quad (7.41)$$

so that the error components, u_i^* and ω_{it}, are mean zero, homoscedastic random variables. Thus, the model in (7.41) fits into the standard one-way error component panel model. As under the fixed effects framework we can recover parameter estimates as well as time-varying and persistent inefficiency through a four step procedure.

Step 1: Deploy GLS to estimate $\boldsymbol{\beta}$ and β_0^*.

Step 2: Construct the pseudo residuals $\tilde{r}_{it} = y_{it} - \boldsymbol{x}_{it}'\hat{\boldsymbol{\beta}} - \beta_0^*$. The best linear unbiased predictor (BLUP) of u_i^* is

$$\tilde{u}_i^* = -\left\{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_\omega^2 + T\hat{\sigma}_u^2}\right\}\sum_{t=1}^{T}\tilde{r}_{it}, \quad (7.42)$$

where σ_u^2 and σ_ω^2 are the variances of u_i^* and ω_{it}. Once u_i^* is estimated, we can get an estimate of u_i from $u_i = \max_i\{u_i^*\} - u_i^*$.

Steps 3 and 4 are identical to those under the fixed effects framework. Again, distributional assumptions are required to separate v_{it} from τ_{it}.

7.6. **Models that Separate Firm Effects, Persistent Inefficiency and Time-varying Inefficiency.** The model introduced in (7.38) views all time constant effects as persistent inefficiency. Thus, all time constant variables are treated as persistent inefficiency even if it should be classified as time constant firm heterogeneity. Consequently, in the presence of unobserved firm heterogeneity, the model is misspecified and is likely to produce an upward bias in inefficiency (since all unobserved heterogeneity is viewed as inefficiency). Alternatively, the models introduced in Section 7.4 view firm effects as something other than inefficiency. Thus, these models fail to capture persistent inefficiency which is confounded within firm effects. Consequently, these models are also misspecified and tend to produce a downward bias in the estimate of overall inefficiency.

Given that the underlying assumptions pertaining to firm level inefficiency in the models introduced in Sections 7.4 and 7.5 are not fully satisfactory, we introduce the model of Kumbhakar, Lien & Hardaker (2014) and Colombi et al. (2014) that overcomes several of the limitations of these models. In this model the error term is split into four components to take into account different factors affecting output, given the inputs. The first component captures firms' latent heterogeneity Greene (2005*b*, 2005*a*), which has to be disentangled from inefficiency; the second component

captures short-run (time-varying) inefficiency. The third component captures persistent or time-invariant inefficiency as in Kumbhakar & Hjalmarsson (1993), Kumbhakar & Heshmati (1995), and Kumbhakar & Hjalmarsson (1998) while the last component captures random shocks.

This model is specified as:

$$y_{it} = \beta_0 + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \mu_i - \eta_i + v_{it} - u_{it} \tag{7.43}$$

The two components $\eta_i > 0$ and $u_{it} > 0$, reflect time constant and time varying inefficiency, respectively, while μ_i captures unobserved, time constant firm heterogeneity and v_{it} is the classical random shock. Previously, some of these components appeared in the models described, though not all simultaneously.

This new model improves upon the previous models in several ways. First, although some of the time-varying inefficiency models presented above can accommodate firm effects, these models fail to take into account the possible presence of some factors that might have permanent (i.e., time-invariant) effects on a firm's inefficiency. Here we call them permanent/time-invariant components of inefficiency. Second, SF models allowing time-varying inefficiency assume that a firm's inefficiency at time t is independent of its previous level inefficiency. It is more sensible to assume that a firm may eliminate part of its inefficiency by removing some of the short-run rigidities, while some other sources of inefficiency might stay with the firm over time. The former is captured by the time-varying component, η_i, and the latter by the time-varying component, u_{it}. Finally, many panel SF models do consider permanent/ time-invariant inefficiency effects, but do not take into account the effect of unobserved firm heterogeneity on output. By doing so, these models confound permanent/ time-invariant inefficiency with firm effects (heterogeneity). Models proposed by Greene (2005*b*, 2005*a*), Kumbhakar & Wang (2005), Wang & Ho (2010) and Chen et al. (2014) decompose the error term in the production function into three components: a producer-specific time-varying inefficiency term; a producer-specific random- or fixed effects capturing latent heterogeneity; and a producer- and time-specific random error term. However, these models consider any producer-specific, time-invariant component as unobserved heterogeneity. Thus, although firm heterogeneity is now accounted for, it comes at the cost of ignoring long-term (persistent) inefficiency. In other words, long-run inefficiency is again confounded with latent heterogeneity.

Many interesting panel SF models can be obtained as special cases of model (7.43) by eliminating one or more random components. For a somewhat easy reference to all these models we consider a three letter identifier system. Each identifier refers to an error component. Since every model contains a random shock component we do not put an identifier for it. Thus, although we have a maximum of four-way error components model a three letter model identifier is used. The first letter in the identifier pertains to the presence (T = True) or absence (F = False) of random firm (cross-sectional unit) effects in the SF model; the second letter (again, T or F) is related to the presence/absence of the time-varying (short-run) inefficiency term; and the third letter indicates the presence/absence of the time-invariant (persistent) inefficiency term. Using this system the four component model in (7.43) is labeled as TTT. Greene's true random-effect SF model (2005*b*, 2005*a*)

is obtained by dropping the η_i term from (7.43) and it is labeled as TTF. Similarly, the Kumbhakar & Heshmati (1995) model, which accommodates both short-run and long-run inefficiency terms but not latent firm heterogeneity, is labeled as FTT. The Pitt & Lee (1981), Schmidt & Sickles (1984), Kumbhakar (1987), and Battese & Coelli (1988) time-invariant inefficiency model is obtained by dropping the μ_i and u_{it} terms and is labeled as FFT. The pooled SF model (Pitt & Lee 1981) is labeled as FTF (i.e., b_i and η_i are dropped). Within this nomenclature, we could also have TFT, a model which accommodates latent firm heterogeneity and long-run (time-invariant) inefficiency, but not time-varying inefficiency (by omitting u_{it}).

Estimation of the model in (7.43) can be undertaken in a single stage ML method based on distributional assumptions on the four components (Colombi et al. 2011). Here, we first describe a simpler multi-step procedure (Kumbhakar, Lien & Hardaker 2014) before discussing full maximum likelihood estimation. For this, we rewrite the model in (7.43) as

$$y_{it} = \beta_0^* + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \tag{7.44}$$

where $\beta_0^* = \beta_0 - E(\eta_i) - E(u_{it})$; $\alpha_i = \mu_i - \eta_i + E(\eta_i)$; and $\varepsilon_{it} = v_{it} - u_{it} + E(u_{it})$. With this specification α_i and ε_{it} have zero mean and constant variance. This model can be estimated in three steps.

Step 1: Since (7.44) is the familiar panel data model, in the first step the standard random effect panel regression is used to estimate $\hat{\beta}$. This procedure also gives predicted values of α_i and ε_{it}, which we denote by $\hat{\alpha}_i$ and $\hat{\varepsilon}_{it}$.

Step 2: In the second step, the time-varying technical inefficiency, u_{it}, is estimated. For this we use the predicted values of ε_{it} from Step 1. Since

$$\varepsilon_{it} = v_{it} - u_{it} + E(u_{it}), \tag{7.45}$$

by assuming v_{it} is i.i.d. $N(0, \sigma_v^2)$ and u_{it} is $N^+(0, \sigma_u^2)$, which means $E(u_{it}) = \sqrt{2/\pi}\,\sigma_u$), and ignoring the difference between the true and predicted values of ε_{it} (which is the standard practice in any two- or multi-step procedure), we can estimate (7.45) using the standard SF technique. This procedure gives prediction of the time-varying residual technical inefficiency components, (i.e., Jondrow et al. 1982) or residual technical efficiency (i.e., Battese & Coelli 1988).

Step 3: In the final step we can estimate η_i following a similar procedure as in Step 2. For this we use the best linear predictor of α_i from Step 1. Since

$$\alpha_i = \mu_i - \eta_i + E(\eta_i), \tag{7.46}$$

by assuming μ_i is i.i.d. $N(0, \sigma_\mu^2)$, η_i is i.i.d. $N^+(0, \sigma_\eta^2)$, which in turn means $E(\eta_i) = \sqrt{2/\pi}\,\sigma_\eta$, we can estimate (7.46) using the standard normal-half normal SF model cross-sectionally and obtain estimates of the persistent technical inefficiency components, η_i, using the Jondrow et al. (1982) procedure. Persistent technical efficiency can then be estimated from PTE $= e^{-\eta_i}$, where $\hat{\eta}_i$ is the Jondrow et al. (1982) estimator of η_i. The

overall technical efficiency, OTE, is then obtained from the product of PTE and RTE, i.e., $OTE = PTE \times RTE$.

It is possible to extend this model (in steps 2 and 3) to include persistent and time-varying inefficiency that has non-zero mean as well as allowing for heteroscedasticity in both types of inefficiency. We now describe the estimation of the four component model via maximum likelihood estimation proposed in (Colombi et al. 2014).

7.6.1. *Full Maximum Likelihood Estimation.* While the multi-step approach of Kumbhakar, Lien & Hardaker (2014) discussed above allows one to control for latent firm effects and time varying and invariant inefficiency, it still imposes distributional assumptions in several steps and is, overall, inefficient relative to full maximum likelihood. However, given the structure of the four error component model we need to discuss how we can specify distributions for each component to ensure identification. For this we turn to the closed-skew normal distribution.

7.6.1.1. **The closed-skew normal distribution**

To obtain a tractable likelihood function, Colombi et al. (2014) use skew normal distributional assumptions for both the time variant and invariant random components of (7.43). The skew normal distribution, as alluded to in Section 2, is a more general distribution than the normal distribution, allowing for asymmetry (Azzalini 1985). Assuming v_{it} is *i.i.d.* standard normal and u_{it} is *i.i.d.* half normal, the composed error $v_{it} - u_{it}$ has a skew normal distribution. The same set of assumptions can be used for μ_i and η_i. Thus, model (7.43)'s likelihood can be derived.

To more precisely detail the log-likelihood function we describe the assumptions deployed in Colombi et al. (2014):

CKMV 1: For $i = 1, 2, ..., n$ and $t = 1, \cdots, T$, the $2(T+1)$ random variables $\eta_i,\ \mu_i,\ u_{it},\ v_{it}$ $(t = 1, 2, ..., T)$ are independent in probability;

CKMV 2: for every i, η_i is a normal random variable with expected value zero and variance σ^2_{1u} left-truncated at zero while μ_i is a normal random variable with zero mean and variance σ^2_{μ};

CKMV 3: for every i and t, u_{it} is a normal random variable with mean zero and variance σ^2_{2u} left-truncated at zero while v_{it} is a normal random variable with zero mean and variance σ^2_{v};

CKMV 4: $\mathbf{x}'_{it}$ is exogenous with respect to $\eta_i,\ \mu_i,\ u_{it},\ v_{it}$.

The following matrix representation of model (7.43) will be useful in our presentation of the log likelihood function. Let $\mathbf{1}_T$ be a vector of ones, $\mathbf{0}_T$ a vector of zeros; and $\boldsymbol{I}_T$ the identity matrix of dimension T. Moreover, $\boldsymbol{y}_i$ is a vector of the T observations on the *i-th* unit; $\boldsymbol{X}_i$ is the $T \times p$ matrix with rows $\boldsymbol{x}'_{it}$, $\boldsymbol{u}_i$ is the $(T+1)$ vector with components $\eta_i, u_{i1}, u_{i2}, ..., u_{iT}$; and $\boldsymbol{v}_i$ is the vector of the idiosyncratic random components of the *i-th* unit. From (7.43), it follows that: $\boldsymbol{y}_i = \mathbf{1}_T(\beta_0 + \mu_i) + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{u}_i + \boldsymbol{v}_i$, where the matrix $\boldsymbol{A}$ is defined as: $\boldsymbol{A} = -[\ \mathbf{1}_T \quad \boldsymbol{I}_T\ ]$.

Let $\phi_q(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Omega})$ be the density function of a q-dimensional normal random variable with expected value $\boldsymbol{\mu}$ and variance $\boldsymbol{\Omega}$, while $\overline{\Phi}_q(\boldsymbol{\mu}, \boldsymbol{\Omega})$ is the probability that a q-variate normal random variable of expected value $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Omega}$ belongs to the positive orthant.

A random vector $\boldsymbol{z}$, $-\infty < \boldsymbol{z} < \infty$, has an (o, q) closed-skew normal distribution with parameters $\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{D}, \boldsymbol{\nu}, \boldsymbol{\Delta}$ if its probability density function is (Arellano-Valle & Azzalini 2006, González-Farías, Domínguez-Molina & Gupta 2004):

$$f(\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{D}, \boldsymbol{\nu}, \boldsymbol{\Delta}, o, q) = \frac{\phi_o(\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Gamma}) \overline{\Phi}_q(\boldsymbol{D}(\boldsymbol{z} - \boldsymbol{\mu}) - \boldsymbol{\nu}, \boldsymbol{\Delta})}{\overline{\Phi}_q(-\boldsymbol{\nu}, \boldsymbol{\Delta} + \boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{D}')}. \tag{7.47}$$

Note that the dimensions of the matrices $\boldsymbol{\Gamma}, \boldsymbol{D}, \boldsymbol{\Delta}$ and of the vectors $\boldsymbol{\mu}, \boldsymbol{\nu}$ are determined by the dimensionality o of the multi-normal probability density function and by the dimensionality q of the multi-normal distribution function. Aside from the boldface and matrix notation, this is nothing more that the multivariate generalization of the univariate skew normal distribution we discussed in Section 2. The $\overline{\Phi}_q(-\boldsymbol{\nu}, \boldsymbol{\Delta} + \boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{D}')$ term appearing in the denominator of (7.47) is to ensure integration to 1 so that we have a theoretically consistent probability density function.

To keep the notational burden to a minimum we introduce the following matrices:

$$\begin{aligned}
\boldsymbol{V} &= \begin{bmatrix} \sigma_{1u}^2 & \mathbf{0}_T' \\ \mathbf{0}_T & \boldsymbol{\Psi} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \sigma_v^2 \boldsymbol{I}_T + \sigma_\mu^2 \mathbf{1}_T \mathbf{1}_T' \\
\boldsymbol{\Lambda} &= \boldsymbol{V} - \boldsymbol{V}\boldsymbol{A}'(\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{V} = (\boldsymbol{V}^{-1} + \boldsymbol{A}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}, \\
\boldsymbol{R} &= \boldsymbol{V}\boldsymbol{A}'(\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}')^{-1} = \boldsymbol{\Lambda}\boldsymbol{A}'\boldsymbol{\Sigma}^{-1},
\end{aligned}$$

where $\boldsymbol{\Psi}$ is the diagonal matrix with σ_{2u}^2 on the main diagonal.

The relevance of the closed-skew normal density function in the context of the TTT model stems from the following result of Colombi et al. (2014): Under assumptions CKMV1-CLMV 4, conditional on $\boldsymbol{X}_i$, the random vector $\boldsymbol{y}_i$ has a $(T, T+1)$ closed-skew normal distribution with the parameters: $\boldsymbol{\nu} = \mathbf{0}$, $\boldsymbol{\mu} = \mathbf{1}_T\beta_0 + \boldsymbol{X}_i\boldsymbol{\beta}$, $\boldsymbol{\Gamma} = \boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}'$, $\boldsymbol{D} = \boldsymbol{R}$; and $\boldsymbol{\Delta} = \boldsymbol{\Lambda}$. From this we have that the conditional on $\boldsymbol{X}_i$ density of $\boldsymbol{y}_i$ is

$$f(\boldsymbol{y}_i) = \phi_T(\boldsymbol{y}_i, \mathbf{1}_T\beta_0 + \boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}') \frac{\overline{\Phi}_{T+1}(\boldsymbol{R}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \mathbf{1}_T\beta_0), \boldsymbol{\Lambda})}{2^{-(T+1)}}. \tag{7.48}$$

It can be easily checked that it is not necessary to include both time invariant and time varying inefficiency to obtain a closed-skew normal distribution of the error components. For example a a (T, T) closed-skew normal results in both the TTF (Greene 2005*b*, Greene 2005*a*) and FTF (Kumbhakar 1987, Battese & Coelli 1988) models. Further still, when time-varying inefficiency is omitted, a $(T, 1)$ closed-skew normal density arises; and, when the random firm-effects are omitted, the joint distribution is given by the previous results with $\sigma_\mu^2 = 0$.

With the density in hand, the log-likelihood for the nT observations from (7.43) is:

$$\ln \mathcal{L} = \sum_{i=1}^{n} \left[\ln \phi_T(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}, \mathbf{1}_T\beta_0, \boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}') + \ln \overline{\Phi}_{T+1}(\boldsymbol{R}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \mathbf{1}_T\beta_0), \boldsymbol{\Lambda}) \right] \tag{7.49}$$

which is the nothing more than the sum of the log-likelihood for each of the n independent closed-skew normal random variables $\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}$. We mention here that for $T > 2$ the computational complexity involved to maximize the log likelihood function is high. This stems from the T integrals in $\overline{\Phi}_{T+1}(\boldsymbol{R}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \mathbf{1}_T\beta_0), \boldsymbol{\Lambda})$.

7.6.1.2. **Prediction of the random components**

Aside from estimating $\boldsymbol{\beta}$ and the parameters of the distributions of the random components, we still need to construct predictors of technical inefficiency and firm effects. To do this we introduce some more notation:

$$\boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \mathbf{1}_T\beta_0, \quad \tilde{\sigma}_\mu^2 = \sigma_\mu^2 - \sigma_\mu^4\mathbf{1}_T'\boldsymbol{\Delta}\mathbf{1}_T$$

$$\boldsymbol{\Delta} = (\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}')^{-1}, \tilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda} - \boldsymbol{R}\mathbf{1}_T\mathbf{1}_T'\boldsymbol{R}'\frac{\sigma_\mu^4}{\tilde{\sigma}_\mu^2}.$$

Colombi et al. (2014) list the conditional distributions (on $\boldsymbol{y}_i$) for μ_i and $\boldsymbol{u}_i$ as

$$f(\mu_i|\boldsymbol{y}_i) = \phi(\mu_i, \sigma_\mu^2\mathbf{1}'\boldsymbol{\Delta}\boldsymbol{r}_i, \tilde{\sigma}_\mu^2)\frac{\overline{\Phi}_{T+1}(\boldsymbol{R}\boldsymbol{r}_i - \boldsymbol{R}\mathbf{1}_T\sigma_\mu^2\tilde{\sigma}_\mu^{-2}(\mu_i - \sigma_\mu^2\mathbf{1}_T'\boldsymbol{\Delta}\boldsymbol{r}_i), \tilde{\boldsymbol{\Lambda}})}{\overline{\Phi}_{T+1}(\boldsymbol{R}\boldsymbol{r}_i, \boldsymbol{\Lambda})}; \tag{7.50}$$

$$f(\boldsymbol{u}_i|\boldsymbol{y}_i) = \frac{\phi_{T+1}(\boldsymbol{u}_i, \boldsymbol{R}\boldsymbol{r}_i, \boldsymbol{\Lambda})}{\overline{\Phi}_{T+1}(\boldsymbol{R}\boldsymbol{r}_i, \boldsymbol{\Lambda})}, \quad \boldsymbol{u}_i \geq 0. \tag{7.51}$$

These can in turn be used to derive the conditional moments of both the unobserved firm effects and time varying and time invariant technical inefficiency. This is done using the moment generating function of the (o, q) closed-skew normal distribution:

$$E(\exp\{\boldsymbol{t}'\boldsymbol{z}\}) = \frac{\overline{\Phi}_q(\boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{t} - \boldsymbol{\nu}, \boldsymbol{\Delta} + \boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{D}')}{\overline{\Phi}_q(-\boldsymbol{\nu}, \boldsymbol{\Delta} + \boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{D}')}\exp\{\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Gamma}\boldsymbol{t}\}. \tag{7.52}$$

Using the moment generating function, Colombi et al. (2014) provide the conditional means of the random effects as (in their model $\boldsymbol{y}$ is in logarithmic form):

$$E(e^{\mu_i}|\boldsymbol{y}_i) = \frac{\overline{\Phi}_{T+1}(\boldsymbol{R}\boldsymbol{r}_i - \boldsymbol{R}\mathbf{1}_T\sigma_\mu^2, \boldsymbol{\Lambda})}{\overline{\Phi}_{T+1}(\boldsymbol{R}\boldsymbol{r}_i, \boldsymbol{\Lambda})}e^{\sigma_\mu^2\mathbf{1}_T'\boldsymbol{\Delta}\boldsymbol{r}_i + \frac{1}{2}\tilde{\sigma}_\mu^2}; \tag{7.53}$$

$$E(e^{\boldsymbol{t}'\boldsymbol{u}_i}|\boldsymbol{y}_i) = \frac{\overline{\Phi}_{T+1}(\boldsymbol{R}\boldsymbol{r}_i + \boldsymbol{\Lambda}\boldsymbol{t}, \boldsymbol{\Lambda})}{\overline{\Phi}_{T+1}(\boldsymbol{R}\boldsymbol{r}_i, \boldsymbol{\Lambda})}e^{\boldsymbol{t}'\boldsymbol{R}\boldsymbol{r}_i + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Lambda}\boldsymbol{t}}. \tag{7.54}$$

The first element of (7.54) is the conditional expected value of time-invariant inefficiency for firm i. We also note that conditional on the observation $\boldsymbol{r}_i$, the firm effect μ_i does not have a normal distribution as is the case in the standard random effects panel model (Baltagi 2013).

7.6.2. *A Simulated Maximum Likelihood Estimator.* While the log-likelihood of the TTT stochastic frontier model appears daunting to implement, Greene & Fillipini (2014) recently proposed a simulation based optimization routine which circumvents many of the challenges with brute force

optimization in this setting. Using the insights of Butler & Moffitt (1982), Greene & Fillipini (2014) note that the density in (7.48) can be greatly simplified by conditioning on μ_i and η_i. In this case, the conditional density is simply the product over time of T univariate closed-skew normal densities. Thus, only a single integral, as opposed to T integrals needs to be calculated.

The conditional density using the Butler & Moffitt (1982) trick is

$$f(\boldsymbol{y}_i) = \int_{-\infty}^{\infty} \prod_{t=1}^{T} \left[\frac{2}{\sigma} \phi(\epsilon_{it}/\theta) \Phi(-\lambda \epsilon_{it}/\sigma) \right] \frac{2}{\theta} \phi(\delta_i/\theta) \Phi(-\gamma \delta_i/\theta) d\delta_i, \tag{7.55}$$

where $\epsilon_{it} = y_{it} - \alpha - \boldsymbol{x}_{it}'\boldsymbol{\beta} - \delta_i$ and $\delta_i = \mu_i - \eta_i$. Here we use the common $\lambda = \sigma_{2u}/\sigma_v$ and $\sigma = \sqrt{\sigma_v^2 + \sigma_{2u}^2}$ notation for the time varying skew normal density and $\theta = \sqrt{\sigma_\mu^2 + \sigma_{1u}^2}$ and $\gamma = \sigma_{1u}/\sigma_\mu$ for the time constant skew normal density. Following our discussion in Section 2, provided we can generate random draws for δ_i, we can replace the one dimension integral in (7.55) with a simple average.

Our simulated log likelihood function is

$$\ln \mathcal{L}_s = \sum_{i=1}^{N} \ln \left(R^{-1} \sum_{r=1}^{R} \prod_{t=1}^{T} \left[\frac{2}{\sigma} \phi(\tilde{\epsilon}_{it}/\theta) \Phi(-\lambda \tilde{\epsilon}_{it}/\sigma) \right] \right), \tag{7.56}$$

where $\tilde{\epsilon}_{it} = y_{it} - \beta_0 - \boldsymbol{x}_{it}'\boldsymbol{\beta} - \tilde{\delta}_{ir}$ and $\tilde{\delta}_{ir} = \sigma_\tau W_{ir} - \sigma_\eta |H_{ir}|$ and W_{ir} and H_{ir} are independent draws from a standard normal distribution. Maximization of this simulated log likelihood is not more complicated from the cross sectional case, aside from the additional parameters. With the milestone work of Colombi et al. (2014) and Greene & Fillipini (2014), estimation of fully general panel data stochastic frontiers is now quite accessible to applied researchers.

7.6.3. *Inference Across the TTT Model.* The most general stochastic frontier model in the panel context is the TTT. Models appearing in the literature are special cases of TTT. For example, the widely used true random effects model, TTF, (Greene 2005*b*) is a special case of the TTT model. The same holds for all of the models we have discussed above. Naturally, inference is necessary to determine the best model for the data at hand.

Testing any of the previous models against the most general TTT model is a non-standard problem because under the null hypothesis one or more parameters are on the boundary of the parameter space. In fact, the asymptotic distribution of the log-likelihood ratio test statistic, under reasonable assumptions, is a mixture of chi-square distributions known as the chi-bar-square distribution (Silvapulle & Sen 2005). For example, to test the FTT model against the TTT model the log-likelihood ratio test statistic is asymptotically distributed as a 0.5 mixture of a chi-square distribution with zero degrees of freedom and a chi-square distribution with one degree of freedom and the p-value is found by dividing the p-value corresponding to a chi-square distribution with one degree of freedom by two.

Future research focusing on adapting testing procedures to the TTT framework is important moving forward. As discussed earlier, the presence of both time varying and invariant inefficiency

yields different policy recommendations and so working with models that document their presence, or lack of one, are important for proper analysis.

8. Nonparametric Estimation in the Stochastic Frontier Model

An oft heard criticism of the benchmark, cross-section stochastic frontier model is that it is heavily parameterized, both regarding the specification of the production function as well as the distributional assumptions placed on the components of the convoluted error term. If any of these assumptions break down it is likely that model misspecification will impact estimates and any inferences based off of them. This is widely perceived as one of the major disadvantages of stochastic frontier methods relative to their deterministic counterparts (Simar & Wilson 2013), which are almost exclusively estimated in a nonparametric fashion. To remedy these specification issues, a variety of approaches have been proposed to combat the presence (to various degrees) of model misspecification. Naturally, a first step in this direction is to lessen parametric specification on the production frontier itself as it is widely regarded that the specification of the error distribution has less impact on the analysis.

If interest hinges directly on the shape of the inefficiency distribution then semiparametric deconvolution methods (Horrace & Parmeter 2011) can be deployed to recover this distribution and to construct conditional mean estimates of inefficiency. Further, flexible methods that allow the parameters of the error distribution to depend in an unspecified fashion on the covariates while simultaneously allowing the production frontier to be left unspecified exist. We can think of a hierarchical treatment of existing approaches which focus attention on relaxing assumptions regarding estimation of the frontier, relaxing assumptions on the shape/features of the distribution of inefficiency, or combined approaches. Given the presence of noise, it is currently impossible to have a fully nonparametric approach which can leave all features of the model unspecified and yet recover estimates of inefficiency and the production frontier (Hall & Simar 2002). Thus, many have referred to the existing set of methods as semiparametric since some parametric assumptions/restrictions are required at various stages of the analysis.

In our view the use of nonparametric estimation techniques within the confines of the stochastic frontier model is one of the main areas of stochastic frontier analysis where the largest gains will come from. And while several nonparametric methods have been available to researchers since the early 1990s, their use in applied efficiency settings has only recently risen. Given that many of the existing surveys, books and authoritative reviews of the stochastic frontier model do not contain discussion of the myriad advances using nonparametric methods, our approach here is to detail many of these recent advances and hopefully showcase their applied appeal. We also provide an overview of kernel smoothing methods, which is one of the most popular (but certainly not the only) techniques for relaxing parametric assumptions, which may be new to readers who are well schooled in parametric stochastic frontier methods.

8.1. **A Primer on Kernel Smoothing.**

While a variety of tools exist to implement nonparametric estimators, undoubtedly the most popular are kernel based methods. Kernel methods entail estimation based on taking local averages. Here local refers to the distance between the point of

interest, $\boldsymbol{x}$ and nearby observations, $\boldsymbol{x}_i$, and is measured by a smoothing parameter, more commonly referred to as a bandwidth. Think of constructing the sample mean for a given set of data, $\bar{x} = N^{-1} \sum_{i=1}^{N} x_i$. Each observation is weighted *equally* with respect to the other observations. Nonparametric methods seek to adjust the uniform weighting (the N^{-1} in front of the summand) to more adequately characterize the underlying structure of the data generating process. More generally, nonparametric estimators employ local weighting; thus, for the density of $\boldsymbol{x}$, $f(\boldsymbol{x})$, weighting is changed as $\boldsymbol{x}$ changes, thereby affording flexibility to the method not present in traditional parametric methods. Given the myriad reviews of kernel smoothing our review will be kept to a minimum, focusing on terminology and key concepts. For an advanced treatment see the authoritative treatise of Li & Racine (2007). For more heuristic reviews we recommend Racine (2008) and Henderson & Parmeter (2014).

8.1.1. *Density Estimation.* To obtain a firm grasp of how general nonparametric estimators work it is instructive to consider the kernel density estimator. This estimator provides the necessary intuition for understanding the mechanics of kernel smoothing. First let us consider a basic approach to approximate a density which will flow into the kernel density estimator; the histogram. Specifically, for univariate x, if we assume that x has support $\mathcal{S}(x) = [a, b]$, we can divide $\mathcal{S}(x)$ into B equally spaced boxes where each box will have width $h = (b - a)/B$. We will refer to h as the binwidth of any given box. The intervals can be described as

$$(a + (r - 1)h, a + rh], \qquad \text{for } r = 1, 2, \ldots, B.$$

Let N_r denote the number of observations from the sample that are located in box r. This can be summarized as

$$N_r = \sum_{i=1}^{N} \mathbf{1}\{a + (r - 1)h < x_i \leq a + rh\}.$$

The proportion of observations falling into the r^{th} box is N_r/N. The expectation of the proportion of observations in the r^{th} interval is then

$$E[N_r]/N = Pr\left(a + (r - 1)h < X \leq a + rh\right) = \int_{a+(r-1)h}^{a+rh} f(x)dx, \tag{8.1}$$

where $f(x)$ is the density of x, the object one wishes to estimate.

Now, assuming B is large so that h is small, then on the interval $(a + (r - 1)h, a + rh]$, $f(x)$ is approximately constant, $f(x) \approx f(c)$ for some $c \in (a + (r - 1)h, a + rh]$. Thus,

$$E[N_r]/N = \int_{a+(r-1)h}^{a+rh} f(x)dx \approx f(c)\left(a + rh - (a + (r - 1)h)\right) = hf(c). \tag{8.2}$$

Dividing each side of (8.2) by h suggests that a crude estimator of a density function is

$$\widehat{f}(c) = N_r/(Nh) \tag{8.3}$$

for $c \in (a + (r-1)h, a + rh]$.

Calculation of the density in this fashion fails to account for where the data are located. Thus, since the density estimator is constant in any particular box, certain points will be estimated more precisely in a box than others. A better estimator can be constructed by using the location of the observations to help place the boxes throughout $\mathcal{S}(x)$. This estimator is known as the centered histogram.

If the point of interest is x, then using boxes that are centered over the sample observations, the number of observations that are within $2h$ (the centered binwidth) of x is

$$N_x = \sum_{i=1}^{N} \mathbf{1}\left\{x - h < x_i \leq x + h\right\}.$$

The corresponding probability of falling in this box (centered on x) is N_x/N. A natural estimator of our density is then

$$\widehat{f}(x) = \frac{N_x}{2Nh} = \frac{1}{2Nh} \sum_{i=1}^{N} \mathbf{1}\left\{x - h < x_i \leq x + h\right\},$$

and for convenience, the density estimator is written as

$$\widehat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} (1/2)\mathbf{1}\left\{\left|\frac{x_i - x}{h}\right| \leq 1\right\}. \tag{8.4}$$

For the discussion that follows, we are going to generalize this estimator slightly. First, replace $(1/2)\mathbf{1}\left\{\left|\frac{x_i - x}{h}\right| \leq 1\right\}$ in (8.4) with a 'kernel' function $k(u)$ where

$$k(u) = \begin{cases} 1/2 & \text{if } |u| \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{8.5}$$

Note that $k(u)$ is used for notational convenience. In general, u is taken to be $(x_i - x)/h$. One issue with the kernel described here is that it is discontinuous at -1 and 1 and has a derivative of 0 everywhere except at these two points (where it is undefined). This suggests that the corresponding density estimate is going to be non-smooth. While there exist many non-smooth densities, typically one envisions the underlying density being estimated as smooth, which is certainly the case for the error distributions we have considered in our previous discussion.

With this specific kernel density estimator, notice a general form for the kernel density estimator

$$\widehat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} k\left(\frac{x_i - x}{h}\right). \tag{8.6}$$

Given the generality of this definition regarding what the kernel function, $k(u)$, may look like, h is typically referred to as a bandwidth as opposed to a binwidth. In practice one will want to select

kernel functions that are much smoother than the uniform kernel. The bandwidth corresponds to the smoothness of the estimate now instead of the direct width that the kernel covers.

Several generic properties on the kernel function involve the moments of the kernel. Letting

$$\kappa_j(k) = \int_{-\infty}^{\infty} u^j k(u) du, \tag{8.7}$$

a kernel is of 2^{nd} order if $\kappa_0(k) = 1$, $\kappa_1(k) = 0$ and $\kappa_2(k) < \infty$. Imposing $\kappa_0(k) = 1$ means that any weighting function must integrate to unity. This is why probability density functions unanimously appear as kernels in applied work. Any symmetric kernel will satisfy $\kappa_1(k) = 0$. Now, a natural question to ask is how this discussion extends if we have more than a single covariate. While there are many options for constructing kernels to smooth multivariate data, we advocate for the product kernel, which is simply the product of the individual kernel functions, applied to each component of our data.

The product kernel function can be constructed as

$$K_{i\boldsymbol{x}} = K_h\left(\boldsymbol{x}_i, \boldsymbol{x}\right) = \prod_{d=1}^{q} k\left(\frac{x_{id} - x_d}{h_d}\right).$$

$k(\cdot)$ is the individual kernel function which smooths a specific covariate. One of the most popular kernel functions is the Gaussian kernel, $k(z) = \phi(z)$. This gives rise to the multivariate kernel density estimator

$$\widehat{f}(\boldsymbol{x}) = \frac{1}{N\left|\mathbf{h}\right|} \sum_{i=1}^{N} K_{i\boldsymbol{x}}, \tag{8.8}$$

where $|\mathbf{h}|$ is the product of the individual bandwidths, $h_1 h_2 \cdots h_q$. This notation and setup may seem somewhat foreign to the untrained reader but what we see in (8.8) is nothing more than a weighted average where the kernel allows each point to provide a different amount of weighing in the construction of the average for a given point based on how close the observations are to this point, rather than treat all of the data equally.

One aspect of nonparametric estimation that practitioners must confront is that unlike parametric modeling, nonparametric modeling typically entails a bias in finite samples. Now, to be clear, parametric methods do not impose a bias when they are correctly specified. So for example, if we assume the production frontier is Cobb-Douglas, and it actually is, then the parametric estimates from maximum likelihood are going to be unbiased. However, the issue with parametric methods is that if the functional form is misspecified then even as the sample gets larger, the bias does not disappear. Alternatively, for nonparametric methods, even though they are biased estimators for any given sample size, as the sample size grows the bias diminishes. This is the tradeoff that the applied research will confront when deciding how best to construct the model and estimate it.

The bias of the kernel density estimator is Li & Racine (2003, Theorem 3.1)

$$Bias\left[\widehat{f}(\boldsymbol{x})\right] \approx \frac{\kappa_2(k)}{2}\sum_{d=1}^{q} h_d^2 \frac{\partial^2 f(\boldsymbol{x})}{\partial x_j^2}.$$

The key feature of the bias that we draw attention to is the h_d^2 term in the summation. Notice that the bias of our nonparametric density estimator will decrease as these bandwidths shrink to zero. This is the first point that we will make about the bandwidth, as the bandwidth gets smaller, the bias of the estimator will also get smaller. On the surface this seems great, you can simply set they bandwidth to zero and obtain a nonparametric estimator with no bias.

This is not quite right however. The reason is that in all of statistical analysis there is always a tradeoff between bias and variance. Here is no different. If we consider the variance of the kernel density estimator, which is

$$Var\left(\widehat{f}(\boldsymbol{x})\right) \approx \frac{f(\boldsymbol{x})}{N|\mathbf{h}|}$$

we see that if we were to deploy the strategy of making our bandwidths exceptionally small, that while we would eliminate the bias of our estimator, we would also have an estimator that is not useful for inference, given that the variance would explode.

This tradeoff raises the natural question of what we should do regarding the size of the bandwidth. With biased estimators it is common to examine the mean squared error of the estimator, the sum of the squared bias and the variance. In our current setting this is

$$MSE\left(\widehat{f}(\boldsymbol{x})\right) \approx \frac{\kappa_2^2(k)}{4}\left(\sum_{d=1}^{q} h_d^2 \frac{\partial^2 f(\boldsymbol{x})}{\partial x_j^2}\right)^2 + \frac{f(\boldsymbol{x})}{N|\mathbf{h}|}. \tag{8.9}$$

This formula is typically minimized with respect to the bandwidths to determine the set of bandwidths which provide the right balance between the bias and the variance of the estimator. Implementing this minimization in practice can be challenging given that many of the quantities in (8.9) depend on the exact thing the estimator is trying to estimate, in this case the unknown density of $\boldsymbol{x}$, $f(\boldsymbol{x})$. Ignoring this issue, a key feature of the bandwidths that will be obtained from this optimization is that $h_{opt} \sim N^{-1/(4+q)}$. That is, the optimal rate of decay for the bandwidths is $1/(4+q)$. The dependence of this rate of decay on the number of variables in the model is commonly referred to as the curse of dimensionality and is an issue to be aware of in empirical settings. While parametric misspecification is a thorny issue, the curse of dimensionality can also prevent the applied researcher from obtaining meaningful estimates in a nonparametric setting depending upon on how large N is.

8.1.2. *Regression Estimation.* In a stochastic frontier analysis our interest will hinge on the frontier function, which for our purposes here we will think of as a regression function. Ignoring for the moment the structure of the error term our nonparametric regression model, as given in Racine &

Li (2004), is

$$y_i = m(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, 2, \ldots, N. \tag{8.10}$$

Using the product kernel described above, it is straightforward to construct kernel regression estimators. The local constant least squares (LCLS) estimator is defined as

$$\widehat{m}(\boldsymbol{x}) = \left(\sum_{i=1}^{N} y_i K_{i\boldsymbol{x}} \right) / \left(\sum_{i=1}^{N} K_{i\boldsymbol{x}} \right). \tag{8.11}$$

The intuition behind the construction of this estimator follows from a simple example. If we were estimating the expected output for a single farmer who uses 2.5 kilograms of fertilizer per hectare, we could simply use the observed output for that given farmer. However, in this case this is likely to be a quite noisy estimate of the farmer's output given that it is based on a single observation. If instead we averaged the observed output for all farms who used two to three kilograms of fertilizer per hectare we could mitigate some of the noise stemming from this single observation. Clearly the selection of farmers who use two to three kilograms is arbitrary, but the bandwidth is used here to help us eliminate this arbitrariness. Naturally though we can see that for farmers using 20 kilograms of fertilizer they are quite different from this farmer using 2.5 kilograms and so this farmer should receive less weight.

Heuristically, what the LCLS estimator amounts to is running the regression

$$y_i = a + v_i; \quad i \in N_{close}(\boldsymbol{x}) \tag{8.12}$$

where $N_{close}(\boldsymbol{x})$ is the set of observations that is deemed to be close to the point of interest, $\boldsymbol{x}$. This is nothing more than OLS estimation with a constant for a subset of our data. Whereas typical OLS, say what the analyst would run if they elected to use MOLS, involves a single regression, LCLS involves N regressions, one for each observation, $\boldsymbol{x}_i$. Or, in matrix form, if we were to denote the column of N ones as $\boldsymbol{\imath}$, then the OLS estimator of $\boldsymbol{y}$ on $\boldsymbol{\imath}$ is simply

$$\bar{y} = (\boldsymbol{\imath}'\boldsymbol{\imath})^{-1}\boldsymbol{\imath}'\boldsymbol{y}, \tag{8.13}$$

whereas the LCLS estimator is

$$\bar{y}(\boldsymbol{x}) = (\boldsymbol{\imath}'\mathcal{K}(\boldsymbol{x})\boldsymbol{\imath})^{-1}\boldsymbol{\imath}'\mathcal{K}(\boldsymbol{x})\boldsymbol{y}, \tag{8.14}$$

where $\mathcal{K}(\boldsymbol{x})$ is the square matrix with the elements $K_{i\boldsymbol{x}}$ along the diagonal.

The theoretical properties of the LCLS estimator are well known (see for example Racine & Li 2004) and they nearly mimic those discussed briefly for the density estimator. The LCLS estimator has a bias which can be shrunk by making the bandwidth smaller, at the expense of increasing the variance. Moreover, the curse of dimensionality presents itself for the regression setting in equal force as in the density case, thus, care must be taken with datasets that have a large number of continuous covariates.

8.1.2.1. **Local-linear least-squares**

As an alternative to the LCLS estimator, which approximates the unknown conditional mean with a constant, the local linear least squares (LLLS) estimator, using the same intuition as the LCLS estimator, estimates the underlying conditional mean locally, but by fitting a line instead of a constant. To see how this works, we take a Taylor expansion around the point of interest, $\boldsymbol{x}$, in (8.10) as

$$y_i \approx m(\boldsymbol{x}) + (\boldsymbol{x}_i - \boldsymbol{x})' \boldsymbol{\beta}(\boldsymbol{x}) + \varepsilon_i, \tag{8.15}$$

where $\boldsymbol{\beta}(\boldsymbol{x}) \equiv \partial m(\boldsymbol{x}) / \partial \boldsymbol{x}$. We can rewrite Equation (8.15) as

$$y_i = \boldsymbol{X}' \boldsymbol{\delta}(\boldsymbol{x}) + \varepsilon_i, \tag{8.16}$$

where the ith row of $\boldsymbol{X}$ is $[1, (\boldsymbol{x}_i - \boldsymbol{x})]$ and $\boldsymbol{\delta}(\boldsymbol{x}) \equiv [m(\boldsymbol{x}), \beta(\boldsymbol{x})]'$.

It can be shown (Racine & Li 2004) that the LLLS estimator is

$$\widehat{\boldsymbol{\delta}}(\boldsymbol{x}) = \begin{pmatrix} \widehat{m}(\mathbf{x}) \\ \widehat{\boldsymbol{\beta}}(\boldsymbol{x}) \end{pmatrix} = \left(\boldsymbol{X}' \mathcal{K}(\boldsymbol{x}) \boldsymbol{X}\right)^{-1} \boldsymbol{X}' \mathcal{K}(\boldsymbol{x}) \boldsymbol{y}. \tag{8.17}$$

An additional benefit of the LLLS estimator in practice is that it will estimate nonparametrically both the conditional mean and the vector of first derivatives of the conditional mean. These gradients are usually of interest to practitioners, for example, when studying banks, one may wish to calculate returns to scale, which is simply a function of the first derivatives. Further, we could extend the Taylor expansion to second or third order (or even higher) and estimate this model locally. In doing so we have what is termed the local polynomial least squares (LPLS) estimator.

8.1.3. *Bandwidth Selection.* As in the density setting, selection of the bandwidth parameters is paramount for practical implementation of the estimator. The most common approach is to use least squares cross validation (LSCV). LSCV proceeds by selecting the vector of bandwidths that minimizes the squared prediction errors from the LCLS (or LLLS) estimator which omits the ith observation when constructing nonparametric estimates for the ith observation. That is, the criterion function of LSCV is

$$CV(h) = \sum_{i=1}^{N} \left[y_i - \widehat{m}_{-i}(\boldsymbol{x}_i)\right]^2,$$

where $\widehat{m}_{-i}(\boldsymbol{x}_i)$ is the leave-one-out LCLS or LLLS estimator. The leave-one-out LCLS estimator is defined as

$$\widehat{m}_{-i}(\boldsymbol{x}_i) = \left(\sum_{\substack{j=1 \\ j \neq i}}^{N} y_j K_{ji}\right) / \left(\sum_{\substack{j=1 \\ j \neq i}}^{N} K_{ji}\right).$$

Here $K_{ji} = K_h(\boldsymbol{x}_j, \boldsymbol{x}_i)$. The LSCV approach has been throughly studied and scrutinized and is generally recognized as a practical approach (but surely not the only approach) to selecting smoothing parameters in practice.

While the above discussion may seem to suggest that the results of our stochastic frontier analysis will be too dependent on arbitrary choices about the degree of smoothness, we argue that there is no more arbitrariness in this process of statistical estimation than there is in traditional parametric stochastic frontier analysis. The choice of specification for the production function, coupled with the distributional assumptions on both u and v is typically as *ad hoc* as picking the smoothing parameter. This is because production theory rarely provides guidance on these unknown quantities. It is just as likely that key insights into these important economic relationships are driven via the choice of parametric density assumed for inefficiency or the specification of the conditional mean of the production structure.

8.2. **Estimation of the Frontier.**

8.2.1. *Semiparametric Approaches.* Nonparametric estimation of the stochastic production frontier was first proposed by Banker & Maindiratta (1992) and Fan, Li & Weersink (1996). Banker & Maindiratta (1992) propose a nonparametric approach that is quite similar to data envelopment analysis but is embedded in a maximum likelihood framework to allow for both noise and inefficiency whereas Fan et al. (1996) use standard kernel methods coupled with maximum likelihood. Fan et al. (1996) note that direct nonparametric estimation of the conditional mean would result in a biased estimate when one ignores the inefficiency term. That is, the key condition required for consistent estimation of the production frontier in a regression setting, ignoring the composed error structure, is $E[\varepsilon|\boldsymbol{x}] = 0$. However, given the one-sided nature of u, this condition cannot be satisfied for any $\boldsymbol{x}$. Thus, the level of the production frontier cannot be identified in the regression setup as

$$y_i = m(\boldsymbol{x}_i) + \varepsilon_i = m(\boldsymbol{x}_i) + \mu + (\varepsilon_i - \mu) \equiv m^*(\boldsymbol{x}_i) + \varepsilon_i^*. \tag{8.18}$$

Note that this is true whether we wish to estimate the frontier in a parametric or nonparametric fashion.

Fan et al.'s (1996) solution is to correct the (downward) bias in the estimation of $m(\boldsymbol{x})$ by retaining standard distributional assumptions from the SFA literature (e.g., normal noise, half-normal inefficiency) and estimating the corresponding distributional parameters via maximum likelihood on the nonparametric residuals from a standard kernel regression. Once these parameters are determined, the estimated conditional mean can be shifted (bias-corrected) by the estimated mean of the inefficiency distribution (mean correction factor). Under weak conditions Fan et al. (1996) show that the parameters of the composed error distribution can be estimated at the parametric $\sqrt{n}$ rate. Their simulations reveal that the semiparametric method produces estimates of the distributional parameters that are competitive with the same distributional parameter estimates produced from correctly specified production frontiers in the standard maximum likelihood framework.

Fan et al. (1996) assume that noise follows a normal distribution (as per usual) and that technical inefficiency stems from a half-normal distribution. Given these distributional assumptions and the (biased) nonparametric estimate of the frontier, one would estimate the stochastic semiparametric frontier model and the unknown distributional parameters as:

Step 1: Compute the conditional expectation, $E[y_i|\boldsymbol{x}_i]$, using either the local constant or local linear estimator. Call this $\hat{m}(\boldsymbol{x}_i)$. Let the residuals be denoted $\hat{\varepsilon}_i = y_i - \hat{m}(\boldsymbol{x}_i)$.

Step 2: Define the concentrated variance of the composed error term $\sigma^2(\lambda)$ as a function of $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_u^2 + \sigma_v^2$, as follows:

$$\hat{\sigma}^2(\lambda) = \frac{N^{-1}\sum_{i=1}^{N}\hat{\varepsilon}_i^2}{1 - \frac{2\lambda^2}{\pi(1+\lambda^2)}}. \tag{8.19}$$

Step 3: Define the mean correction factor $\mu = \sqrt{2/\pi}\sigma_u$ as a function of λ, i.e.,

$$\hat{\mu}(\lambda) = \frac{\sqrt{2}\hat{\sigma}(\lambda)\lambda}{(\pi(1+\lambda^2))^{1/2}}. \tag{8.20}$$

Step 4: Estimate λ by maximizing the concentrated log likelihood function consistent with the normal, half normal distributional assumptions which is

$$\hat{\lambda} = \max_{\lambda}\left(-N\ln\hat{\sigma}(\lambda) + \sum_{i=1}^{N}\ln(\Phi(-\tilde{\varepsilon}_i\lambda/\hat{\sigma}(\lambda)) - (2\hat{\sigma}^2(\lambda))^{-1}\sum_{i=1}^{N}\tilde{\varepsilon}_i^2\right), \tag{8.21}$$

where $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \hat{\mu}(\lambda)$.

Step 5: The stochastic production frontier $m(\boldsymbol{x}_i)$ is consistently estimated by

$$\tilde{m}(\boldsymbol{x}_i) = \hat{m}(\boldsymbol{x}_i) + \hat{\mu}, \tag{8.22}$$

where $\hat{\mu} = \sqrt{2}\hat{\sigma}\hat{\lambda}/\left(\pi(1+\hat{\lambda}^2)\right)$ and $\hat{\sigma} = \sqrt{\hat{\sigma}^2(\hat{\lambda})}$.

The concentration of the regular normal, half normal likelihood function is for simplicity. One could just as easily maximize the traditional maximum likelihood function. However, in this case the user now has to worry about only one dimensional optimization which is now commonly thought of as a trivial problem given current computing power.

Note that this style of estimation would work with a parametric functional form. One would simply engage in OLS estimation of the unknown frontier and then bias correct the estimates upwards, which amounts to an intercept correction similar to COLS or MOLS.[37]

Recently, Martins-Filho & Yao (2011) demonstrated that while the kernel estimator of the frontier detailed in Fan et al. (1996) is consistent, the parametric estimator for the parameters of the

[37]The approach of Kuosmanen & Kortelainen (2012) and Parmeter & Racine (2012) is essentially identical to this setup except that they require the production frontier to obey traditional axioms such as monotonicity and concavity something that Fan et al. (1996) did not accommodate in their approach.

convoluted density produce an asymptotic bias when normalized by $\sqrt{n}$. As an alternative, Martins-Filho & Yao (2011) proposed an estimator which jointly estimates the distributional parameters and the unknown frontier.

To be more precise, consider again the likelihood function proposed by Fan et al. (1996):

$$\mathcal{L}_N(\theta, m) = N^{-1} \sum_{i=1}^{N} \ln f_\varepsilon \left(y_i - m(\boldsymbol{x}_i; \theta, g) - \gamma(\theta); \theta \right). \tag{8.23}$$

Fan et al.'s (1996) estimator of θ is based on the idea that if $m(\boldsymbol{x}_i; \theta, g)$ were known, a standard parametric ML estimator for θ could be obtained in routine fashion by maximizing $\bar{\mathcal{L}}(\theta, m)$. However, since $m(\cdot)$ is unknown, the exact likelihood function is replaced with the approximation

$$\tilde{\mathcal{L}}_N(\theta, \widehat{m}) = N^{-1} \sum_{i=1}^{N} \ln f_\varepsilon \left(y_i - \widehat{m}(\boldsymbol{x}_i) - \gamma(\theta); \theta \right) \tag{8.24}$$

and the parametric estimator is $\widehat{\theta} = \max_{\theta} \tilde{\mathcal{L}}_N \left(\theta, \widehat{m} + \gamma(\theta) \right)$. The bias associated with estimating the frontier in a first stage independent of the structure of f_ε is what produces an asymptotic bias in the ML estimates. There are two remedies to handle the bias in $\widehat{\theta}$. First, one can use a suboptimal bandwidth when estimating $m(\boldsymbol{x}_i)$; this suboptimal bandwidth will allow the bias in $\widehat{m}$ to decay at an appropriate rate such that $\widehat{\theta}$ no longer carries an asymptotic bias when normalized by $\sqrt{n}$. Second, a joint estimation procedure that explicitly connects estimation of θ and $m(\boldsymbol{x}_i)$ can be deployed; this procedure is termed profile likelihood estimation. This is precisely the approach of Martins-Filho & Yao (2011).

Martins-Filho & Yao's (2011) joint estimation approach relies on local likelihood estimation. In much the same way that kernel regression constructs local averages to estimate the conditional mean, local likelihood local averages the likelihood surface to construct a smoothed likelihood function. In the current setup we have

$$\check{\mathcal{L}}_N(\theta, m_{\boldsymbol{x}}) = (N|h|)^{-1} \sum_{i=1}^{N} \ln f_\varepsilon(y_i - m(\boldsymbol{x}_i); \theta) K_{i\boldsymbol{x}}. \tag{8.25}$$

Notice that this likelihood function looks almost like the parametric likelihood, except, as per usual with kernel methods, we allow each observation to be weighted differently, as opposed to the uniform N^{-1} weighting that appeared in our earlier parametric discussion.

Estimation involves two repeated steps. First, for a fixed $\boldsymbol{x}$ and θ, $m(\boldsymbol{x}_i)$ is estimated as

$$\widehat{m}(\boldsymbol{x}) = \max_{m(\boldsymbol{x})} \check{\mathcal{L}}_N(\theta, m).$$

Second, using this estimator of $m(\boldsymbol{x})$, an estimator of θ is determined as

$$\widehat{\theta} = \max_{\theta} \check{\mathcal{L}}_N(\theta, \widehat{m}_{\boldsymbol{x}}).$$

This two-step procedure requires iteration to obtain the final estimates. As starting values the estimated parameters from simple application of Fan et al. (1996). From here, the two-step

approach is implemented for each point of interest; keep in mind that there if you evaluate the frontier at 75 points, then 76 optimizations need to be solved, one for each evaluation point and a final optimization to determine θ. The iterations can be terminated once there is sufficiently small movement in $\widehat{\theta}$ from iteration to iteration. Martins-Filho & Yao (2011) use a tolerance of 0.001 in their Monte Carlo simulations. As a last step, the final estimate of θ is then used to reestimate $m(\boldsymbol{x})$ for each of the evaluation points.

8.2.2. *Alternative Methods.* While Fan et al. (1996) is commonly thought of as the first attempt to lessen parametric assumptions in the stochastic frontier model, Banker & Maindiratta (1992) is often overlooked in this regard. They proposed a nonparametric estimator of the production frontier, embedded in a normal, truncated normal likelihood setting (though they did not implement this estimator). This model can actually be thought of as a non-smooth, single stage equivalent to the method of Martins-Filho & Yao (2011). Banker & Maindiratta (1992) suggest estimating the frontier using piecewise linear segments subject to monotonicity and concavity constraints (much the same way that the DEA estimator is constructed). These constraints ensure that the estimated frontier obeys the crucial axioms of production. An unfortunate aspect of Banker & Maindiratta's (1992) model was, at the time, the inability to reliably implement this estimator. Another, still extant, issue with their model is the nonsmoothness of the resultant estimator. It is not clear how important aspects of the frontier, returns to scale for example, are calculated.

Recently, a number of estimators have been proposed that build upon the work of Banker & Maindiratta (1992) in various dimensions. Parmeter & Racine (2012) propose imposing monotonicity and convexity constraints within the confines of the Fan et al. (1996) estimator and, consider a variant which simply deploys either COLS or MOLS rather than make distributional assumptions. Kuosmanen & Kortelainen (2012) use a similar piecewise linear framework as Banker & Maindiratta (1992) to impose monotonicity and concavity but rely on minimizing a sum of squared errors criterion instead of maximizing a likelihood function, which they refer to as stochastic nonparametric envelopment of data (StoNED). The distributional parameters are either recovered using MOLS or with a similar approach as in Fan et al. (1996). Another nonsmooth approach which has ties back to Banker & Maindiratta (1992) is the work of Kuosmanen (2008) and Kuosmanen & Johnson (2010) which embeds DEA in a standard regression setting via convex nonparametric least squares. This approach has the benefit of not requiring *a priori* distributional assumptions on the error term to estimate the production frontier.

It remains to be seen the empirical benefit of directly imposing monotonicity and concavity constraints on the production frontier. In both Parmeter & Racine (2012) and Kuosmanen & Kortelainen (2012) these constraints appear to have an impact on what one learns, however, the empirical work of Parmeter, Sun, Henderson & Kumbhakar (2014) did not find large differences on estimates of the technology when imposing these constraints. More work is clearly needed to determine the usefulness of imposing these constraints in practice.

Another, quite general approach, is that of Simar, Van Keilegom & Zelenyuk (2014) who propose local polynomial estimation of the frontier, including determinants of inefficiency. Their approach is similar to Fan et al. (1996) except that they do not recover estimates of the distributional parameters in a similar second stage, as they have determinants of inefficiency, rather, they use nonparametric regression methods.

8.3. **Estimation of Inefficiency.**

8.3.1. *Semiparametric Estimation of the Distribution of Inefficiency.* Virtually all stochastic frontier analyses eschew specification issues of the distributive law of inefficiency. While recent research has spent effort to test for the correct form or to determine which set of determinants belong in the parameterizations of the inefficiency distribution, the shape of this distribution is often an afterthought. The recent work of Horrace & Parmeter (2011) has proposed an estimator for the distribution of inefficiency that does not require assumptions on u (but still requires specification of v).

Horrace & Parmeter (2011) use the basic stochastic production frontier model:

$$y_i = m(\boldsymbol{x}_i) + \varepsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i - u_i. \tag{8.26}$$

To help detail how Horrace & Parmeter (2011) construct an estimator for the distribution of u, we mention that a standard result in statistics is that if two random variables are independent, which is a standard operating assumption in the stochastic frontier model, then the characteristic function of the convolution of these two random variables is simply the product of the individual characteristic functions. Since the characteristic function uniquely specifies a distribution, Horrace & Parmeter (2011) used this result to back out the characteristic function for u, and subsequently, the distribution of u.

For clarity, we let the probability densities of the error components be $f_v(z)$, $f_u(z)$ and $f_\varepsilon(z)$ with corresponding characteristic functions $\varphi_v(\tau)$, $\varphi_u(\tau)$, and $\varphi_\varepsilon(\tau)$. We use the subscripts here so that which density we are referring to is clear. Based on the assumption of independence between v_j and u_j, the key statistical condition that Horrace & Parmeter (2011) use is

$$\varphi_\varepsilon(\tau) = \varphi_v(\tau)\varphi_u(\tau). \tag{8.27}$$

To build a proper estimator for $\varphi_u(\tau)$, Horrace & Parmeter (2011) require v to be from a normal distribution. In essence, this is their identification conditional so that they can separate φ_v from φ_u. Assuming that v is from the normal family guarantees that they possess a nonzero characteristic function everywhere. Another identification assumption that is required is that u cannot look too normal. In practice this will not matter, but theoretically this is an important assumption. Essentially, it requires that u come from what is referred to as the class of ordinary smooth densities (Fan 1991). Examples of distributions that fall within the ordinary smooth family are the Laplace, gamma and exponential. This rules out a truncated normal density given that for μ large, this will look almost identical to a normal density.

Using these two distributional restrictions, the Fourier inversion formula identifies the density of u as

$$f_u(z) = \frac{1}{2\pi} \int e^{-i\tau z + \frac{1}{2}\sigma^2\tau^2} \varphi_\varepsilon(\tau) d\tau, \tag{8.28}$$

where $i = \sqrt{-1}$, see Lukacs (1968, p. 14). If φ_ε were known, we could use Equation (8.28) to recover the density of u. Unfortunately, we do not observe φ_ε. However, we can consistently estimate this using the empirical characteristic function,

$$\widehat{\varphi}_\varepsilon(\tau) = \frac{1}{N} \sum_{j=1}^{N} e^{i\tau\varepsilon_j}. \tag{8.29}$$

Unfortunately, ε_j is unobserved in (8.29), but we can estimate it by consistently estimating $\boldsymbol{\beta}$. That is, for $\widehat{\boldsymbol{\beta}}$, define residuals $\hat{\varepsilon}_j = y_j - \boldsymbol{x}_j'\widehat{\boldsymbol{\beta}}$. We will use the empirical characteristic function of the residuals which is defined as

$$\widehat{\varphi}_{\hat{\varepsilon}}(\tau) = \frac{1}{N} \sum_{j=1}^{N} e^{i\tau\hat{\varepsilon}_j}. \tag{8.30}$$

Replacing φ_ε with $\widehat{\varphi}_{\hat{\varepsilon}}$ in Equation (8.28) does not ensure that the integration will exist, so we convolute the integrand with a smoothing kernel (see Stefanski & Carroll 1990). Define a random variable z with kernel density $K(z)$ and corresponding (invertible) characteristic function $\varphi_K(\tau)$. The characteristic function, $\varphi_K(\tau)$, must have finite support to ensure that the integration exists and that the resulting estimate represents a density function. Using $K(z) = (\pi z)^{-1}\sin(z)$, $(\varphi_K(\tau) = 1\{|\tau| \leq 1\})$, our estimator of the density of u is,

$$\widehat{f}_u(z) = \frac{1}{2\pi} \int_{-1/h}^{1/h} e^{-i\tau z + \frac{1}{2}\hat{\sigma}_N^2\tau^2} \widehat{\varphi}_{\hat{\varepsilon}}(\tau) d\tau, \tag{8.31}$$

where the limits of integration are a function of a sequence of the bandwidth, which represents the degree of smoothing. Horrace & Parmeter (2011) set h to be a multiple of $\frac{\ln k_N}{k_N}$ where $k_N = \sqrt{\frac{\ln N}{\ln(\ln N)}}$. They require the bandwidth to shrink slowly as the sample size grows, much slower than the $N^{-1/5}$ rate that we discussed earlier. The reason for this is that deconvolution is a much more difficult estimation problem and larger amounts of data are required to achieve the same level of precision as say estimation of a standard density when there is no noise present.

The variance estimator is defined as (Meister 2006)

$$\hat{\sigma}_N^2 = \begin{cases} 0, & \text{if } \tilde{\sigma}_N^2 < 0 \\ \tilde{\sigma}_N^2, & \text{if } \tilde{\sigma}_N^2 \in [0, \sigma_N^2] \\ \sigma_N^2, & \text{if } \tilde{\sigma}_N^2 > \sigma_N^2, \end{cases} \tag{8.32}$$

where $\tilde{\sigma}_N^2 = -2k_N^{-2} \ln\left(\frac{|\hat{\varphi}_{\hat{\varepsilon}}(k_N)|}{C_1 k_N^{\delta}}\right)$ and $\sigma_N^2 = \ln(\ln(N))/4$. Note that the estimator for the variance of v is truncated, however, as the sample size grows this truncation becomes irrelevant. Here $\delta > 1$ and

$C_1 > 0$ are arbitrary. Both Meister (2006) and Horrace & Parmeter (2011) show that inappropriate choice of these constants has a negligible effect on the performance of the estimator.

A still unresolved issue with the density estimator of u is how to construct predictions of inefficiency as in Jondrow et al. (1982). This would be a useful construct as f_u's exact shape is only of concern as it pertains to recovering $E[u_i|\varepsilon_i]$.

8.3.2. *Nonparametric Estimation of the Mean of Inefficiency.* The standard setup of the deconvolution estimator of Horrace & Parmeter (2011) assumes the distribution of u has support everywhere, but the assumptions of the stochastic frontier model imply that the distribution of u will have a jump discontinuity at $u = 0$. The estimated residuals $\widehat{\varepsilon}$ are shifted away from this discontinuity however, by $\mu = E(u)$, which is unknown. Thus, to properly recover this shift we need to estimate $E(u)$. This is straightforward with parametric assumptions, however, if one wishes to avoid invoking distributional assumptions then a more sophisticated approach is necessary. Hall & Simar (2002) develop a procedure for detecting the location of the jump discontinuity, which allows for estimation of the mean of inefficiency.

The estimator proposed by Hall & Simar (2002) is quite intuitive. If we observed data generated by a convolution where one random variable has a jump discontinuity (as does our inefficiency variable)[38], say μ, while the other is continuous almost surely, then the point of this discontinuity causes a severe change in the derivative of the convoluted density.

Define the ordered regression residuals, $\hat{\varepsilon}_{(1)} \leq \cdots \leq \hat{\varepsilon}_{(N)}$ then Hall and Simar propose to estimate the jump discontinuity by

$$\hat{\mu} = \underset{x \in \mathcal{N}(\hat{\varepsilon}_{(\ell)})}{argmax} \quad |\widehat{f'}_{\hat{\varepsilon}}(x)|, \tag{8.33}$$

where $\mathcal{N}(\hat{\varepsilon}_{(\ell)})$ is a neighborhood around either the left ($\ell = 1$) or right tail ($\ell = N$) of the distribution where the jump discontinuity exists. That is, while a local maximum of the derivative of the kernel density estimate can occur anywhere, Hall & Simar (2002) suggest explicitly looking in a region where the jump discontinuity is likely to occur. So, for a jump discontinuity appearing on the right hand side of the distribution (as in the production frontier setting) we search near the N^{th} order statistic $\hat{\varepsilon}_{(N)}$, whereas if the jump discontinuity appears on the left hand side of the distribution (as in the cost frontier setting) we search around the first order statistic $\hat{\varepsilon}_{(1)}$. A kernel density estimator and its associated derivative is used to construct $\widehat{f'}_{\hat{\varepsilon}}(x)$.

Hall & Simar (2002) show that as $N \to \infty$, $\hat{\mu} = \mu + O(\sigma_v^2)$. Notice that this bias does not diminish as the sample size increases. They provide conditions to further reduce the bias, however, unless one is willing to assume the variance of the noise is diminishing as the sample increases, one can do no better than a biased estimate of the jump discontinuity. In simulations they show that the bias is actually quite small and so this estimator shows promise as a viable means of estimating the boundary of the distribution of u. Bandwidth selection for the kernel density estimate and

[38]The jump discontinuity occurs due to the fact that no probability is assigned to points to the left of the boundary, assuming the boundary is a lower boundary.

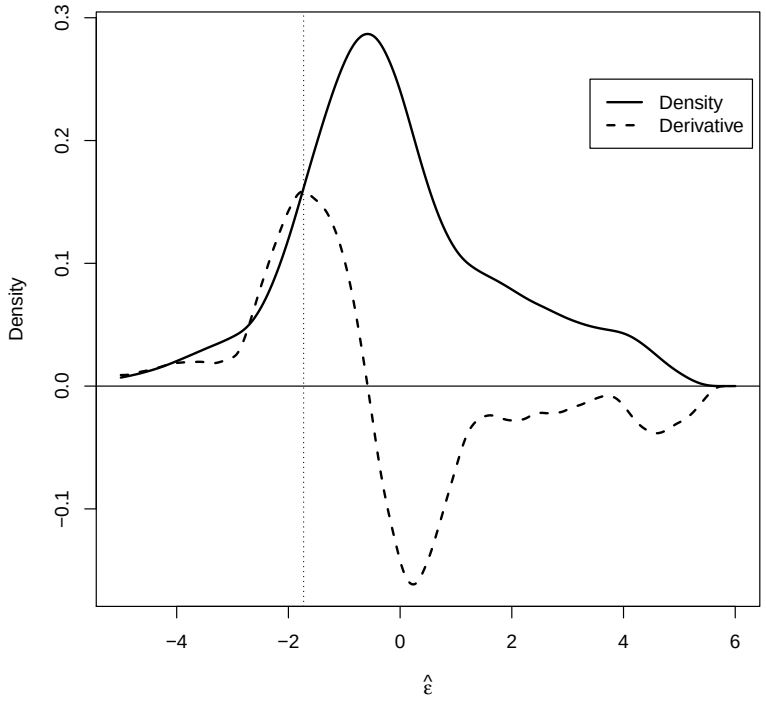

FIGURE 5. Illustration of the Detection of the Jump Discontinuity in Hall and Simar (2002).

the selection of the neighborhood are discussed in Hall & Simar (2002) and Delaigle & Gijbels (2006*b*, 2006*a*).

8.3.3. *Nonparametric Estimation of the Determinants of Inefficiency.* So far we have focused our discussion on estimation of the frontier or estimation of the distribution of inefficiency in a nonparametric fashion. An alternative, when determinants of inefficiency are available, $\boldsymbol{z}_u$, is to estimate $E(u|\boldsymbol{z}_u)$ in a nonparametric fashion. In this case we can estimate the frontier without requiring distributional assumptions, and we can recover the effect of covariates on expected inefficiency without invoking the scaling property or specifying the scaling function. Parmeter, Wang & Kumbhakar (2014) recently proposed a partly linear regression estimator for exactly this setup.

Beginning with the traditional stochastic production frontier model, with determinants of inefficiency, $\boldsymbol{z}_u$ available we have,

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i - u_i, \tag{8.34}$$

where $E[u_i] = g(\boldsymbol{z}_{ui}) \geq 0$. The main focus of Parmeter, Wang & Kumbhakar (2014) is on estimation of $g(\boldsymbol{z}_{ui})$ without imposing distributional assumptions on u_i, such as half-normal or truncated half-normal. This model is very similar to that proposed by Deprins & Simar (1989*a*, 1989*b*) and

extended by Deprins (1989). Additionally, if one were to parametrically specify $g(\boldsymbol{z}_{ui})$ and invoke the scaling property, then this model becomes that in Simar et al. (1994), Caudill et al. (1995) and Wang & Schmidt (2002). The key here is that Parmeter, Wang & Kumbhakar (2014) leave $g(\boldsymbol{z}_{ui})$ unspecified and do not invoke the scaling property.

To estimate the model in (8.34) add and subtract $g(\boldsymbol{z}_{ui})$ to obtain

$$\begin{aligned} y_i =& \boldsymbol{x}_i'\boldsymbol{\beta} + v_i - u_i \\ y_i =& \boldsymbol{x}_i'\boldsymbol{\beta} - g(\boldsymbol{z}_{ui}) + v_i - (u_i - g(\boldsymbol{z}_{ui})) \\ y_i =& \boldsymbol{x}_i'\boldsymbol{\beta} - g(\boldsymbol{z}_{ui}) + \varepsilon_i, \end{aligned} \tag{8.35}$$

where ε_i is independently but not identically distributed. The model in (8.35) is nothing more than the partly linear model of Robinson (1988) (see also Fan, Li & Stengos 1992). A key identification condition in this setup is that $\boldsymbol{x}$ and $\boldsymbol{z}_u$ do not contain common elements.

Notice that if $\boldsymbol{\beta}$ were known, $g(\boldsymbol{z}_{ui})$ could be identified as the conditional mean of $\tilde{\varepsilon}_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$ given $\boldsymbol{z}_{ui}$. However, $\boldsymbol{\beta}$ is unknown and must be estimated. To estimate $\boldsymbol{\beta}$ note that by conditioning only on $\boldsymbol{z}_{ui}$ in equation (8.35) we have

$$E[y_i|\boldsymbol{z}_{ui}] = E[\boldsymbol{x}_i|\boldsymbol{z}_{ui}]'\boldsymbol{\beta} - g(\boldsymbol{z}_{ui}). \tag{8.36}$$

Subtracting (8.36) from (8.35) yields

$$y_i - E[y_i|\boldsymbol{z}_{ui}] = (\boldsymbol{x}_i - E[\boldsymbol{x}_i|\boldsymbol{z}_{ui}])'\boldsymbol{\beta} + \varepsilon_i. \tag{8.37}$$

If $E[y_i|\boldsymbol{z}_{ui}]$ and $E[x_i|\boldsymbol{z}_{ui}]$ were known, $\boldsymbol{\beta}$ could be estimated via OLS. The intuition behind recovering a consistent estimator for $\boldsymbol{\beta}$ for the partly linear model of Robinson (1988) is to replace the unknown conditional means with nonparametric estimates.

However, an additional issue that Parmeter, Wang & Kumbhakar (2014) have to deal with that is not present in Robinson (1988), is that ε_i depends on $\boldsymbol{z}_{ui}$ through $g(\boldsymbol{z}_{ui})$. If $\boldsymbol{z}_{ui}$ and $\boldsymbol{x}_i$ are correlated this may impact OLS estimation of $\boldsymbol{\beta}$ in (8.37). Parmeter, Wang & Kumbhakar (2014) demonstrate that $\boldsymbol{x} - E[\boldsymbol{x}|\boldsymbol{z}_u]$ is uncorrelated with ε and so this is no issue whatsoever. The elegance of the framework of Parmeter, Wang & Kumbhakar (2014) is that u_i is not required to satisfy the scaling property for identification or estimation. Thus, the scaling property is imposed *ex poste* through the interpretation of $g(\boldsymbol{z}_{ui})$. Whereas the scaling property was required for parametric identification of the conditional mean, no such restriction is needed in this setup. The rationale for this result is that the scaling model provides a parametric framework where no distributional assumption is required. In general, if the scaling property is not required to hold then parametric identification of the conditional mean of inefficiency requires a distributional assumption to avoid parametric misspecification of the mean.

For example, consider the standard setting where u_i is a truncated (at zero) normal random variate with mean, $\mu(\boldsymbol{z}_{ui})$ and variance, $\sigma^2(\boldsymbol{z}_{ui})$. It is well known that the conditional mean of u_i

given $\boldsymbol{z}_{ui}$ is (as discussed in Section 5)

$$E[u_i|\boldsymbol{z}_{ui}] = \mu(\boldsymbol{z}_{ui}) + \xi(-\mu(\boldsymbol{z}_{ui})/\sigma(\boldsymbol{z}_{ui})))\sigma(\boldsymbol{z}_{ui}) \tag{8.38}$$

where $\xi(\alpha) = \phi(\alpha)/(1-\Phi(\alpha))$ is the inverse Mill's ratio. It becomes apparent from this conditional mean when the scaling property is appealing from a parametric perspective. When the scaling property holds only a single function must be correctly specified and no distributional assumption is required. When the scaling property does not hold then not only must a distributional assumption be correctly specified, but the mean and/or variance must be correctly specified with respect to dependence upon $\boldsymbol{z}_{ui}$. Note that for a truncated distribution, the mean of the truncated distribution will depend on *both* the pre-truncated mean and variance, so parameterization of each becomes crucial for appropriate estimation.

To estimate both $\boldsymbol{\beta}$ and $g(\boldsymbol{z}_{ui})$ in this framework replace $E[y_i|\boldsymbol{z}_{ui}]$ and $E[\boldsymbol{x}_i|\boldsymbol{z}_{ui}]$ in (8.37) with

$$\hat{E}[y|\boldsymbol{z}_{ui}] = \left(\sum_{j=1}^{N} K_{j\boldsymbol{z}_u} y_i\right) / \left(\sum_{j=1}^{N} K_{j\boldsymbol{z}_u}\right) \tag{8.39}$$

$$\hat{E}[\boldsymbol{x}|\boldsymbol{z}_{ui}] = \left(\sum_{j=1}^{N} K_{j\boldsymbol{z}_u} x_i\right) / \left(\sum_{j=1}^{N} K_{j\boldsymbol{z}_u}\right), \tag{8.40}$$

which is nothing more than the local constant estimator applied to y and each element of $\boldsymbol{x}$. Once bandwidths have been selected, which can be done through LSCV (Gau, Liu & Racine 2013), the conditional expectations for y and each element of $\boldsymbol{x}$ are easily constructed and can be used for OLS estimation of $\boldsymbol{\beta}$. That is, instead of the usual regression y on $\boldsymbol{x}$, one performs the modified regression $\tilde{y}$ on $\tilde{\boldsymbol{x}}$, where $\tilde{w} = w - \hat{E}[w|\boldsymbol{z}_u]$. The estimates for $\boldsymbol{\beta}$ can then be used to obtain consistent estimates of our conditional mean of inefficiency via standard nonparametric regression techniques.

Let $\check{\varepsilon}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is our estimate from the regression of $\tilde{y}$ on $\tilde{\boldsymbol{x}}$. We then estimate $g(\boldsymbol{z}_{ui})$ via local linear least squares as

$$\hat{g}(\boldsymbol{z}_{ui}) = e'(R_z'\mathcal{K}(\boldsymbol{z}_{ui})R_z)^{-1}R_z'\mathcal{K}(\boldsymbol{z}_{ui})\check{\varepsilon}_i, \tag{8.41}$$

where $e' = (1, \mathbf{0})$, where $\mathbf{0}$ is a vector of zeros of length d. Additionally, let $1_N' = (1, \ldots, 1)$ be a vector of ones of length N, $R_z = (1_N, \boldsymbol{z}_u - 1_N\boldsymbol{z}_{ui})$. We use the local linear as opposed to a local constant estimator to recover $g(\boldsymbol{z}_{ui})$ so that we not only obtain estimates of $g(\boldsymbol{z}_{ui})$ but that we obtain information on $\partial\hat{g}(\boldsymbol{z}_{ui})/\partial\boldsymbol{z}_{ui}$.

An identification issue that is inherent in the framework of Parmeter, Wang & Kumbhakar (2014) is that the intercept of technology cannot be separated from $g(\boldsymbol{z}_u)$. That is, $\boldsymbol{x}$ cannot contain the 1 that represents the intercept. Thus, while one can investigate the shape of $g(\boldsymbol{z}_u)$ and readily interpret the partial derivatives of $g(\boldsymbol{z}_u)$, the actual level of $g(\boldsymbol{z}_u)$ cannot be directly taken as an absolute measure of inefficiency. This is not a concern however as differences between $g(\boldsymbol{z}_u)$ across firms can be used as measures of relative inefficiency. For example, in practice regulators define

benchmarks in terms of top performers and are interested in efficiency of firms relative to the benchmark. In relative efficiency measures, the intercept plays no role. Another issue is that it is possible that one will obtain estimates of $g(\boldsymbol{z}_u)$ which may be negative; however, given that $g(\boldsymbol{z}_u)$ must be nonnegative this is inconsistent with the idea that $g(\boldsymbol{z}_u)$ represent average inefficiency.[39]

We also mention two important extensions of the basic model in Parmeter, Wang & Kumbhakar (2014). First, it is possible to relax the parametric specification on the production technology in (8.35). The key here is to note that in this case the model can now be written as

$$y_i = m(\boldsymbol{x}_i) - g(\boldsymbol{z}_{ui}) + \varepsilon_i, \tag{8.42}$$

where $m(\boldsymbol{x}_i)$ is the unspecified production technology. This model can be estimated using kernel methods following the additively separable kernel estimator of Kim, Linton & Hentgartner (1999). However, it is paramount in this setup that $\boldsymbol{x}$ and $\boldsymbol{z}$ do not overlap. In this setting one can relax nearly all of the heavily criticized assumptions that appear in the stochastic frontier literature.

An alternative estimator quite similar to that in (8.42) is the recent model of Simar et al. (2014), which allows both $m(\cdot)$ and $g(\cdot)$ to depend on $\boldsymbol{x}$ and $\boldsymbol{z}_u$. In this case they cannot fully identify $g(\cdot)$ but they can identify $\partial g(\cdot)/\partial \boldsymbol{x}$ or $\partial g(\cdot)/\partial \boldsymbol{z}_u$ under specific assumptions about the class of distributions to which u belongs. This is not as restrictive as assuming the exact distribution of u.

The models of Parmeter, Wang & Kumbhakar (2014) and Simar et al. (2014) offer a tradeoff in modeling assumptions. Parmeter, Wang & Kumbhakar (2014) require that traditional inputs cannot influence inefficiency while Simar et al. (2014) allow this feature, but then must restrict the class of distributions to which u can belong.

8.3.3.1. **Testing for Correct Specification of the Scaling Function**

If we assume that the scaling property holds, then we can develop a consistent model specification test for the scaling function. This is similar to the suite of tests presented in Alvarez et al. (2006) to test for the scaling property, albeit in a maximum likelihood framework. Thus, if we assume the scaling property holds, then we can estimate the production technology and conditional inefficiency using nonlinear least squares. Subsequently, we can test the functional form used for the scaling function against a nonparametric alternative using the conditional moment based test of Li & Wang (1998).

Consider the partly linear model given in (8.35). The null hypothesis of interest is

$$H_0 : P(g(\boldsymbol{z}_{ui}) = g(\boldsymbol{z}_{ui}; \boldsymbol{\gamma}_0)) = 1, \quad \text{for almost all } \boldsymbol{z}_{ui}, \quad \boldsymbol{\gamma}_0 \in \mathcal{B} \subset \mathbb{R}^d,$$

where $g(\boldsymbol{z}_{ui}; \boldsymbol{\gamma})$ is a known function with $\boldsymbol{\gamma}$ being a $d \times 1$ vector of unknown parameters and $\mathcal{B}$ is a compact subset of $\mathbb{R}^d$.

Defining $\widehat{\varepsilon}_i = y_i - \boldsymbol{x}_i'\widehat{\beta} + g(\boldsymbol{z}_{ui}; \widehat{\boldsymbol{\gamma}})$, where the estimator $\widehat{\boldsymbol{\gamma}}$ is obtained using NLS, a feasible test statistic for the null of correct parametric specification (versus the nonparametric alternative) is

[39] One can estimate $g(\boldsymbol{z}_{ui})$ ensuring positivity via the constrained regression method of Du, Parmeter & Racine (2013), but that is beyond the scope of our discussion.

given as (Equation 5 of Li & Wang 1998)

$$\widehat{I}_N = \frac{1}{N(N-1)\,|h|}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\widehat{\varepsilon}_i\widehat{\varepsilon}_j K_{ij}.$$

Alternatively, we could replace $\widehat{\boldsymbol{\beta}}$ with $\widetilde{\boldsymbol{\beta}}_{SP}$ in our definition for $\widehat{\varepsilon}$. Li & Wang (1998) note that this 'mixed' residual produces a test statistic that does not converge to 0 under the alternative without requiring further technical assumptions. With proper normalization, $\widehat{I}_N$ tends towards the standard normal in distribution for which an estimator of the variance of the test statistic is

$$\widehat{\sigma}_N^2 = \frac{2}{N\,(N-1)\,|h|}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\widehat{\varepsilon}_i^2\widehat{\varepsilon}_j^2 K_{ij}^2.$$

Given these results, the test statistic is constructed as

$$\widehat{T}_N = \frac{N\,|h|^{1/2}\,\widehat{I}_N}{\widehat{\sigma}_N}$$

and converges to the standard normal distribution under the null. Li & Wang (1998) note that we could use $\hat{\boldsymbol{\beta}}_{OLS}$ to construct the residuals and still obtain this result. Similar to other conditional moment tests in the nonparametric specification testing literature, we use a bootstrap procedure to implement the test. Given this, we use the NLS residuals as opposed to the mixed residuals as this requires estimation of the partly linear model for each bootstrap iteration. The steps for the two-point wild bootstrap procedure are as follows:

Step 1: Compute the test statistic $\widehat{T}_N$ for the original sample of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, $\{\boldsymbol{z}_{u1}, \boldsymbol{z}_{u2}, \ldots, \boldsymbol{z}_{uN}\}$ and $\{y_1, y_2, \ldots, y_n\}$ and save the re-centered residuals from the null model $\widehat{\varepsilon}_i - \overline{\widehat{\varepsilon}}$, $i = 1, 2, \ldots, N$ where $\widehat{\varepsilon}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + g(\boldsymbol{z}_{ui}; \widehat{\boldsymbol{\gamma}})$ and $\overline{\widehat{\varepsilon}} = N^{-1}\sum_{i=1}^{N}\widehat{\varepsilon}_i$.

Step 2: For each observation i, construct the bootstrapped residual ε_i^*, where $\varepsilon_i^* = \frac{1-\sqrt{5}}{2}\left(\widehat{\varepsilon}_i - \overline{\widehat{\varepsilon}}\right)$ with probability $\frac{1+\sqrt{5}}{2\sqrt{5}}$ and $\varepsilon_i^* = \frac{1+\sqrt{5}}{2}\left(\widehat{\varepsilon}_i - \overline{\widehat{\varepsilon}}\right)$ with probability $1 - \frac{1+\sqrt{5}}{2\sqrt{5}}$. Construct the bootstrapped left-hand-side variable by adding the bootstrapped residuals to the fitted values under the null as $y_i^* = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - g(\boldsymbol{z}_{ui}, \widehat{\boldsymbol{\gamma}}) + \varepsilon_i^*$. Call $\{y_i^*, \boldsymbol{x}_i, \boldsymbol{z}_{ui}\}_{i=1}^N$ the bootstrap sample.

Step 3: Calculate $\widehat{T}_N^*$ where $\widehat{T}_N^*$ is calculated the same way as $\widehat{T}_N$ except that $\widehat{\varepsilon}_i$ is replaced by $\widehat{\varepsilon}_i^*$.

Step 4: Repeat steps (2)-(3) a large number (B) of times and then construct the sampling distribution of the bootstrapped test statistics. We reject the null that the parametric model is correctly specified if the estimated test statistic $\widehat{T}_N$ is greater than the upper α-percentile of the bootstrapped test statistics.

8.4. **Almost Fully Nonparametric Approaches.** Perhaps the most general method for relaxing parametric restrictions in a stochastic frontier setting is the local likelihood approach of Kumbhakar, Park, Simar & Tsionas (2007). The local likelihood method is semiparametric as it requires

distributional assumptions to derive the likelihood function (as in Fan et al. 1996, Martins-Filho & Yao 2011), but Kumbhakar et al. (2007) allow the parameters of the density of ε to be smooth functions of the covariates. Kumbhakar et al.'s (2007) local likelihood approach is similar to the profile likelihood approach of Martins-Filho & Yao (2011), however, two separate steps are not required; this is due to the fact that in Martins-Filho & Yao (2011), the parameters λ and σ are constant, and so profiling is necessary to ensure this, whereas in Kumbhakar et al. (2007) λ and σ can depend on $\boldsymbol{x}$ and so no separate profiling step is needed.

To describe the approach of Kumbhakar et al. (2007), we begin with the likelihood function of Martins-Filho & Yao (2011) in (8.25), except that θ is now a function of $\boldsymbol{x}$:

$$\check{\mathcal{L}}_N(\theta(\boldsymbol{x}), m_{\boldsymbol{x}}) = (N|h|)^{-1} \sum_{i=1}^{N} \ln f_\varepsilon(y_i - m(\boldsymbol{x}_i); \theta(\boldsymbol{x})) K_{i\boldsymbol{x}}. \tag{8.43}$$

Kumbhakar et al. (2007) suggest using a local linear approximation for the unknown functions; coupled with the assumption of a normal, half normal convoluted error term. The full local log likelihood function is

$$\begin{aligned}\check{\mathcal{L}}(\theta(\boldsymbol{x}), m_{\boldsymbol{x}}) =& (N|h|)^{-1} \sum_{i=1}^{N} \Big[-0.5\tilde{\sigma}^2_{\boldsymbol{x}}(\boldsymbol{x}_i) - 0.5\tilde{\varepsilon}_i e^{-\tilde{\sigma}^2_{\boldsymbol{x}}(\boldsymbol{x}_i)} \\ &+ \ln \Phi\left(-\tilde{\varepsilon}_i e^{\tilde{\lambda}_{\boldsymbol{x}}(\boldsymbol{x}_i) - 0.5\tilde{\sigma}^2_{\boldsymbol{x}}(\boldsymbol{x}_i)} \right) \Big] K_{i\boldsymbol{x}}\end{aligned} \tag{8.44}$$

where we have defined $\tilde{m}_{\boldsymbol{x}}(\boldsymbol{x}_i) = \tilde{m}_0 - \tilde{m}_1'(\boldsymbol{x}_i - \boldsymbol{x})$, $\tilde{\sigma}^2_{\boldsymbol{x}}(\boldsymbol{x}_i) = \tilde{\sigma}^2_0 + \tilde{\sigma}^{2\prime}_1(\boldsymbol{x}_i - \boldsymbol{x})$ and $\tilde{\lambda}_{\boldsymbol{x}}(\boldsymbol{x}_i) = \tilde{\lambda}_0 + \tilde{\lambda}_1'(\boldsymbol{x}_i - \boldsymbol{x})$ and $\tilde{\varepsilon}_i = y_i - \tilde{m}_{\boldsymbol{x}}(\boldsymbol{x}_i)$. In this framework, for each point of interest, we have $3 + 3q$ parameters to compute, the three function estimates, $\tilde{m}_0$, $\tilde{\sigma}^2_0$ and $\tilde{\lambda}_0$ and the $3q$ derivative estimates of the functions, $\tilde{m}_1$, $\tilde{\sigma}^2_1$ and $\tilde{\lambda}_1$.

The local-likelihood function can be solved using standard optimization algorithms. As with all nonlinear optimization, judicious choice of starting values is warranted. Kumbhakar et al. (2007) suggest starting with the local linear estimates for $\tilde{m}_0$ and $\tilde{m}_1$ and the global parametric maximum likelihood estimates for σ^2 and λ. Note that $\tilde{m}_0$ needs to be corrected to reflect the fact that the conditional mean of ε is not zero from the local linear estimates. That is, the starting value for the unknown frontier should be $\widehat{m}_0^{MOLS}(\boldsymbol{x}) = \widehat{m}_0(\boldsymbol{x}) + \sqrt{2\widehat{\sigma}^2\widehat{\lambda}^2/\pi(1+\widehat{\lambda}^2)}$. Thus, the starting values are $(\widehat{m}_0^{MOLS}, \widehat{m}_1(\boldsymbol{x}))$ for $(\tilde{m}_0(\boldsymbol{x}), \tilde{m}_1(\boldsymbol{x}))$, $(\widehat{\sigma}, \mathbf{0})$ for $(\tilde{\sigma}_0(\boldsymbol{x}), \tilde{\sigma}_1(\boldsymbol{x}))$ and $(\widehat{\lambda}, \mathbf{0})$ for $(\tilde{\lambda}_0(\boldsymbol{x}), \tilde{\lambda}_1(\boldsymbol{x}))$.

One important note is the presence of $e^{-\tilde{\sigma}^2_{\boldsymbol{x}}(\boldsymbol{x}_i)}$ (or $e^{-\tilde{\lambda}_{\boldsymbol{x}}(\boldsymbol{x}_i)}$) in the likelihood function. The reason for this is that instead of operating with a constrained parameter space for $\tilde{\sigma}^2_0 > 0$ and $\tilde{\lambda}_0 > 0$, Kumbhakar et al. (2007) transform the parameter space to the real line using $\tilde{\sigma}^2 = \ln(\sigma^2)$ and $\tilde{\lambda} = \ln(\lambda)$. Thus, unconstrained optimization can be undertaken, which is computationally easier to deploy. Once estimates for $\tilde{\sigma}^2_0$ and $\tilde{\lambda}_0$ have been obtained, they can be transformed to construct estimates of σ^2_0 and λ_0 via:

$$\widehat{\sigma}^2_0 = e^{\ln \widehat{\tilde{\sigma}}^2_0}; \quad \widehat{\lambda}_0 = e^{\ln \widehat{\tilde{\lambda}}_0}.$$

With these estimates, firm level inefficiency can be estimated following the framework of Jondrow et al. (1982). The JLMS efficiency scores are

$$\widehat{u}_i = \frac{\widehat{\sigma}_0(\boldsymbol{x}_i)\widehat{\lambda}_0(\boldsymbol{x}_i)}{1+\widehat{\lambda}_0^2(\boldsymbol{x}_i)}\left[\frac{\phi\left(-\xi(\boldsymbol{x}_i)\right)}{\Phi\left(-\xi(\boldsymbol{x}_i)\right)}-\xi(\boldsymbol{x}_i)\right], \tag{8.45}$$

where $\xi(\boldsymbol{x}_i)=\widehat{\varepsilon}_i\widehat{\lambda}_0(\boldsymbol{x}_i)/\widehat{\sigma}_0(\boldsymbol{x}_i)$ and $\widehat{\varepsilon}_i = y_i - \widehat{m}_0(\boldsymbol{x}_i)$.

8.5. **Which Approach is Best?** With all of the different approaches to nonparametrically estimate the frontier, it would seem challenging to select an appropriate estimator. Firstly, we mention that little research exists on how well the above methods compare to one another across various applied dimensions, such as actual estimates but also computation time, bandwidth selection and prediction. These are all important aspects to consider when debating alternative methods.

Given the straightforwardness of Fan et al. (1996) it would seem this method is perhaps the most amenable to practice (as well as Simar et al. 2014). Both the local constant and local linear estimators can be easily deployed using available statistical software,[40] and a similar MOLS type approach could be deployed, or the Hall & Simar (2002) boundary correction can be used. This combination essentially allows one to estimate the frontier with no distributional assumptions. However, as we have stated repeatedly, in a cross section, if one wishes to predict firm level inefficiency, unless determinants are available, distributional assumptions are required.

The likelihood approaches of Kumbhakar et al. (2007) and Martins-Filho & Yao (2011) hold promise, but few empirical appearances have been made. One notable exception is Kumbhakar & Tsionas (2008), who apply the local-likelihood estimator (assuming truncated normal as opposed to half normal) to study efficiency across U.S. commercial banks. An issue that Martins-Filho & Yao (2011) raise regarding the specification of Kumbhakar et al. (2007) is that estimation may prove difficult given that all of the parameters of the likelihood function are allowed to depend on $\boldsymbol{x}$. Given Waldman's (1982) wrong skew issue, this might be too flexible of a method in some settings.

A current drawback (for potential users) for all of the nonparametric methods described here is that these estimators are not available in commercial software designed for frontier estimation. While these approaches are flexible, they entail particular coding nuances that more applied users may shy from. Future widespread availability of these methods should help to further push nonparametric stochastic frontier methods to the forefront of applied efficiency analysis and make these methods more commonplace.

[40] Of note is the `np` package (Hayfield & Racine 2008) available in the `R` programming language.

9. The Environmental Production Function and Efficiency

The production of undesirable outputs or so-called "bad" outputs is an inherent attribute of many production processes such as electric power generation, where the production of electricity (desirable output) is accompanied by the emission of pollutants (undesirable outputs). A separate literature has grown to address efficiency issues related to the so called environmental production function. The modeling of the technology and quantifying inefficiency in the presence of undesirable outputs is not a trivial issue. A standard approach is to use inputs, good and outputs as arguments in the transformation or distance function and then impose regularity conditions on the arguments to separate undesirable/bad outputs from the desirable/good outputs. Since the regularity conditions in terms of inputs and bad outputs are the same, bad outputs are effectively treated as inputs (e.g., Reinhard, Lovell & Thijssen 1999, Hailu & Veeman 2001). Such a treatment of undesirable outputs has since been heavily criticized due the implied strong disposability of undesirable outputs (Färe et al. 2005).

The literature on modeling bad outputs followed two distinct paths. The most common approach to model undesirable outputs is to specify a directional distance function (DDF) (Chung, Färe & Grosskopf 1997, Färe et al. 2005) which allows one to consider the expansion in desirable outputs and a simultaneous contraction in undesirable outputs. In this approach expansion of good outputs and contraction of bad outputs are not by the same proportion and in parametric models a quadratic function is used. In the second approach production of bad outputs is viewed as a production process and the technology of good output is separated from the production of bad outputs. In this approach one can specify and estimate both technical and environmental inefficiency. Like the stochastic frontier technology, by-production technologies specify that there is a certain minimal amount of the by-product that is produced, given the quantities of certain inputs and intended outputs. Presence of (environmental) inefficiency in by-production could generate more than this minimal amount of the unintended output. Similarly, presence of technical inefficiency in the production of intended outputs may imply that less than the maximal amount is produced, or alternatively it may imply that more than the minimal amount of inputs are used to produce a given level of intended output. In this section we briefly detail several of the extant modeling approaches.

9.1. Directional output distance function (DDF) approach.

Consider the production process in which N inputs $\boldsymbol{x}(t) \in \mathbb{R}^N_+$ are being transformed into M desirable outputs $\boldsymbol{y}(t) \in \mathbb{R}^M_+$ and P undesirable outputs $\boldsymbol{b}(t) \in \mathbb{R}^P_+$, where t denotes the time. This production process can be represented by the output set $S(\boldsymbol{x}(t)) \equiv \{(\boldsymbol{y}(t), \boldsymbol{b}(t)) : \ \boldsymbol{x}(t) \text{ can produce } (\boldsymbol{y}(t), \boldsymbol{b}(t))\}$. Following Färe et al. (2005), the maximal distance between the observed output vector $(\boldsymbol{y}(t), \boldsymbol{b}(t))$ and the frontier of the output set $S(\boldsymbol{x}(t))$ in a given direction $\boldsymbol{g} \equiv (\boldsymbol{g}_y, -\boldsymbol{g}_b) \in \mathbb{R}^M_+ \times \mathbb{R}^P_+$ is given by the value of the DODF defined as

$$\vec{D}_o(\boldsymbol{z}(t), t; \boldsymbol{g}) = \max \{\beta : \ (\boldsymbol{y}(t), \boldsymbol{b}(t)) + (\beta \boldsymbol{g}_y, -\beta \boldsymbol{g}_b) \in S(\boldsymbol{x}(t))\} \ , \tag{9.1}$$

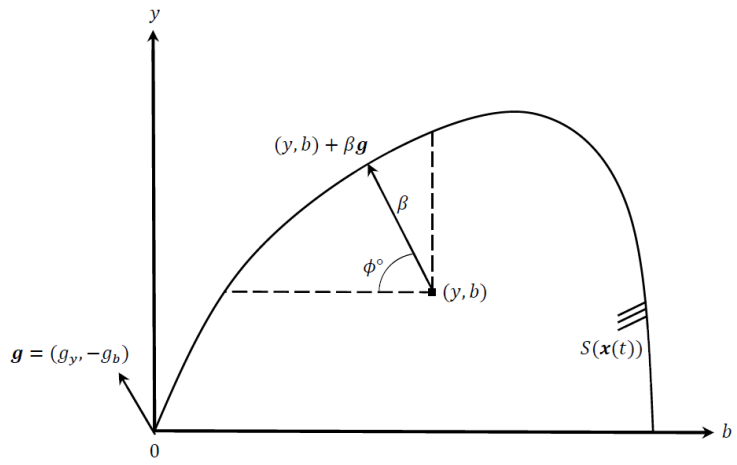

FIGURE 6. DODF with Undesirable Outputs

where $\boldsymbol{z}(t) \equiv (\boldsymbol{y}(t), \boldsymbol{b}(t), \boldsymbol{x}(t))$.

The DODF in (9.1) seeks the simultaneous maximal expansion in desirable outputs and maximal reduction in undesirable outputs. Note that, unlike the traditional RODF, the DODF constitutes an additive measure of inefficiency in a given direction $\boldsymbol{g}$, where the zero value of $\vec{D}_o(\boldsymbol{z}(t), t; \boldsymbol{g})$ implies full efficiency. The function $\vec{D}_o(\boldsymbol{z}(t), t; \boldsymbol{g})$ also needs to satisfy the following theoretical properties: non-negativity in $(\boldsymbol{y}(t), \boldsymbol{b}(t))$, negative monotonicity in $\boldsymbol{y}(t)$, positive monotonicity in $(\boldsymbol{x}(t), \boldsymbol{b}(t))$, concavity in $(\boldsymbol{y}(t), \boldsymbol{b}(t))$ and the translation property, i.e.,

$$\vec{D}_o\left(\boldsymbol{y}(t) + \kappa \boldsymbol{g}_y, \boldsymbol{b}(t) - \kappa \boldsymbol{g}_b, \boldsymbol{x}(t), t; \boldsymbol{g}\right) = \vec{D}_o\left(\boldsymbol{y}(t), \boldsymbol{b}(t), \boldsymbol{x}(t), t; \boldsymbol{g}\right) - \kappa \ , \tag{9.2}$$

for some arbitrary scalar $\kappa \in \mathbb{R}$. According to (9.2), efficiency improves by the amount κ if desired outputs are expanded by $\kappa \boldsymbol{g}_y$ while undesired outputs are reduced by $\kappa \boldsymbol{g}_b$, given inputs $\boldsymbol{x}(t)$.

Figure 6 graphically illustrates the DODF in a given direction $(g_y, -g_b)$ for the case of scalar outputs $y(t)$ and $b(t)$. A production unit (y, b) produces inside the output set $S(\boldsymbol{x}(t))$ and thus is inefficient. If it were to operate efficiently in a given direction $\boldsymbol{g}$ which corresponds to the angle ϕ, it could move to the frontier of $S(\boldsymbol{x}(t))$ by expanding the desirable output y by βg_y and simultaneously contracting the undesirable output b by βg_b. For more details, see Färe et al. (2005).

9.2. **DODF with Undesirable Outputs.** Since the translation property in (9.2) holds for any $\kappa \in \mathbb{R}$, FS suggest setting $\kappa = -y$ (hereafter referred to as normalization) which, after adding an *i.i.d.* error $\xi_y \in \mathbb{R}$ and some rearranging, yields the normalized stochastic DODF that satisfies the translation property by construction, i.e.,

$$\begin{aligned} y &= \vec{D}_o\left(y(1 - g_y), b + y g_b, \boldsymbol{x}, t; \boldsymbol{g}\right) - \vec{D}_o\left(y, b, \boldsymbol{x}, t; \boldsymbol{g}\right) + \xi_y \\ &\equiv \vec{D}_o\left(y(1 - g_y), b + y g_b, \boldsymbol{x}, t; \boldsymbol{g}\right) - \zeta_y + \xi_y \ , \end{aligned} \tag{9.3}$$

where $\zeta_y \equiv \vec{D}_o\left(y, b, \boldsymbol{x}, t; \boldsymbol{g}\right) \geq 0$ is the unobserved inefficiency.

Next, note that the direction $\boldsymbol{g}$ is uniquely defined by the ratio of g_y and g_b. Without the loss of generality, we can therefore normalize $g_y = 1$ and control the direction via g_b only. The latter makes y disappear from the right-hand side of (9.3). After assuming the quadratic form of $\vec{D}_o\left(y(1-g_y), b+yg_b, \boldsymbol{x}, t; \boldsymbol{g}\right)$, we can rewrite (9.3) as

$$y = \beta_0 + \beta_1\widetilde{b} + \sum_{n=1}^{N}\gamma_n x_n + \beta_\tau t \ + \frac{1}{2}\beta_{11}\widetilde{b}^2 + \frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\gamma_{nn'}x_n x_{n'} + \\ \frac{1}{2}\beta_{\tau\tau}t^2 \ + \sum_{n=1}^{3}\varphi_{n1}x_n\widetilde{b} + \beta_{\tau 1}t\widetilde{b} + \sum_{n=1}^{N}\gamma_{\tau n}t x_n - \zeta_y + \xi_y \tag{9.4}$$

where $\widetilde{b} \equiv b + yg_b$. Equation (9.4) can adopt not only the 45° direction used by Feng & Serletis (2014) [$g_b = 1$] but any feasible direction. Note that the above formulation is no different from a standard frontier production function since $\zeta_y \geq 0$ and $\xi_y \gtreqless 0$. Note that none of the variables are in log and inefficiency is not interpreted as a simultaneous percentage increase in good outputs and percentage decrease in bad outputs.

Tsionas et al. (2014) estimate equation (9.4) subject to symmetry, monotonicity and curvature constraints using the Bayesian approach as in Feng & Serletis (2014). Agee, Atkinson, Crocker & Williams (2014) considered a DDF in which the endogeneity of inputs and outputs is recognized (using a Bayesian GMM approach). In their model the DDF is specified as a quadratic function which is estimated subject to imposition of the theoretical properties of the DDF (see Färe et al. 2005). Atkinson & Tsionas (2014) addressed endogeneity of inputs and outputs in the DDF model that uses a quadratic functional form. Given the complicated nature of the econometric model, the Bayesian approach is commonly used to estimate this model (Atkinson & Tsionas 2014, Tsionas et al. 2014, Feng & Serletis 2014).

There are several problems in the above formulation and estimation of the quadratic DDF. These are: the choice of direction, the appropriate scaling of the variables, and the presence of a non-unitary Jacobian. The DDF-based productivity index is an implicit function of the prespecified direction, and its estimates are unlikely to be invariant to the directional choice. To mitigate this problem, a data-driven mechanism to determine the direction is advocated. The FS productivity index is sensitive to the choice of either of the two normalizations[41] that ensure that the stochastic DDF satisfies the translation property. In particular, the normalization results in the appearance of the left-hand-side "dependent variable" on the right-hand side of the regression equation. To remedy the latter endogeneity problem, the estimation of the normalized DDF requires that the *non-unit* (observation-specific) Jacobian of the transformation be taken into account in the Bayesian estimation.

Tsionas et al. (2014) suggest that the DDF-based productivity index is *not* invariant to these three issues. This is disconcerting given the policy implications that indices constructed from this

[41]That is, in the case of one desirable and one undesirable outputs, one can have either the desirable or the undesirable output as the "dependent variable" in the normalized DDF. The questions is whether this choice affects the results.

method may have. It is thus recommended that practitioners exert caution when formalizing their findings based on a specific direction for a given normalization. Further details on these issues can be found in Tsionas et al. (2014).

Given inputs, if more of good outputs are produced more bad outputs are automatically produced because of the by-production nature of bad outputs. The monotonic relationship between good and bad outputs is similar to the relationship between inputs and good outputs. This led some authors to treat bad outputs as inputs (Reinhard et al. 1999, Reinhard, Lovell & Thijssen 2000) and to define environmental inefficiency in terms of technical inefficiency. In this approach, a standard production or distance function with OO or IO inefficiency is estimated first. The estimated OO or IO inefficiency is then converted to maximum reduction in the bad output, which is labeled as environmental inefficiency. In this approach, there will be no environmental inefficiency, an outcome at odds with much of the literature in environmental efficiency if the production process is technically fully efficient. This is true in the model proposed by Fernández, Koop & Steel (2005). Note that none of these model use the DDF approach.

An alternative modeling strategy avoids some of these problems. In this approach production of good output (Y) is separated from the production of bad outputs (Z) simply because of the fact that production of Z is affected by Y, not the other way around. Given this by-production nature of bad outputs, Fernández, Koop & Steel (2002, FKS) use separate technologies for the production of each type of output. By doing so they separate technical efficiency from environmental efficiency. Instead of using the standard transformation function formulation $F(Y, X, t) = 1$ to describe the production of good outputs, they assume separability of good outputs and specify the technology as: $\Theta(Y) = h_1(X)$, where X is the vector of inputs. Similarly, the by-production of bad outputs (Z) is assumed to be separable, and the technology for bad output is specified as $\kappa(Z) = h_2(Y)$. Thus, in their formulation Θ and κ are simply aggregator functions. Technical and environmental inefficiencies in their model are introduced through the Θ and κ functions. Fernández et al. (2005) went backward. They specified the technology without the separability assumption but didn't separate technical inefficiency from environmental inefficiency because a single equation is used for the production of good and bad outputs.

Murty, Russell & Levkoff (2012, MRL hereafter) criticized the use of a single equation to describe any technology that produces both good and bad outputs as in Fernández et al. (2005) and Atkinson & Dorfman (2005). They strongly advocated the use of the by-production approach to model production of bad outputs. In their by-production approach, $F(Y, X, X^b) = 1$ is the production technology for good outputs, where X^b is the pollution generating input vector (coal or sulfur burned) and $Z = h(X^b)$ or $Z = h(X^b, Y)$ is the technology for bad outputs.

9.3. **The By-Production Model.** Before detailing the by-production model, it might be worth consider the single equation representation of the technology in which bad outputs and inputs are simply added as regressors in the transformation and/or the distance function (Atkinson & Dorfman 2005, Fernández et al. 2005, Agee et al. 2014). The model is $F(Y, X, X^b, Z) = 1$, where

the disposability assumptions on these variables are: $F_Y \geq 0$, $F_X \leq 0, F_X^b \leq 0$ and $F_Z \leq 0$ – F_Y, F_X and F_Z are partial derivatives of $F(\cdot)$. Since $F_X \leq 0, F_X^b \leq 0$ and $F_Z \leq 0$, from a pure mathematical point of view there is no difference between Z, X and X^b in $F(Y, X, X^b, Z) = 1$. That is, bad outputs can be treated as inputs (both X and X^b), and if inputs are freely disposable so are bad outputs. This is something that is highly criticized in the environmental production literature (Färe et al. 2005).

There are some other problems with this model. First, a flexible functional form on $F(Y, X, X^b, Z) = 1$ will always allow substitutability/complementarity among Y, Z, X and X^b. For example, with 2 good outputs $\partial \ln Y_1 / \partial \ln Y_2 \leq 0$, ceteris paribus, since $F_{Y_1} \geq 0$ and $F_{Y_2} \geq 0$. This might be intuitive because when less of Y_1 is produced some resources will be released which can be used to the production of Y_2. For two bad outputs $\partial \ln Z_1 / \partial \ln Z_2 \leq 0$. Is it intuitive? Does more SO_2 emission necessarily mean less NO_x? This will always be the case empirically if the monotonicity restrictions are imposed. However, from an engineering production point of view Z_1 and Z_2 might not be complementary. Also, the concave relationship between Y and Z in Färe et al. (2005) means that when less of Z_1 is produced, less good output is also produced, which in turn means less of Z_2. Thus $\partial \ln Z_1 / \partial \ln Z_2$ is expected to be positive.

The problem with the single equation representation is that Z is related to Y, viz., $Z = Z(Y)$ or more generally $Z = Z(Y, X^b)$ and this relationship has to be incorporated into the model by appending another equation or in some other form. Although Färe et al. (2005) emphasized this relationship they did not show how this relationship can be incorporated in to the model other than imposing the monotonicity constraints.

We now consider the by-production model which avoids the problems mentioned above. Following Caves, Christensen & Swanson (1981), Kumbhakar & Tsionas (2014) start with the transformation function representation of the underlying technology and extend it to accommodate IO inefficiency for the production of good outputs (Kumbhakar 2012):

$$F(Y, \theta X, t) = 1 \tag{9.5}$$

where $X \in \Re^J$ is the vector of good inputs, $Y \in \Re^M$ represents the good output vector and $\theta \leq 1$ is input-oriented technical inefficiency. The transformation function $F(\cdot)$ is assumed to satisfy all the standard monotonicity properties.

Kumbhakar & Tsionas (2014) used one pollution-generating technology for each bad output $Z_q, q = 1, \ldots, Q$. By doing so they do not allow substitutability among bad outputs. That is, the technologies for production of bad outputs are specified as:

$$z_{q,it} = g_q\left(y_{it}, x_{it}^b, t\right) + \xi_{q,it} + \eta_{q,it},\ q = 1, \ldots, Q \tag{9.6}$$

where $z_{q,it}$ represents bad output Z_q (in log) $(q = 1, \ldots, Q)$ and $x_{it}^b \in \Re^K$ are (log of) bad inputs. Furthermore, $\eta_{q,it} \geq 0$ represents environmental inefficiency in the sense that it gives the percentage over production of Z_q, *ceteris paribus*. Finally, a stochastic error, $\xi_{q,it} \lesseqgtr 0$ (similar to $v_{j,it}$), for each q is added.

Write $\hat{X}_{jit} = \theta X_{jit} \Rightarrow \hat{x}_{jit} = x_{jit} + \ln\theta_{it}$ and assume a translog form of the transformation function, i.e.,

$$\begin{aligned}\ln F\left(Y_{it}, \theta_{it}X_{j,it}, t\right) =& \alpha_o + \sum_{j=1}^{J}\alpha_j\,\hat{x}_{j,it} + \sum_{m=1}^{M}\beta_m\,y_{m,it} + \tfrac{1}{2}\sum_{j=1}^{J}\sum_{j'=1}^{J}\alpha_{jj'}\,\hat{x}_{j,it}\,\hat{x}_{j',it} + \alpha_t\,t \\ &+ \tfrac{1}{2}\alpha_{tt}\,t^2 + \sum_{m}\beta_{mt}\,y_{m,it}\,t + \tfrac{1}{2}\sum_{m=1}^{M}\sum_{m'=1}^{M}\beta_{mm'}\,y_{m,it}\,y_{m',it} \\ &+ \sum_{j=1}^{J}\sum_{m=1}^{M}\delta_{jm}\,\hat{x}_{j,it}\,y_{m,it} + \sum_{j}\alpha_{jt}\,\hat{x}_{j,it}\,t. \end{aligned} \tag{9.7}$$

Imposition of linear homogeneity with respect to inputs on it amounts to imposing the following restrictions:

$$\sum_{j=1}^{J}\alpha_j = 1, \sum_{j=1}^{J}\alpha_{jj'} = 0\ \forall\ j', \sum_{j=1}^{J}\delta_{jm} = 0,\ \forall\ m, \sum_{j}\alpha_{jt} = 0. \tag{9.8}$$

With these restrictions in place the transformation function takes the form of an input distance function (IDF), which can be expressed as $\ln F\left(Y_{it}, \theta_{it}X_{j,it}, t\right) = \ln F(Y_{it}, X_{it}, t) + \ln\theta_{it}$. For the translog function in (9.7), after some rearrangement, we get[42]

$$\begin{aligned} x_{1,it} =& \alpha_o + \sum_{j=2}^{J}\alpha_j\,\tilde{x}_{j,it} + \sum_{m=1}^{M}\beta_m\,y_{m,it} + \tfrac{1}{2}\sum_{j=2}^{J}\sum_{j'=2}^{J}\alpha_{jj'}\tilde{x}_{j,it}\tilde{x}_{j',it} + \alpha_t\,t + \tfrac{1}{2}\alpha_{tt}\,t^2 + \sum_{m}\beta_{mt}\,y_{m,it}\,t \\ &+ \tfrac{1}{2}\sum_{m=1}^{M}\sum_{m'=1}^{M}\beta_{mm'}\,y_{m,it}\,y_{m',it} + \sum_{j=2}^{J}\sum_{m=1}^{M}\delta_{jm}\,\tilde{x}_{j,it}\,y_{m,it} + \sum_{j}\alpha_{jt}\,\hat{x}_{j,it}\,t + v_{1,it} + u_{1,it} \end{aligned} \tag{9.9}$$

where $\tilde{x}_{j,it} = x_{j,it} - x_{1,it}, j = 2, \ldots, J$ and $u_{1,it} = -\ln\theta_{it} \geq 0$. We also added a stochastic noise term $v_{1,it} \gtrless 0$ in (9.9).

Assume translog functional forms on $g_q(\cdot)$ in (9.6), i.e.,

$$z_{q,it} = a_{q,0} + \boldsymbol{W}_{it}'\boldsymbol{\beta}_q + \tfrac{1}{2}\boldsymbol{W}_{it}'\boldsymbol{\Gamma}_q\boldsymbol{W}_{it} + \xi_{q,it} + \eta_{q,it},\ q = 1, \ldots, Q \tag{9.10}$$

where $\boldsymbol{W}_{it}' = \left(y_{it}', x_{it}^{b'}, t\right)$. The above equations can be written more compactly as

$$z_{q,it} = a_{q,0} + \boldsymbol{V}_{it}'\boldsymbol{\gamma}_q + \xi_{q,it} + \eta_{q,it},\ q = 1, \ldots, Q \tag{9.11}$$

The coefficients of each of the Q equations are unrestricted and the coefficients $\boldsymbol{\gamma}_q$ are different for different equations. Thus our econometric model consists of the technologies for the production of good and bad outputs. The system consists of (9.9) and (9.11). We write the error terms of this

[42]Strictly speaking, there should be a minus sign in front of $x_{1,it}$. To remove the negative sign we multiply both sides of (9.9) by -1 which changes the sign of all the coefficients and the inefficiency term $u_{1,it}$ in (9.9). The sign changes will be automatically absorbed by the estimated parameters.

system in vector form as $\boldsymbol{v}_{it} = [v_{1,it}, v_{2,it}, \ldots, v_{J,it}, \xi_{1,it}, \ldots, \xi_{Q,it}]'$ and assume

$$\boldsymbol{v}_{it} \sim N_{J+Q}\left(\boldsymbol{O}, \boldsymbol{\Sigma}\right). \tag{9.12}$$

For the one-sided inefficiency terms we assume:

$$u_{1,it} \sim N^{+}\left(0, \sigma_u^2\right), \eta_{q,it} \sim N^{+}\left(0, \omega_q^2\right), q = 1, \ldots, Q, \tag{9.13}$$

distributed independently of each other. We assume them to be independent of all other random variables and the regressors. Since there are no cross-equation restrictions in terms of the parameters and the error components (both inefficiency and noise) are assumed to be independent of each other, one can estimate these equations separately using standard stochastic cost models (because the one-sided inefficiency terms in these models are non-negative and each of them appears with a positive sign). However, it is possible to allow the noise terms to follow a multivariate normal distribution in which case the system defined by (9.9), and (9.10) or (9.11) is to be estimated jointly.

9.4. **A Single Equation Representation of the Bad Outputs Technology.** Now we consider an alternative model which is similar in spirit to the one proposed by FKS. In this model, no distinction is made between good and pollution generating inputs. The technology for good output is specified in terms of the transformation function in (9.5) in which the input vector X includes both good and bad inputs. The technology for the production of bad outputs is specified as a single equation, viz., $H(Y, \lambda Z) = 1$ where $\lambda \leq 1$ is environmental inefficiency. More specifically, $(1 - \lambda)$ 100% is the rate at which all the bad output can be reduced without reducing good outputs. Note that since there are no pollution generating inputs, the $H(\cdot)$ does not include X as arguments. Similar to the transformation function, we assume, for identification purposes, that $H(\cdot)$ is homogeneous of degree 1 in Z and a translog form for $\ln H(\cdot)$.[43] Thus, the system of equations in Model 2 are:

$$\begin{aligned} x_{1,it} =& \alpha_o + \sum_{2=1}^{J} \alpha_j \tilde{x}_{j,it} + \sum_{m=1}^{M} \beta_m y_{m,it} + \tfrac{1}{2} \sum_{j=2}^{J} \sum_{j'=1}^{J} \alpha_{jj'} \tilde{x}_{j,it} \tilde{x}_{j',it} \\ &+ \tfrac{1}{2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \beta_{mm'}\, y_{m,it}\, y_{m',it} + \sum_{j=2}^{J} \sum_{m=1}^{M} \delta_{jm}\, \tilde{x}_{j,it}\, y_{m,it} + v_{1,it} + u_{1,it} \end{aligned} \tag{9.14}$$

$$z_{1,it} = a_0 + \boldsymbol{W}_{it}'\boldsymbol{\delta} + \tfrac{1}{2}\boldsymbol{W}_{it}'\boldsymbol{\Delta}\boldsymbol{W}_{it} + \zeta_{it} + \tau_{it}, \tag{9.15}$$

where $\boldsymbol{W}_{it}' = (y_{it}', \tilde{z}_{it}', t)$, $\tilde{z}_{qit} = z_{qit} - z_{1it}$, $\tau_{it} = \ln \lambda_{it} \geq 0$ and ζ_{it} is an error term. Note that the lower case letters are logs of their uppercase counterparts.

If the noise terms as well as the inefficiency terms are assumed to be uncorrelated then one can use standard stochastic cost frontier approach to estimate the parameters and the JLMS formula to compute technical and environmental inefficiency.

[43] FKS used a more restrictive form of $H(\cdot)$, viz., $H(Y, \lambda Z) = h(Y) \cdot g(\lambda Z)$.

Ideally, the above models should satisfy the monotonicity conditions, viz.,

$$\alpha_j + \sum_{j'=2}^{J} \alpha_{jj'} \tilde{x}_{j',it} + \sum_{m=1}^{M} \delta_{jm} y_{m,it} \geq 0, j = 2, \ldots, J \tag{9.16}$$

$$\beta_m + \sum_{m'=1}^{M} \beta_{mm'} y_{m',it} + \sum_{j=2}^{J} \delta_{jm} \tilde{x}_{j,it} \leq 0, m = 1, \ldots, M. \tag{9.17}$$

If we rewrite (9.11) explicitly as:

$$\begin{aligned} z_{q,it} =& a_{q,0} + \sum_{m=1}^{M} \zeta_m y_{m,it} + \sum_{k=1}^{K} \tau_k x^b_{k,it} + \tfrac{1}{2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \zeta_{mm'} y_{m,it} y_{m',it} + \tfrac{1}{2} \sum_{k=1}^{K} \sum_{k'=1}^{K} \tau_{kk'} x^b_{k,it} x^b_{k',it} \\ &+ \sum_{k=1}^{K} \sum_{m=1}^{M} \varrho_{km} y_{m,it} x^b_{k,it} \end{aligned} \tag{9.18}$$

then

$$\zeta_m + \sum_{m'=1}^{M} \zeta_{mm'} y_{m',it} + \sum_{k=1}^{K} \varrho_{km} x^b_{k,it} \geq 0, m = 1, \ldots, M \tag{9.19}$$

$$\tau_k + \sum_{k'=1}^{K} \tau_{kk'} x^b_{k',it} + \sum_{m=1}^{M} \varrho_{km} y_{m,it} \geq 0, k = 1, \ldots, K \tag{9.20}$$

Imposition of these constraints makes estimation via maximum likelihood complicated. Kumbhakar & Tsionas (2014) resort to Bayesian methods of inference as in FKS. Further econometric details can be found in Kumbhakar & Tsionas (2014).

10. Concluding Remarks

Since their birth in 1977, stochastic frontier models have matured. Aigner et al. (1977) has surpassed 6800 citations in Google Scholar and over 840,000 hits pop up for the simple `stochastic frontier analysis` search in Google. However, maturity here does not imply there is not room for growth. In fact, some of the most important advances in the field have happened in the last 10-15 years.

The original cross-sectional sectional stochastic frontier model has been extended in many directions. Perhaps the most promising is in the area of nonparametric estimation and inference. We expect to see a big surge in these areas and our review of these methods is far from comprehensive. Future studies comparing and contrasting estimates should help to further our understanding of how these methods work (see the recent study of Andor & Hesse 2014). The recent three and four component panel data models should also help to shed light on the behavior of firm efficiency over time. Coupled with models that account for selection and the presence of fully efficient firms, there is an exciting array of new methods for one to cut their teeth on for efficiency analysis.

Our intent here was not to provide an encyclopedic coverage of the entire field of efficiency analysis. Further, there are some areas that we have not covered at all (Bayesian approaches for example). We and some of which are better covered in recent surveys and books that we have cited. We also decided not to add codes and real applications to keep the size reasonable. Many of the models we discussed can be estimated using canned softwares such as LIMDEP and STATA. Some of the standard models can be estimated in a variety of commercial softwares. We hope that this survey will be helpful to both the practitioners and researchers in understanding modeling issues in SFA.

References

Afriat, S. N. (1972), 'Efficiency estimation of production functions', *International Economic Review* **13**(3), 568–598.

Agee, M. D., Atkinson, S. E., Crocker, T. D. & Williams, J. W. (2014), 'Non-separable pollution control: Implications for a CO_2 emissions cap and trade system', *Resource and Energy Economics* **36**(1), 64–82.

Aigner, D. & Chu, S. (1968), 'On estimating the industry production function', *American Economic Review* **58**, 826–839.

Aigner, D. J., Lovell, C. A. K. & Schmidt, P. (1977), 'Formulation and estimation of stochastic frontier production functions', *Journal of Econometrics* **6**(1), 21–37.

Ali, M. & Flinn, J. C. (1989), 'Profit efficiency among Basmati rice producers in Pakistan Punjab', *American Journal of Agricultural Economics* **71**(2), 303–310.

Almanidis, P., Qian, J. & Sickles, R. C. (2014), Stochastic frontier models with bounded inefficiency, *in* R. C. Sickles & W. C. Horrace, eds, 'Festschrift in Honor of Peter Schmidt Econometric Methods and Applications', Springer: New York, pp. 47–82.

Almanidis, P. & Sickles, R. C. (2011), The skewness issue in stochastic frontier models: Fact or fiction?, *in* I. van Keilegom & P. W. Wilson, eds, 'Exploring Research Frontiers in Contemporary Statistics and Econometrics', Springer Verlag, Berlin.

Altunbas, Y., Evans, L. & Molyneux, P. (2001), 'Bank ownership and efficiency', *Journal of Money, Credit and Banking* **33**(4), 926–954.

Alvarez, A., Amsler, C., Orea, L. & Schmidt, P. (2006), 'Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics', *Journal of Productivity Analysis* **25**(2), 201–212.

Andor, M. & Hesse, F. (2014), 'The StoNED Age: The departure into a new era of efficiency analysis? A Monte Carlo comparison of StoNED and the "oldies" (SFA and DEA)', *Journal of Productivity Analysis* **41**(1), 85–109.

Arellano-Valle, R. B. & Azzalini, A. (2006), 'On the unification of families of skew-normal distributions', *Scandinavian Journal of Statistics* **33**(3), 561–574.

Atkinson, S. E. & Dorfman, J. H. (2005), 'Bayesian measurement of productivity and efficiency in the presence of undesirable outputs: crediting electric utilities for reducing air pollution', *Journal of Econometrics* **126**(3), 445–468.

Atkinson, S. & Tsionas, E. G. (2014), Directional distance functions: optimal endogenous directions. Unpublished working paper.

Azzalini, A. (1985), 'A class of distributions which includes the normal ones', *Scandinavian Journal of Statistics* **12**(2), 171–178.

Baccouche, R. & Kouki, M. (2003), 'Stochastic production frontier and technical inefficiency: A sensitivity analysis', *Econometric Reviews* **22**(1), 79–91.

Baltagi, B. H. (2013), *Econometric Analysis of Panel Data*, 5th edn, John Wiley & Sons, Great Britain.

Banker, R. D. & Maindiratta, A. (1992), 'Maximum likelihood estimation of monotone and concave production frontiers', *Journal of Productivity Analysis* **3**(4), 401–415.

Battese, G. E. & Coelli, T. J. (1988), 'Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data', *Journal of Econometrics* **38**, 387–399.

Battese, G. E. & Coelli, T. J. (1992), 'Frontier production functions, technical efficiency and panel data: With application to paddy farmers in India', *Journal of Productivity Analysis* **3**, 153–169.

Battese, G. E. & Coelli, T. J. (1995), 'A model for technical inefficiency effects in a stochastic frontier production function for panel data', *Empirical Economics* **20**(1), 325–332.

Battese, G. E. & Corra, G. S. (1977), 'Estimation of a production frontier model: With application to the pastoral zone off Eastern Australia', *Australian Journal of Agricultural Economics* **21**(3), 169–179.

Beckers, D. E. & Hammond, C. J. (1987), 'A tractable likelihood function for the normal-gamma stochastic frontier model', *Economics Letters* **24**(1), 33–38.

Bera, A. K. & Sharma, S. C. (1999), 'Estimating production uncertainty in stochastic frontier production function models', *Journal of Productivity Analysis* **12**(2), 187–210.

Bos, J. & Schmiedel, H. (2007), 'Is there a single frontier in a single European banking market?', *Journal of Banking & Finance* **31**(7), 2081–2102.

Bos, J. W. B., Economidou, C. & Koetter, M. (2010), 'Technology clubs, R&D and growth patterns: Evidence from EU manufacturing', *European Economic Review* **54**(1), 60–79.

Bos, J. W. B., Economidou, C., Koetter, M. & Kolari, J. W. (2010), 'Do all countries grow alike?', *Journal of Development Economics* **91**(1), 113–127.

Bradford, D., Kleit, A., Krousel-Wood, M. & Re, R. (2001), 'Stochastic frontier estimation of cost models within the hospital', *Review of Economics and Statistics* **83**(2), 302–309.

Bravo-Ureta, B. E. (1986), 'Technical efficiency measures for dairy farms based on a probabilistic frontier function', *Canadian Journal of Agricultural Economics* **34**(2), 400–415.

Bravo-Ureta, B. E. & Rieger, L. (1991), 'Dairy farm efficiency measurement using stochastic frontiers and neoclassical duality', *American Journal of Agricultural Economics* **73**(2), 421–428.

Butler, J. & Moffitt, R. (1982), 'A computationally efficient quadrature procedure for the one factor multinomial probit model', *Econometrica* **50**, 761–764.

Carree, M. A. (2002), 'Technological inefficiency and the skewness of the error component in stochastic frontier analysis', *Economics Letters* **77**(1), 101–107.

Caudill, S. B. (2003), 'Estimating a mixture of stochastic frontier regression models via the EM algorithm: A multiproduct cost function application', *Empirical Economics* **28**(1), 581–598.

Caudill, S. B. & Ford, J. M. (1993), 'Biases in frontier estimation due to heteroskedasticity', *Economics Letters* **41**(1), 17–20.

Caudill, S. B., Ford, J. M. & Gropper, D. M. (1995), 'Frontier estimation and firm-specific inefficiency measure in the presence of heteroskedasticity', *Journal of Business & Economic Statistics* **13**(1), 105–111.

Caves, D., Christensen, L. & Swanson, J. A. (1981), 'Productivity growth, scale economies, and capacity utilization in U.S. railroads, 1955-74', *American Economic Review* **71**(4), 994–1002.

Chamberlain, G. (1987), 'Asymptotic efficiency in estimation with conditional moment restrictions', *Journal of Econometrics* **34**(2), 305–334.

Chen, Y.-Y., Schmidt, P. & Wang, H.-J. (2014), 'Consistent estimation of the fixed effects stochastic frontier model', *Journal of Econometrics* **181**(2), 65–76.

Chew, B., Clark, K. & Bresnahan, T. (1990), Measurement, coordination and learning in a multiplant network, *in* R. Kaplan, ed., 'Measures for Manufacturing Excellence', Harvard Business School Press, Boston, pp. 129–162.

Christensen, L. R. & Greene, W. H. (1976), 'Economies of scale in U.S. electric power generation', *Journal of Political Economy* **84**(4), 655–676.

Chung, Y., Färe, R. & Grosskopf, S. (1997), 'Productivity and undesirable outputs: A directional distance function approach', *Journal of Environmental Management* **51**(3), 229–240.

Colombi, R., Kumbhakar, S., Martini, G. & Vittadini, G. (2014), 'Closed-skew normality in stochastic frontiers with individual effects and long/short-run efficiency', *Journal of Productivity Analysis* **42**(2), 123–136.

Colombi, R., Martini, G. & Vittadini, G. (2011), A stochastic frontier model with short-run and long-run inefficiency random effects. Department of Economics and Technology Management, University of Bergamo, Working Paper Series.

Cornwell, C. & Schmidt, P. (1992), 'Models for which the MLE and the conditional MLE coincide', *Empirical Economics* **17**(2), 67–75.

Cornwell, C., Schmidt, P. & Sickles, R. C. (1990), 'Production frontiers with cross-sectional and time-series variation in efficiency levels', *Journal of Econometrics* **46**(2), 185–200.

Cuesta, R. A. (2000), 'A production model with firm-specific temporal variation in technical inefficiency: With application to Spanish dairy farms', *Journal of Productivity Analysis* **13**, 139–152.

Cuesta, R. A. & Zofio, J. L. (2005), 'Hyperbolic efficiency and parametric distance functions: With application to Spanish savings banks', *Journal of Productivity Analysis* **24**(1), 31–48.

Delaigle, A. & Gijbels, I. (2006*a*), 'Data-driven boundary estimation in deconvolution problems', *Computational Statistics and Data Analysis* **50**, 1965–1994.

Delaigle, A. & Gijbels, I. (2006*b*), 'Estimation of boundary and discontinuity points in deconvolution problems', *Statistica Sinica* **16**, 773–788.

Deprins, D. (1989), *Estimation de frontieres de Production et Mesures de l'Efficacite Technique*, Louvain-la-Neuve, Belgium: CIACO.

Deprins, P. & Simar, L. (1989*a*), 'Estimating technical efficiencies with corrections for environmental conditions with an application to railway companies', *Annals of Public and Cooperative Economics* **60**(1), 81–102.

Deprins, P. & Simar, L. (1989*b*), 'Estimation de frontieres deterministes avec factuers exogenes d'inefficacite', *Annales d'Economie et de Statistique* **14**, 117–150.

Diewert, W. E. & Wales, T. J. (1987), 'Flexible functional forms and global curvature conditions', *Econometrica* **55**(1), 43–68.

Domínguez-Molina, J. A., González-Farías, G. & Ramos-Quiroga, R. (2003), Skew normality in stochastic frontier analysis. Comunicación Técnica No I-03-18/06-10-2003 (PE/CIMAT).

Du, P., Parmeter, C. F. & Racine, J. S. (2013), 'Nonparametric kernel regression with multiple predictors and multiple shape constraints', *Statistica Sinica* **23**(3), 1347–1371.

Dugger, R. (1974), An application of bounded nonparametric estimating functions to the analysis of bank cost and production functions, PhD thesis, University of North Carolina , Chapel Hill.

Fan, J. (1991), 'On the optimal rates of convergence for nonparametric deconvolution problems', *Annals of Statistics* **19**(3), 1257–1272.

Fan, Y., Li, Q. & Stengos, T. (1992), Root-n consistent semiparametric regression with conditionally heteroscedastic disturbances. Working Paper 1992-17, University of Guelph, Department of Economics.

Fan, Y., Li, Q. & Weersink, A. (1996), 'Semiparametric estimation of stochastic production frontier models', *Journal of Business & Economic Statistics* **14**(4), 460–468.

Färe, R., Grosskopf, S., Noh, D.-W. & Weber, W. (2005), 'Characteristics of a polluting technology: Theory and practice', *Journal of Econometrics* **126**(3), 469–492.

Farrell, M. J. (1957), 'The measurement of productive efficiency', *Journal of the Royal Statistical Society Series A, General* **120**(3), 253–281.

Feng, G. & Serletis, A. (2014), 'Undesirable outputs and a primal Divisia productivity index based on the directional output distance function', *Journal of Econometrics* . Forthcoming.

Feng, Q., Horrace, W. C. & Wu, G. L. (2013), Wrong skewness and finite sample correction in parametric stochastic frontier models. Center for Policy Research Working Paper 154, Syracuse University.

Fernández, C., Koop, G. & Steel, M. (2002), 'Multiple-output production with undesirable outputs: An application to nitrogen surplus in agriculture', *Journal of the American Statistical Association* **97**(458), 432–442.

Fernández, C., Koop, G. & Steel, M. (2005), 'Alternative efficiency measures for multiple-output production', *Journal of Econometrics* **126**(3), 411–444.

Flores-Lagunes, A., Horrace, W. C. & Schnier, K. E. (2007), 'Identifying technically efficient fishing vessels: A non-empty, minimal subset approach', *Journal of Applied Econometrics* **22**(4), 729–745.

Førsund, F. R. & Hjalmarsson, L. (1974), 'On the measurement of productive efficiency', *The Swedish Journal of Economics* **76**(2), 141–154.

Gagnepain, P. & Ivaldi, M. (2002), 'Stochastic frontiers and asymmetric information models', *Journal of Productivity Analysis* **18**(2), 145–159.

Gau, Q., Liu, L. & Racine, J. S. (2013), 'A partially linear kernel estimator for categorical data', *Econometric Reviews* . forthcoming.

González-Farías, G., Domínguez-Molina, J. A. & Gupta, A. K. (2004), The closed skew normal distribution, *in* M. Genton, ed., 'Skew Elliptical Distributions and their Applications: A Journal beyond Normality', Chapman and Hall/CRC, Boca Raton, Florida, chapter 2.

Grassetti, L. (2011), A novel mixture based stochastic frontier model with application to hospital efficiency. Unpublished manuscript, University of Udine.

Greene, W. H. (1980*a*), 'Maximum likelihood estimation of econometric frontier functions', *Journal of Econometrics* **13**(1), 27–56.

Greene, W. H. (1980*b*), 'On the estimation of a flexible frontier production model', *Journal of Econometrics* **13**(1), 101–115.

Greene, W. H. (1990), 'A gamma-distributed stochastic frontier model', *Journal of Econometrics* **46**(1-2), 141–164.

Greene, W. H. (2003), 'Simulated likelihood estimation of the normal-gamma stochastic frontier function', *Journal of Productivity Analysis* **19**(2), 179–190.

Greene, W. H. (2005*a*), 'Fixed and random effects in stochastic frontier models', *Journal of Productivity Analysis* **23**(1), 7–32.

Greene, W. H. (2005*b*), 'Reconsidering heterogeneity in panel data estimators of the stochastic frontier model', *Journal of Econometrics* **126**(2), 269–303.

Greene, W. H. (2010), 'A stochastic frontier model with correction for sample selection', *Journal of Productivity Analysis* **34**(1), 15–24.

Greene, W. H. & Fillipini, M. (2014), Persistent and transient productive inefficiency: A maximum simulated likelihood approach. CER-ETH - Center of Economic Research at ETH Zurich, Working Paper 14/197.

Hadri, K. (1999), 'Estimation of a doubly heteroscedastic stochastic frontier cost function', *Journal of Business & Economic Statistics* **17**(4), 359–363.

Hafner, C. M., Manner, H. & Simar, L. (2013), The "wrong skewness" problem in stochastic frontier models: A new approach. Université Catholique de Louvain Working Paper.

Hailu, A. & Veeman, T. S. (2001), 'Non-parametric productivity analysis with undesirable outputs: An application to the Canadian pulp and paper industry', *American Journal of Agricultural Economics* **83**(3), 605–616.

Hall, P. & Simar, L. (2002), 'Estimating a changepoint, boundary or frontier in the presence of observation error', *Journal of the American Statistical Association* **97**, 523–534.

Hayfield, T. & Racine, J. S. (2008), 'Nonparametric econometrics: The np package', *Journal of Statistical Software* **27**(5).
URL: *http://www.jstatsoft.org/v27/i05/*

Heckman, J. J. (1976), 'Sample selection bias as a specification error', *Econometrica* **47**(1), 153–161.

Henderson, D. J. & Parmeter, C. F. (2014), *Applied Nonparametric Econometrics*, Cambridge University Press, Cambridge, Great Britain.

Hjalmarsson, L., Kumbhakar, S. C. & Heshmati, A. (1996), 'DEA, DFA, and SFA: A comparison', *Journal of Productivity Analysis* **7**(2), 303–327.

Horrace, W. C. (2005), 'Some results on the multivariate truncated normal distribution', *Journal of Multivariate Analysis* **94**(2), 209–221.

Horrace, W. C. & Parmeter, C. F. (2011), 'Semiparametric deconvolution with unknown error variance', *Journal of Productivity Analysis* **35**(2), 129–141.

Horrace, W. C. & Parmeter, C. F. (2014), A Laplace stochastic frontier model. University of Miami Working Paper.

Horrace, W. C. & Schmidt, P. (1996), 'Confidence statements for efficiency estimates from stochastic frontier models', *Journal of Productivity Analysis* **7**, 257–282.

Horrace, W. C. & Schmidt, P. (2000), 'Multiple comparisons with the best, with economic applications', *Journal of Applied Econometrics* **15**(1), 1–26.

Hsiao, C. (2014), *Analysis of Panel Data*, 3rd edn, Cambridge University Press, Cambridge, Great Britain.

Hsieh, C.-T. & Klenow, P. J. (2009), 'Misallocation and manufacturing TFP in China and India', *Quarterly Journal of Economics* **124**(4), 1403–1448.

Huang, C. J. & Liu, J.-T. (1994), 'Estimation of a non-neutral stochastic frontier production function', *Journal of Productivity Analysis* **5**(1), 171–180.

Jondrow, J., Lovell, C. A. K., Materov, I. S. & Schmidt, P. (1982), 'On the estimation of technical efficiency in the stochastic frontier production function model', *Journal of Econometrics* **19**(2/3), 233–238.

Kalirajan, K. P. (1990), 'On measuring economic efficiency', *Journal of Applied Econometrics* **5**(1), 75–85.

Kaparakis, E., Miller, S. & Noulas, A. (1994), 'Short run cost inefficiency of commercial banks: A flexible stochastic frontier approach', *Journal of Money, Credit and Banking* **26**(1), 21–28.

Kim, M. & Schmidt, P. (2008), 'Valid test of whether technical inefficiency depends on firm characteristics', *Journal of Econometrics* **144**(2), 409–427.

Kim, W., Linton, O. B. & Hengartner, N. W. (1999), 'A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals', *Journal of Computational and Graphical Statistics* **8**(2), 278–297.

Koetter, M. & Poghosyan, T. (2009), 'The identification of technology regimes in banking: Implications for the market power-fragility nexus', *Journal of Banking & Finance* **33**, 1413–1422.

Kumbhakar, S. (2012), 'Specification and estimation of primal production models', *European Journal of Operational Research* **217**(4), 509–518.

Kumbhakar, S. C. (1987), 'The specification of technical and allocative inefficiency in stochastic production and profit frontiers', *Journal of Econometrics* **34**(1), 335–348.

Kumbhakar, S. C. (1990), 'Production frontiers, panel data, and time-varying technical inefficiency', *Journal of Econometrics* **46**(1), 201–211.

Kumbhakar, S. C. (1991), 'The measurement and decomposition of cost-inefficiency: The translog cost system', *Oxford Economic Papers* **43**(6), 667–683.

Kumbhakar, S. C. (2001), 'Estimation of profit functions when profit is not maximum', *American Journal of Agricultural Economics* **83**(1), 1–19.

Kumbhakar, S. C., Ghosh, S. & McGuckin, J. T. (1991), 'A generalized production frontier approach for estimating determinants of inefficiency in US diary farms', *Journal of Business & Economic Statistics* **9**(1), 279–286.

Kumbhakar, S. C. & Heshmati, A. (1995), 'Efficiency measurement in Swedish dairy farms: An application of rotating panel data, 1976-88', *American Journal of Agricultural Economics* **77**(3), 660–674.

Kumbhakar, S. C. & Hjalmarsson, L. (1993), Technical efficiency and technical progress in Swedish dairy farms, *in* K. L. H. Fried & S. Schmidt, eds, 'The Measurement of Productive Efficiency', Oxford University Press, Oxford, United Kingdom.

Kumbhakar, S. C. & Hjalmarsson, L. (1998), 'Relative performance of public and private ownership under yardstick competition: Electricity retail distribution', *European Economic Review* **42**(1), 97–122.

Kumbhakar, S. C., Lien, G. & Hardaker, J. B. (2014), 'Technical efficiency in competing panel data models: A study of Norwegian grain farming', *Journal of Productivity Analysis* **41**(2), 321–337.

Kumbhakar, S. C. & Lovell, C. A. K. (2000), *Stochastic Frontier Analysis*, Cambridge University Press.

Kumbhakar, S. C., Park, B. U., Simar, L. & Tsionas, E. G. (2007), 'Nonparametric stochastic frontiers: A local maximum likelihood approach', *Journal of Econometrics* **137**(1), 1–27.

Kumbhakar, S. C., Parmeter, C. F. & Tsionas, E. (2013), 'A zero inefficiency stochastic frontier estimator', *Journal of Econometrics* **172**(1), 66–76.

Kumbhakar, S. C. & Sun, K. (2013), 'Derivation of marginal effects of determinants of technical efficiency', *Economics Letters* **120**(2), 249–253.

Kumbhakar, S. C. & Tsionas, E. G. (2006), 'Estimation of stochastic frontier production functions with input-oriented technical inefficiency', *Journal of Econometrics* **133**(1), 71–96.

Kumbhakar, S. C. & Tsionas, E. G. (2008), 'Scale and efficiency measurement using a semiparametric stochastic frontier model: evidence from the U.S. commercial banks', *Empirical Economics* **34**(3), 585–602.

Kumbhakar, S. C. & Tsionas, E. G. (2014), The good, the bad and the inefficiency: A system approach to model environmental production technology. Advanced Lecture at the Taiwan Efficiency and Productivity Conference, Unpublished working paper.

Kumbhakar, S. C., Tsionas, E. G. & Sipiläinen, T. (2009), 'Joint estimation of technology choice and technical efficiency: an application to organic and conventional dairy farming', *Journal of Productivity Analysis* **31**(2), 151–161.

Kumbhakar, S. C. & Wang, H.-J. (2005), 'Production frontiers, panel data, and time-varying technical inefficiency', *Journal of Econometrics* **46**(1), 201–211.

Kumbhakar, S. C. & Wang, H.-J. (2006), 'Estimation of technical and allocative inefficiency: A primal system approach', *Journal of Econometrics* **134**(3), 419–440.

Kumbhakar, S. C., Wang, H.-J. & Horncastle, A. (2014), *A Practitioner's Guide to Stochastic Frontier Analysis*, Cambridge University Press, Cambridge, England.

Kuosmanen, T. (2008), 'Representation theorem for convex nonparametric least squares', *Econometrics Journal* **11**(2), 308–325.

Kuosmanen, T. (2012), 'Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model', *Energy Economics* **34**, 2189–2199.

Kuosmanen, T. & Fosgerau, M. (2009), 'Neoclassical versus frontier production models? Testing for the skewness of regression residuals', *The Scandinavian Journal of Economics* **111**(2), 351–367.

Kuosmanen, T. & Johnson, A. (2010), 'Data envelopment analysis as nonparametric least-squares regression', *Operations Research* **58**(1), 149–160.

Kuosmanen, T. & Kortelainen, M. (2012), 'Stochastic non-smooth envelopment of data: semi-parametric frontier esitmation subject to shape constraints', *Journal of Productivity Analysis* **38**(1), 11–28.

Lai, H.-P. & Huang, C. J. (2010), 'Likelihood ratio tests for model selection of stochastic frontier models', *Journal of Productivity Analysis* **34**(1), 3–13.

Lai, H., Polachek, S. & Wang, H.-J. (2009), Estimation of a stochastic frontier model with sample selection. Working Paper, Department of Economics, National Chung Cheng University, Taiwan.

Lau, L. (1978), Applications of profit functions, *in* M. Fuss & D. L. McFadden, eds, 'Production Economics: A Dual Approach to Theory and Applications Volume I: The Theory of Production', North Holland: Elsevier, Amsterdam, The Netherlands.

Lee, L. (1983), 'A test for distributional assumptions for the stochastic frontier function', *Journal of Econometrics* **22**(2), 245–267.

Lee, Y. & Schmidt, P. (1993), A production frontier model with flexible temporal variation in technical efficiency, *in* K. L. H. Fried & S. Schmidt, eds, 'The Measurement of Productive Efficiency', Oxford University Press, Oxford, United Kingdom.

Leibenstein, H. (1966), 'Allocative efficiency vs. 'X-efficiency'', *American Economic Review* **56**(3), 392–415.

Li, Q. (1996), 'Estimating a stochastic production frontier when the adjusted error is symmetric', *Economics Letters* **52**(3), 221–228.

Li, Q. & Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Li, Q. & Racine, J. S. (2003), 'Nonparametric estimation of distributions with categorical and continuous data', *Journal of Multivariate Analysis* **86**, 266–292.

Li, Q. & Wang, S. (1998), 'A simple consistent bootstrap test for a parametric regression function', *Journal of Econometrics* **87**(2), 145–165.

Lukacs, E. (1968), *Stochastic Convergence*, Ratheon Education Company, Lexington, Massachusetts.

Martins-Filho, C. B. & Yao, F. (2011), 'Semiparametric stochastic frontier estimation via profile likelihood', *Econometric Reviews* . Forthcoming.

McFadden, D. (1989), 'A method of simulated moments for estimation of discrete response models without numerical integration', *Econometrica* **57**(5), 995–1026.

McLachlan, G. J. & Peel, D. (2000), *Finite Mixture Models*, Wiley, New York, NY.

Meeusen, W. & van den Broeck, J. (1977), 'Efficiency estimation from Cobb-Douglas production functions with composed error', *International Economic Review* **18**(2), 435–444.

Meister, A. (2006), 'Density estimation with normal measurement error with unknown variance', *Statistica Sinica* **16**(1), 195–211.

Mester, L. (1993), 'Efficiency in the savings and loan industry', *Journal of Banking & Finance* **17**(2/3), 267–286.

Mester, L. (1996), 'A study of bank efficiency taking into account risk preferences', *Journal of Banking & Finance* **20**(6), 1025–1045.

Mundlak, Y. (1961), 'Empirical production function free of management bias', *Journal of Farm Economics* **43**(1), 44–56.

Murty, S., Russell, R. R. & Levkoff, S. B. (2012), 'On modeling pollution-generating technologies', *Journal of Environmental Economics and Management* **64**(1), 117–135.

Newman, C. & Matthews, A. (2006), 'The productivity performance of Irish dairy farms 1984-2000: a multiple output distance function approach', *Journal of Productivity Analysis* **26**(2), 191–205.

Neyman, J. & Scott, E. L. (1948), 'Consistent estimation from partially consistent observations', *Econometrica* **16**, 1–32.

Nguyen, N. B. (2010), Estimation of technical efficiency in stochastic frontier analysis, PhD thesis, Bowling Green State University.

O'Hagan, A. & Leonard, T. (1976), 'Bayes estimation subject to uncertainty about parameter constraints', *Biometrika* **63**(1), 201–203.

Olson, J. A., Schmidt, P. & Waldman, D. A. (1980), 'A Monte Carlo study of estimators of stochastic frontier production functions', *Journal of Econometrics* **13**, 67–82.

Ondrich, J. & Ruggiero, J. (2001), 'Efficiency measurement in the stochastic frontier model', *European Journal of Operational Research* **129**(3), 434–442.

Orea, L. & Kumbhakar, S. C. (2004), 'Efficiency measurement using a latent class stochastic frontier model', *Empirical Economics* **29**(1), 169–183.

Parmeter, C. F. & Racine, J. S. (2012), Smooth constrained frontier analysis, *in* X. Chen & N. Swanson, eds, 'Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr.', Springer-Verlag, New York, New York, chapter 18, pp. 463–489.

Parmeter, C. F., Sun, K., Henderson, D. J. & Kumbhakar, S. C. (2014), 'Regression and inference under economic restrictions', *Journal of Productivity Analysis* **41**(1), 111–129.

Parmeter, C. F., Wang, H.-J. & Kumbhakar, S. C. (2014), Nonparametric estimation of the determinants of inefficiency. Department of Economics, University of Miami, Working Paper Series.

Perelman, M. (2011), 'X-Efficiency', *Journal of Economic Perspectives* **25**(4), 211–222.

Pitt, M. M. & Lee, L.-F. (1981), 'The measurement and sources of technical inefficiency in the Indonesian weaving industry', *Journal of Development Economics* **9**(1), 43–64.

Racine, J. S. (2008), 'Nonparametric econometrics: A primer', *Foundations and Trends in Econometrics* **3**(1), 1–88.

Racine, J. S. & Li, Q. (2004), 'Nonparametric estimation of regression functions with both categorical and continuous data', *Journal of Econometrics* **119**(1), 99–130.

Reifschneider, D. & Stevenson, R. (1991), 'Systematic departures from the frontier: A framework for the analysis of firm inefficiency', *International Economic Review* **32**(1), 715–723.

Reinhard, S., Lovell, C. A. K. & Thijssen, G. (1999), 'Econometric estimation of technical and environmental efficiency: An application to Dutch dairy farms', *American Journal of Agricultural Economics* **81**(1), 44–60.

Reinhard, S., Lovell, C. & Thijssen, G. (2000), 'Environmental efficiency with multiple environmentally detrimental variables; estimated with SFA and DEA', *European Journal of Operational Research* **121**(3), 287–303.

Rho, S. & Schmidt, P. (2013), 'Are all firms inefficient?', *Journal of Productivity Analysis* . forthcoming.

Richmond, J. (1974), 'Estimating the efficiency of production', *International Economic Review* **15**(2), 515–521.

Ritter, C. & Simar, L. (1997), 'Pitfalls of normal-gamma stochastic frontier models', *Journal of Productivity Analysis* **8**(2), 167–182.

Robinson, P. M. (1988), 'Root-n consistent semiparametric regression', *Econometrica* **56**, 931–954.

Ruggiero, J. (1999), 'Efficiency estimation and error decomposition in the stochastic frontier model: A Monte Carlo analysis', *European Journal of Operational Research* **115**(6), 555–563.

Ryan, D. L. & Wales, T. J. (2000), 'Imposing local concavity in the translog and generalized Leontief cost functions', *Economics Letters* **67**(1), 253–260.

Schmidt, P. (1976), 'On the statistical estimation of parametric frontier production functions', *The Review of Economics and Statistics* **58**(2), 238–239.

Schmidt, P. & Lin, T.-F. (1984), 'Simple tests of alternative specifications in stochastic frontier models', *Journal of Econometrics* **24**(3), 349–361.

Schmidt, P. & Sickles, R. C. (1984), 'Production frontiers and panel data', *Journal of Business & Economic Statistics* **2**(2), 367–374.

Shephard, R. W. (1953), *Cost and Production Functions*, Princeton University Press, Princeton, NJ.

Silvapulle, M. & Sen, P. (2005), *Constrained Statistical Inference*, WILEY, Hoboken, New Jersey.

Simar, L., Lovell, C. A. K. & van den Eeckaut, P. (1994), Stochastic frontiers incorporating exogenous influences on efficiency. Discussion Papers No. 9403, Institut de Statistique, Universite de Louvain.

Simar, L., Van Keilegom, I. & Zelenyuk, V. (2014), Nonparameric least squares methods for stochastic frontier models. Centre for Efficiency and Productivity Analysis, Working Paper Series, No. WP03/2014.

Simar, L. & Wilson, P. W. (2007), 'Estimation and inference in two-stage, semi-parametric models of production processes', *Journal of Econometrics* **136**(1), 31–64.

Simar, L. & Wilson, P. W. (2010), 'Inferences from cross-sectional, stochastic frontier models', *Econometric Reviews* **29**(1), 62–98.

Simar, L. & Wilson, P. W. (2013), 'Estimation and inference in nonparametric frontier models: Recent developments and perspectives', *Foundations and Trends in Econometrics* **5**(2), 183–337.

Sipiläinen, T. & Oude Lansink, A. (2005), 'Learning in switching to organic farming', *Nordic Association of Agricultural Scientists NJF Report* **1**(1).

Stefanski, L. & Carroll, R. J. (1990), 'Deconvoluting kernel density estimators', *Statistics* **21**(3), 169–184.

Stevenson, R. (1980), 'Likelihood functions for generalized stochastic frontier estimation', *Journal of Econometrics* **13**(1), 58–66.

Stigler, G. (1976), 'The Xistence of X-Efficiency', *American Economic Review* **66**(1), 213–236.

Syverson, C. (2011), 'What determines productivity?', *Journal of Economic Literature* **49**(2), 326–365.

Taube, R. (1988), Möglichkeiten der effizienzmess ung von öffentlichen verwaltungen. Duncker & Humbolt GmbH, Berlin.

Tauer, L. W. (1998), 'Cost of production for stanchion versus parlor milking in New York', *Journal of Dairy Science* **81**(4), 567–569.

Timmer, C. P. (1971), 'Using a probabilistic frontier production function to measure technical efficiency', *The Journal of Political Economy* **79**(4), 776–794.

Tsionas, E. G. & Kumbhakar, S. C. (2004), 'Markov switching stochastic frontier model', *Econometrics Journal* **7**(2), 398–425.

Tsionas, E. G., Kumbhakar, S. C. & Malikov, E. (2014), Estimation of input distance functions: A system approach. State University of New York at Binghamton Working Paper.

Waldman, D. M. (1982), 'A stationary point for the stochastic frontier likelihood', *Journal of Econometrics* **18**(1), 275–279.

Wang, H.-J. (2002), 'Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model', *Journal of Productivity Analysis* **18**(2), 241–253.

Wang, H.-J. & Ho, C.-W. (2010), 'Estimating fixed-effect panel stochastic frontier models by model transformation', *Journal of Econometrics* **157**(2), 286–296.

Wang, H.-J. & Schmidt, P. (2002), 'One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels', *Journal of Productivity Analysis* **18**, 129–144.

Wang, W. S., Amsler, C. & Schmidt, P. (2011), 'Goodness of fit tests in stochastic frontier models', *Journal of Productivity Analysis* **35**(1), 95–118.

Wang, W. S. & Schmidt, P. (2009), 'On the distribution of estimated technical efficiency in stochastic frontier models', *Journal of Econometrics* **148**(1), 36–45.

Wheat, P., Greene, B. & Smith, A. (2014), 'Understanding prediction intervals for firm specific inefficiency scores from parametric stochastic frontier models', *Journal of Productivity Analysis* **42**, 55–65.

White, H. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48**, 817–838.

Wollni, M. & Brümmer, B. (2012), 'Productive efficiency of speciality and conventional coffee farmers in Costa Rica: Accounting for technological heterogeneity and self-selection', *Food Policy* **37**(1), 67–76.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edn, MIT Press, Cambridge, Massachusetts.