STOCHASTIC FRONTIER ANALYSIS: FOUNDATIONS AND ADVANCES

SUBAL C. KUMBHAKAR, CHRISTOPHER F. PARMETER, AND VALENTIN ZELENYUK

Abstract. This chapter reviews some of the most important developments in the econometric estimation of productivity and efficiency surrounding the stochastic frontier model. We highlight endogeneity issues, recent advances in generalized panel data stochastic frontier models, nonparametric estimation of the frontier, quantile estimation and distribution free methods. An emphasis is placed on highlighting recent research and providing broad coverage, while details are left for further reading in the abundant (although not limited to) list of references provided.

JEL Classification: C10, C13, C14, C50

Contents

Date: September 20, 2017.
Key words and phrases. Efficiency, Productivity, Panel Data, Endogeneity, Nonparametric, Determinants of Inefficiency, Quantile, Identification.
Subal C. 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. Valentin Zelenyuk School of Economics and Centre for Efficiency and Productivity Analysis, University of Queensland; e-mail: v.zelenyuk@uq.edu.au.

1. Introduction and Overview

The primary goal of this chapter is to introduce the wide audience of this Handbook to the range of methods, developed over the last four decades, within one of the most popular paradigms in modern productivity analysis - the approach called Stochastic Frontier Analysis, often abbreviated as SFA.

The first, and one of the most important, questions a reader might wonder about is why a researcher on productivity should ever care about SFA in general and especially about such an enormous variety of different types of SFA models that have been proposed over the last four decades. Our goal in writing this chapter was to provide a reader with a good answer to this important question. Here, we strive to outline the essence of major types of SFA methods, providing minimal and the most essential details, and focusing on advantages and disadvantages of each method for dealing with various aspects that arise in practice. We hope that upon finishing reading this chapter a reader who is barely or even not at all familiar with SFA gets a general understanding of the importance and relevance of different SFA methods, along with useful/key references for further details on each method. Of course, the reader also deserves to get a quick answer now, to decide if it is worth it for a productivity researcher to read this chapter further - we try to give such a quick answer in this section.

The Nobel Laureate Paul Krugman was hardly exaggerating when he once quipped that "Productivity isn't everything, but in the long run, it's almost everything." The root of this statement can be seen when looking at various theoretical models of economic growth, e.g., starting from Solow's growth model, the related variations of more advanced growth theory models or empirical growth accounting approach to productivity measurement, as well as from the more sophisticated measurements of productivity. Regardless of how the productivity is measured, it is inevitably tied to measuring production relationships. Such relationships are usually modeled through the so-called production functions or, more generally, transformation functions (e.g., Shephard's distance functions, Directional distance

functions), cost functions, etc. In the classical growth accounting approach (Solow 1957), all the variation in growth apart from the variation of inputs is attributed to the so-called Solow's residual, which under certain restrictions measures what is referred to as the change in total factor productivity (TFP).

A well known problem of simple growth accounting is that it piles up and hides many sources for growth, the most obvious of which is the statistical error. Standard regression methods can and are often used to, basically, estimate *average* relationships conditional on various factors (inputs, demographic and geographic factors, etc.) to filter out the effect of statistical noise. All the deviations from the estimated regression curves in such approaches are attributed to the statistical error, and all the decision making units (DMUs) represented in the data as observations (e.g., firms, countries, etc.), are typically assumed to be fully efficient or on the frontier of the production relationship. Such full efficiency assumption certainly simplifies the measurement complexity, but is it really an innocent assumption?

Indeed, while many economic models admit the assumption that all firms are efficient, the reality that one can observe in practice usually suggest there are quite a bit of inefficiencies in this world. Such inefficiencies could arise, for example, because of asymmetric information or more generally, the problem of incomplete markets (e.g., see Stiglitz & Greenwald 1986), which to some extent is present almost in every aspect of our lives. Differences in inefficiencies (or in relative productivity levels)[1] across firms or countries can also arise due different managerial practices (e.g., see Bloom et al. 2016), which could in turn be implied by the asymmetric information problem, different cultural beliefs, traditions and expectations (Benabou & Tirole 2016). Does accounting for such inefficiency matter for productivity measurement? Vast literature on the subject suggests that it indeed often matter substantially,

[1]Despite the variety of definitions, intuitively, production efficiency can be understood as a relative measure of productivity. In other words, production efficiency is a productivity measure that is being normalized (e.g., to be between 0 and 1 to reflect percentages) relative to some benchmark, such as the corresponding frontier outcome, optimal with respect to some criteria: e.g., maximal output given certain level of input and technology in the case of technical efficiency, or minimal cost given certain level of output and technology in the case of cost efficiency.

as have been documented in thousands of articles in the last four decades. The difference is in the approach - SFA, data envelopment analysis (DEA), free disposable hull (FDH), etc. - and the goal of this chapter is to give a sense of a few major approaches within the SFA paradigm.

In a nutshell, the main premise of the SFA approach is a recognition that whether all DMUs are efficient or not is an empirical question that can and should be statistically tested against the data, while allowing for a statistical error. To enable such testing, the SFA approach provides a framework where production relationship is estimated also as a conditional average (of outputs given inputs and other factors, in the case of production function) but the total deviation from the regression curve is decomposed into two terms - statistical noise and inefficiency. Both of these terms are unobserved by a researcher but with relatively mild assumptions the different approaches within SFA allow the analyst to estimate them for the sample as a whole (e.g., representing an industry) or for each individual DMU.

Importantly, SFA approach also allows for the inefficiency term to be statistically insignificant, if the data might suggest so, thus encompassing the classical approach with a naive assumption of full efficiency as a special case and, importantly, allowing for this assumption to be tested. Moreover, the SFA approach also encompasses the other extreme where one assumes no statistical noise with all the deviations treated as inefficiency to the frontier. Thus, the SFA approach is a natural compromise between approaches that make two extreme assumptions, yet also encompass them as special cases, which can still be followed if the data and the statistical tests from SFA would not recommend otherwise. If the tests support (or at least cannot reject) the full efficiency hypothesis then one can proceed with the standard regression techniques, or even with the Solow's growth accounting, but if not then accounting for possible inefficiency could be critical for both quantitative and qualitative conclusions and, perhaps more importantly, for the resulting policy implications. Indeed, if statistical tests reject the hypothesis of full efficiency of DMUs, then it can be imperative

to decompose the productivity (be it Solow's residual or any other productivity measure) further - to estimate the inefficiency component for the sample (e.g. representing an industry) and for each individual DMU. Moreover, the SFA also provides a framework to analyze the sources of production inefficiency or in variation of productivity levels, which can give important insights into how to reduce the inefficiency. We discuss these interesting and important issues in this chapter. While some of the stylized facts we present here can be also found in previous reviews (Lovell 1993, Kumbhakar & Lovell 2000, Greene 2008, Parmeter & Kumbhakar 2014, Kumbhakar, Wang & Horncastle 2015), and it is impossible to give a good review without following them to some degree, here we also summarize many of (what we believe to be) the key recent developments as well as (with their help) shed some novel perspectives onto the workhorse methods. So, all in all, our belief is that there is much value added for the reader to complement what was done well in earlier reviews on this theme.

The rest of the chapter is structured as follows: Sections 2-4 focus on stochastic frontier models (SFM) for cross-sectional variation in efficiency (relative productivity), where Section 2 covers the foundation laid by Aigner, Lovell & Schmidt (1977) and some closely related research, with Section 3 discussing endogeneity issues and Section 4 focuses on modeling the determinants of inefficiency. Section 5 focuses on SFA models for analyzing variation of efficiency (or relative productivity) not only across firms but also over time, i.e., in the panel data context. Section 6 reviews several prominent semi- and nonparametric approaches to SFA. Section 7 briefly discusses a recent vein of literature focusing on quantile estimation of the SFM. Section 8 presents some further extensions of the SFM, while Section 9 briefly summarizes some of the available software to estimate SFMs in practice. Section 10 concludes.

2. The Benchmark SFM

One of the main approaches to study productivity and efficiency of a cross-section of firms is the SFM, independently proposed by Aigner et al. (1977) (ALS hereafter) and Meeusen

& van den Broeck (1977*a*) (MvB hereafter).[2] Using conventional notation, let Y_i be the single-output for observation (e.g., firm) i and let $y_i = \ln(Y_i)$. The SFM can be written for a production frontier[3] as

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

Here $m(\boldsymbol{x}_i;\boldsymbol{\beta})$ represents the production frontier of a firm (or more generally a DMU), with given input vector $\boldsymbol{x}_i$. Our use of $\boldsymbol{\beta}$ is to clearly signify that we are parametrically specifying our production function.[4] The main difference between a standard production function setup and the SFM is the presence of two distinct error terms in the model. The u_i term captures inefficiency, shortfall from maximal output dictated by the production technology, while the v_i term captures stochastic shocks. The standard neoclassical production function model assumes full efficiency – so the SFM embraces it as a special case, when $u_i = 0$, $\forall i$, and allows the researcher to test this statistically.[5]

One shortcoming of the benchmark SFM is that the appearance of inefficiency in (2.1) lacks any specific structural interpretation. Where is inefficiency coming from? It could stem from inputs being used sub-optimally: workers may not put forth full effort or capital may be improperly used, e.g., due to asymmetric information or other reasons hidden to the researcher or even the firm. Without a specific structural link it is difficult to know just how to treat inefficiency in (2.1). Thus, to estimate the model, several assumptions need to be imposed. First, it is commonly assumed that inputs are independent of u and v, $u_i \perp \boldsymbol{x}$ and $v_i \perp \boldsymbol{x}$ $\forall \boldsymbol{x}$.[6] Second, u and v are assumed to be independent of one another. Next, given

[2]Battese & Corra (1977) and Meeusen & van den Broeck (1977*b*), while appearing in the same year, are applications of the methods.

[3]Our discussion in this chapter will focus on a production frontier, as it is the most popular object of study, while the framework for dual characterizations (e.g., cost, revenue, profit) or other frontiers is similar and follows with only minor changes in notation.

[4]See Section 6 for a discussion on relaxing parametric restrictions on the production frontier in the SFM.

[5]Prior to the development of the SFM, approaches which intended to model inefficiency typically ignored v_i leading to estimators of the SFM with less desirable statistical properties: see the work of Aigner & Chu (1968), Timmer (1971), Afriat (1972), Dugger (1974), Richmond (1974), and Schmidt (1976).

[6]See Section 3 for a discussion on estimation of the SFM when some inputs are allowed to be endogenous.

that u_i leads directly to a shortfall in output it must come from a one-sided distribution implying that $E[\varepsilon_i|\boldsymbol{x}] \neq 0$. This has two effects if one were to estimate the SFM using OLS. First, the intercept of technology would not be identified, and second, without any additional information, nothing can be said about inefficiency. Additionally, if u_i is an independently and identically distributed random variable, there is no policy implication behind it given that nothing can directly increase or decrease inefficiency. That is, the conclusions of such a study would be descriptive (reporting presence or absence of inefficiency) rather than prescriptive or normative.[7].

Denote $E[u]$ as μ_u and $\varepsilon_i^* = v_i - (u_i - \mu_u)$, the benchmark SFM can be rewritten as

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

and $E[\varepsilon_i^*|x] = 0$. The OLS estimator could be used to recover mean inefficiency adjusted technology $m^*(\boldsymbol{x}_i;\boldsymbol{\beta}) = m(\boldsymbol{x}_i,\boldsymbol{\beta}) - \mu_u$ in this case. However, rarely is the sole focus of an analysis of productivity on the production technology. It is more likely that both the production technology and information about inefficiency for each DMU are the targets of interest; more structure is required on the SFM in this case.

ALS' and MvB's approach to extract information on inefficiency, while also estimating technology, was to impose distributional assumptions on u_i and v_i, recovering the implied distribution for ε_i and then estimating all of the parameters of the SFM with the maximum likelihood estimator (MLE). v_i was assumed to be distributed as a normal with mean 0 and variance σ_v^2 by both sets of researchers, while the distribution of u_i differed across the papers; Aigner et al. (1977) assumed that u_i was generated from a half-normal distribution, $N_+(0,\sigma_u^2)$, whereas MvB assumed u_i was distributed exponentially, with parameter σ_u.[8]

Even though the half-normal and exponential distributions are distinct, they possess several common aspects. Both densities have modes at zero and monotonically decay (albeit at

[7]See Section 4 for models handling determinants of inefficiency

[8]ALS also briefly discussed the exponential distribution, but its use and development is mainly attributed to MvB.

different speeds) as u_i increases. The zero mode property is indicative of an industry where there is a tendency for higher efficiency for the majority of the DMUs. Both densities are what are classified as single parameter distributions, which means that the mean and variance both depend on the single parameter, and these distributions also possess the scaling property, which we will discuss in section 4.3.

2.1. **The Distribution of ε.** Estimation of the SFM in (2.1) with maximum likelihood requires that the density of ε, $f(\varepsilon)$, is known. $f(\varepsilon)$ can be determined through the distributional assumptions invoked for v and u. Not all pairs of distributional assumptions for v and u will lead to a tractable density of $f(\varepsilon)$, permitting estimation via maximum likelihood. Fortunately, the half-normal specification of Aigner et al. (1977) and the exponential specification of MvB (along with the normal assumption for v), produce a density for ε that has a closed form solution; direct application of maximum likelihood is straightforward in this setting. For brevity we report the density of the composed error for the normal-half-normal specification.

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

where $\phi(\cdot)$ is the standard normal probability density function (pdf), $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf), with the parameterization $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$ and $\lambda = \sigma_u/\sigma_v$. λ is commonly interpreted as the proportion of variation in ε due to inefficiency. The density of ε in (2.3) can be characterized as that of a skew normal random variable with location parameter 0, scale parameter σ and skew parameter $-\lambda$.[9] This connection has only recently appeared in the efficiency and productivity literature (Chen, Schmidt & Wang 2014).

[9]The pdf of a skew normal random variable x is $f(x) = 2\phi(x)\Phi(\alpha x)$. 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 pdf of x with location ξ, scale ω, and skew parameter α is $f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right)$. See O'Hagan & Leonard (1976) and Azzalini (1985) for more details.

From $f(\varepsilon)$ in (2.3), along with independence assumptions on u_i and v_i the log-likelihood function is

$$\ln \mathcal{L} = \ln \left(\prod_{i=1}^{n} f(\varepsilon_i) \right) = -n \ln \sigma + \sum_{i=1}^{n} \ln \Phi(-\varepsilon_i \lambda / \sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \varepsilon_i^2, \tag{2.4}$$

where $\varepsilon_i = y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta})$. The SFM can be estimated using the traditional maximum likelihood estimator (MLE). The benefit of this is that under the assumption of correct distributional specification of ε, the MLE is asymptotically efficient (i.e., consistent, asymptotically normal and its asymptotic variance reaches the Cramer-Rao lower bound). A further benefit is that a range of testing options are available. For instance, tests related to $\boldsymbol{\beta}$ can easily be undertaken using any of the classic trilogy of tests: Wald, Lagrange multiplier, or likelihood ratio. The ability to readily and directly conduct asymptotic inference is one of the major benefits of stochastic frontier analysis over DEA.[10]

2.2. **Alternative Specifications.** The half-normal assumption for the one-sided inefficiency term is almost without question the most commonly distribution for inefficiency in practice. This stems partly from posterity, partly from the closed form solution of the likelihood function, and partly from the availability of software to estimate the model for applied researchers. However, none of these reasons are sufficient for blind application of the half-normal density for inefficiency in the SFM.

2.2.1. *The Exponential Distribution.* The exponential assumption on inefficiency is also popular. The exponential density is

$$f(u) = \frac{1}{\sigma_u} e^{-u/\sigma_u}, \quad u \geq 0. \tag{2.5}$$

[10]This in no way suggests that inference cannot be undertaken when the DEA estimator is deployed; rather, the DEA estimator has an asymptotic distribution which is much more complicated that the MLE for the SFM, and so direct asymptotic inference is not available; bootstrapping techniques are required for many of the most popular DEA estimators (Simar & Wilson 2013, Simar & Wilson 2015).

For the normal-exponential distributional pair, the density of ε is

$$f(\varepsilon) = \frac{1}{\sigma_u}\Phi(-\varepsilon/\sigma_v - \sigma_v/\sigma_u)e^{\varepsilon/\sigma_u + \sigma_v^2/2\sigma_u^2}, \tag{2.6}$$

with likelihood function

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

Like the half-normal specification for u, the exponential specification monotonically decreases in u, suggesting that larger levels of inefficiency are less likely to occur than small levels of inefficiency. Both the half-normal and exponential specifications for inefficiency stem from what are known as single parameter distributions; single parameter distributions are the simplest distributions and an unfortunate (yet sometimes very convenient) property of them is that all of their moments depend on this single parameter, which can restrict the shape that the density can potentially take.[11]

2.2.2. *The Truncated Normal Distribution.* To allow more generality into the SFM, while guarding against distribution misspecification, a variety of one-sided distributions have been proposed for modeling u_i in the SFM. Stevenson (1980) proposed the truncated-normal distribution as a generalization of the half-normal distribution; whereas the half-normal distribution is the truncation of the $N(0, \sigma_u^2)$ at 0, the truncated-normal distribution is the truncation of the $N(\mu, \sigma_u^2)$ at 0. The pre-truncation mean parameter, μ, affords the SFM more flexibility in the shape of the distribution of inefficiency.

The truncated-normal density is

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma_u\Phi(\mu/\sigma_u)}e^{-\frac{(u-\mu)^2}{2\sigma_u^2}}, \quad u \geq 0. \tag{2.8}$$

This density reduces to the half-normal distribution when $\mu = 0$ and thus provides a generalization (more specifically a nesting structure), and an opportunity for inference on μ.

[11] See Parmeter & Kumbhakar (2014) for a more detailed analysis of the SFM with u distributed exponentially.

An intuitive appeal of deploying truncated-normal distribution in practice is that, unlike the half-normal and exponential densities, the truncated-normal density has mode at 0 only when $\mu \leq 0$, but otherwise has a mode at μ. When $\mu > 0$, the implication is that producers in a given market would tend to have inefficiency u_i near $\mu > 0$ rather than near 0. This connotation may be more realistic in some settings (e.g., the regulatory environment) than the half-normal assumption, where the probability of being less efficient is much larger than of being grossly inefficient.

For the normal-truncated-normal distributional pair, the density of ε 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

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

2.2.3. *Other Distributions.* Aside from the truncated-normal specification for the distribution of u, a variety of alternatives have been proposed throughout the literature. Greene (1980*a*, 1980*b*) and Stevenson (1980) both proposed a gamma distribution for inefficiency. The gamma distribution generalizes the exponential distribution in much the same way that the truncated-normal distribution nests the half-normal distribution. Ritter & Simar (1997) advocate against use of the gamma specification in practice noting that large samples were required to reliably estimate the parameters of the gamma distribution due to computational identification problem with the constant of the regression. Lee (1983) proposed a four parameter Pearson density for the specification of inefficiency; unfortunately, this distribution is intractable for applied work and until now has not appeared to gain popularity. Li (1996) proposed use of the uniform distribution for inefficiency noting an intriguing feature

of the subsequent composed error density: that it could be positively skewed.[12] Another specification for inefficiency appears in Carree (2002), who assumes the distribution of u follows a binomial specification; this allows the skewness of the composed error to be positive or negative. Gagnepain & Ivaldi (2002) specify inefficiency as being Beta distributed when inefficiency can be defined as a percentage (scaled between 0 and 1), while Almanidis, Qian & Sickles (2014) further generalize Stevenson's (1980) framework by assuming a doubly truncated-normal distribution for inefficiency. This distributional assumption also allows the convolved error term to be either positively or negatively skewed.

A common theme of all of the papers just mentioned is that they focus exclusively on the distribution of inefficiency inside the SFM. A recent literature has shed light on the features of $f(\varepsilon)$ for the SFM when both the density of v and the density of u are changed. 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. While these alternative distributional pairs do provide different insights into the behavior of the composed error, it remains to be seen if they will be regularly adopted in practice and whether they provide substantially different conclusions than the most frequently adopted distributional pairs (normal-half-normal for example); see Section 2.4 for more discussion on the perceived importance of distributional assumptions regarding estimation of the SFM.

It is important to note that the main idea behind the SFM is that nearly any pair of distributions can be used to model u and v. The advantage of the normal-half-normal pair that is dominant in the literature is that the likelihood function has an easily to evaluate expression. In general this should not be expected. More likely than not, for a range

[12]Prior to Li (1996) all of the previously proposed distributions always produced a composed error density that was theoretically negatively skewed. Note that if u is distributed uniformly over the interval $[0, b]$, inefficiency is equally likely to be either 0 or b.

of distributional assumptions the likelihood function will contain one or more intractable integrals, complicating estimation.[13]

2.2.4. *Alternative Estimation Approaches of the SFM.* Given the focus on inefficiency in the SFM, and the impact that the distributional assumption on u is likely to have on the MLE, studying the behavior of the SFM across a range of distributional assumptions is desirable. However, outside of a few specifications (half-normal, exponential, truncated-normal for u_i and normal for v_i), the composed error density will not have a likelihood function that lends itself for easy evaluation. In these cases it can be difficult to estimate all of the parameters of the SFM, but several approaches exist, ranging in complexity, to estimate the SFM when direct estimation of the likelihood function is not feasible. The simplest approach, dubbed corrected OLS (COLS) by Olson, Schmidt & Waldman (1980),[14] recognizes that OLS estimation of the SFM produces consistent estimates of the coefficients of the frontier function aside from the intercept. The intercept is biased downward by the expected level of industry inefficiency $E[u] = \sqrt{2/\pi}\sigma_u$. Olson et al.'s (1980) insight was that for a given pair of distributional assumptions (normal-exponential, say), the central moments of the OLS residuals could be used to construct consistent estimators of the parameters of the convolved error. Once these were estimated, expected inefficiency could be estimated and the bias in the intercept corrected. The beauty of COLS from the applied perspective is that OLS can be used and difficult likelihood functions do not have to be derived nor estimated.[15]

[13]Note that the likelihood function for the normal-half-normal pair is dependent upon the cdf of the normal distribution, $\Phi(\cdot)$ which contains an integral, but this can be quickly and easily evaluated across all modern software platforms.

[14]See also Greene (1980*a*, pg. 31-32). Richmond (1974) also proposed adjusting the intercept from OLS estimation, however, his model differs from that of Olson et al. (1980) by assuming the presence of inefficiency (which follows a gamma distribution) but no noise.

[15]There exists some confusion over the terminology COLS as it relates to another method, modified OLS (MOLS). Beginning with Winsten (1957), and discussed in Gabrielsen (1975) and Greene (1980*a*, pg. 32-34), MOLS shifts the estimated OLS production function until all of the observations lie on or below the 'frontier'. At issue is the appropriate name of these two techniques. Greene (2008) called the bounding approach COLS, crediting Lovell (1993, pg. 21) with the initial nomenclature, and referred to MOLS as the method which bias corrects the intercept based on a specific set of distributional assumptions. Further,

Several newer approaches exist as well. One that is becoming popular is maximum simulated likelihood (MSL) estimation (McFadden 1989). Greene (2003) used MSL estimation to estimate the parameters of the SFM for the normal-gamma convolution. The key to implementation of the SFM when the composed error does not produce a tractable likelihood is to notice that the integrals that commonly remain in the density (from integrating u out of the density) can be treated as expectations and evaluated by simulation rather than analytic optimization. Given that the distribution of u is assumed known (up to unknown parameters), for a given set of parameters, draws can be taken and the expectation evaluated, can then replace the integral. Optimization proceeds by searching over the parameter space until a global maximum is found.

An even more recent approach to evaluating intractable likelihoods is found in Tsionas (2012) who suggested estimation of the parameters of the SFM through the characteristic function of the composed error. The reason that this will work is that the characteristic function is a unique representation of a distribution (whether the density does or does not exist), and following from the convolution theorem, the characteristic function of two independent random variables (here v and u) added together is the product of the individual characteristic functions. The characteristic functions for all of the densities described above are known, and so, using the Fast Fourier Transform, the estimated characteristic function can be mapped to the underlying density, and subsequently, the likelihood function. Tsionas's (2012) method is somewhat computationally complicated, but it offers another avenue to estimate the SFM under alternative distributional assumptions on both v and u.

Kumbhakar & Lovell (2000, pg. 70-71) also adopted this terminology. However, given that Olson et al. (1980, pg. 69) explicitly used the terminology COLS, in our review we will adopt COLS to imply bias correction of the OLS intercept and MOLS as a procedure that shifts up (or down) the intercept to bound all of the data. The truth is both COLS and MOLS are the same in the sense that the OLS intercept is augmented, it is just in how each method corrects, or modifies, the intercept that is important. While we are departing from the more mainstream use of COLS and MOLS currently deployed, given the original use of COLS, coupled with myriad papers written by Peter Schmidt and coauthors that we discuss here, we will use the COLS acronym to imply a bias corrected intercept.

2.3. **Estimation of Individual Inefficiency.** Once the parameters of the SFM have been estimated, estimates of firm level productivity and efficiency can be recovered. Observation-specific estimates of inefficiency are one of the main benefits of the SFM relative to neo-classical models of production. Firms can be ranked according to estimated efficiency; the identity of under-performing firms as well as those who are deemed best practice can also be gleaned from the SFM. All of this information is useful in helping to design more efficient public policy or subsidy programs aimed at improving the market, for example, insulating consumers from the poor performance of heavily inefficient firms.

As a concrete illustration, consider firms operating electricity distribution networks that typically possess a natural local monopoly given that the construction of competing networks over the same terrain is prohibitively expensive.[16] 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 balance the heterogeneity that 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 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,

[16]The current literature is fairly rich on various examples of empirical values of SFA for the estimation and use of efficiency estimates in different fields of research. For example, in the context of electricity providers, see Knittel (2002), Hattori (2002), and Kuosmanen (2012); for banking efficiency, see Case, Ferrari & Zhao (2013) and references cited therein; for the analysis of the efficiency of national health care systems, see Greene (2004) and a review by Hollingsworth (2008); for analyzing efficiency in agriculture, see Bravo-Ureta & Rieger (1991), Battese & Coelli (1992, 1995), and Lien, Kumbhakar & Hardaker (2017), to mention just a few.

units. This is the essence of the determinants of inefficiency approach which we will discuss in Section 4. More directly, efficiency rankings are used in regulated industries such that regulators can set tougher future cost reduction targets for the more inefficient companies, in order to ensure that customers do not pay for the inefficiency of firms.

The only direct estimate coming from the normal-half-normal SFM is $\widehat{\sigma}_u^2$. This provides context regarding the shape of the half-normal distribution on u_i and the industry average efficiency $E[u]$, but not on the absolute level of inefficiency for a given firm. If we are only concerned with the average level of technical efficiency for the population, then this is all the information that is needed. Yet, if we want to know about a specific firm, then something else is required. The main approach to estimating firm level inefficiency is the conditional mean estimator of Jondrow, Lovell, Materov & Schmidt (1982), commonly known as the JLMS estimator. Their idea was to calculate the expected value of u_i conditional on the realization of composed error of the model, $\varepsilon_i \equiv v_i - u_i$, i.e., $E[u_i|\varepsilon_i]$.[17] This conditional mean of u_i given ε_i gives a point prediction of u_i. The composed error contains individual-specific information, and the conditional expectation is one measure of firm-specific inefficiency.

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

$$\mu_{*i} = \frac{-\varepsilon_i \sigma_u^2}{\sigma^2} \tag{2.11}$$

and

$$\sigma_*^2 = \frac{\sigma_v^2 \sigma_u^2}{\sigma^2}. \tag{2.12}$$

Given results on the mean of a truncated-normal density it follows that

$$E[u_i|\varepsilon_i] = \mu_{*i} + \frac{\sigma_* \phi\left(\frac{\mu_{*i}}{\sigma_*}\right)}{\Phi\left(\frac{\mu_{*i}}{\sigma_*}\right)}. \tag{2.13}$$

[17] Jondrow et al. (1982) also suggested an alternative estimator based on the conditional mode.

The individual estimates are then obtained by replacing the true parameters in (2.13) with MLE estimates from the SFM.

Another measure of interest is the Afriat-type level of technical efficiency, defined as $e^{-u_i} = Y_i/e^{m(\boldsymbol{x}_i)}e^{v_i} \in [0,1]$. This is useful in cases where output is measured in logarithmic form. Further, technical efficiency is bounded between 0 and 1, making it somewhat easier to interpret relative to a raw inefficiency score. Since e^{-u_i} is not directly observable, the idea of Jondrow et al. (1982) can be deployed here, and $E\left[e^{-u_i}|\varepsilon_i\right]$ can be calculated (Lee & Tyler 1978, Battese & Coelli 1988). For the normal-half-normal model, we have

$$E\left[e^{-u_i}|\varepsilon_i\right] = e^{\left(-\mu_{*i}+\frac{1}{2}\sigma_*^2\right)}\frac{\Phi\left(\frac{\mu_{*i}}{\sigma_*}-\sigma_*\right)}{\Phi\left(\frac{\mu_{*i}}{\sigma_*}\right)}, \tag{2.14}$$

where μ_{*i} and σ_* were defined in (2.11) and (2.12), respectively. Technical efficiency estimates are obtained by replacing the true parameters in (2.14) with MLE estimates from the SFM. When ranking efficiency scores, one should use estimates of $1 - E\left[u_i|\varepsilon_i\right]$, which is the first order approximation of (2.14). 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); see also Kumbhakar et al. (2015).

2.3.1. *Inference about the Presence of Inefficiency.* Having estimated the benchmark SFM, a natural hypothesis is whether inefficiency is even present. In this case the null hypothesis of interest is $H_0 : \sigma_u^2 = 0$ against $H_1 : \sigma_u^2 > 0$.[18] The direct way to test the H_0 is through a likelihood ratio test, keeping in mind that the unrestricted model is the assumed SFM and the restricted model is the linear regression model (or more specifically the normal regression model). There is a problem with implementation of this test however. Under H_0

[18]One could test if other moments of the distribution where 0 as well, but most of the SFMs parameterize the distribution of u with σ_u and so this seems the most natural.

σ_u^2 is restricted to lie on the boundary of the parameter space and this precludes direct use of a likelihood ratio test.

Coelli (1995) demonstrates that under H_0 the likelihood ratio statistic in this setting is a 50:50 mixture of a χ_1^2 distribution, the distribution of the ordinary likelihood ratio statistic if the parameter were not on the boundary of the parameter space, and a χ_0^2, known as the chi-bar-square distribution, $\bar{\chi}^2$, (Coelli 1995, Silvapulle & Sen 2005). This second piece is what captures the potential presence of the σ_u^2 parameter to lie on the boundary of the parameter space and creates a point mass in the asymptotic distribution of the likelihood ratio statistic.

Calculation of the test statistic itself is invariant to whether the parameter lies on the boundary under H_0. What does change is how one goes about calculating either the p-value or the critical value to assess the outcome of the test. In the case of the 50:50 mixture, the critical values are determined by looking at the 2α-level critical value from a χ_1^2 distribution. For example, whereas the critical value for a 5% significance level is 3.841 for χ_1^2, it is 2.706 for the 50:50 mixture. More specifically, Table 1 presents the critical values of both the χ_1^2 and the 50:50 mixture for a range of significance levels.

[Table 1 about here.]

An alternative type of test for the presence of inefficiency is based on the skewness of the residuals. A variety of tests for skewness exist, notably Ahmad & Li (1997), Kuosmanen & Fosgerau (2009) and Henderson & Parmeter (2015*b*). Henderson & Parmeter (2015*b*) proposed a bootstrap based version of Ahmad & Li's (1997) asymptotic test, noting that in finite samples the bootstrap version is likely to have superior performance. This test involves estimating the SFM using OLS and then testing whether the distribution of the OLS residuals is symmetric. Kuosmanen & Fosgerau's (2009) test of symmetry is also based on the bootstrap, but rather than focus on the estimated distribution of the OLS residuals, their test focuses exclusively on the skewness coefficient of the residuals. Both of these tests of

symmetry are appealing because they do not require parametric distributional assumptions and can be implemented after having estimated the SFM using OLS.

2.3.2. *Inference about the Distribution of Inefficiency.* It is important to recognize, despite the frequent misuse of terminology, that the JLMS or Battese-Coelli (or similar types) efficiency estimators are not estimators of u_i or e^{-u_i}, respectively and do not converge to them for $n \to \infty$. As $n \to \infty$, the new observations represent different firms each with their own level of inefficiency and noise (upon which JLMS conditions), rather than observations from the same firm. Even more importantly, the JLMS estimator was not intended to estimate unconditional inefficiency. The JLMS estimator is however, a consistent estimator for the expected level of inefficiency conditional on the particular realizations of ε.[19]

The JLMS efficiency scores can be used to provide a (limited) test of the distribution of inefficiency. The key insight to understand how a test can be constructed is that if the distributional assumptions are correct, then the distribution of $E[u_i|\varepsilon_i]$ is completely known. Hence a comparison of the distribution of $\widehat{E}[u_i|\varepsilon_i]$ to the true distribution of $E[u_i|\varepsilon_i]$ will shed light into the statistical validity of the assumed distributions for u and v. Wang & Schmidt (2009) derived the distribution of $E[u_i|\varepsilon_i]$ for the normal-half-normal SFM while Wang, Amsler & Schmidt (2011) proposed χ^2 and Komolgorov-Smirnov type test statistics against this distribution.[20]

We caution readers regarding a rejection with use of this test. A rejection does not necessarily imply the distributional assumption on u is incorrect, it could be that the normality distributional assumption on v or some other assumptions about the SFM (e.g., the parametric form of m) is violated, and this is leading to the rejection. Similarly, one must be careful in interpreting tests on the distribution of ε (or functionals of ε) when the distribution of

[19]The JLMS efficiency estimator is known as a shrinkage estimator; on average, it understates the efficiency level of a firm with small u_i while it overstates efficiency for a firm with large u_i.
[20]See also Lee (1983) for a different test based off of the Pearson distributional assumption for u.

v is also assumed to be normal. Alternative tests similar to Wang et al. (2011) could be formulated using the Laplace-exponential SFM of Horrace & Parmeter (2014).

2.3.3. *Predicting Inefficiency.* Aside from testing for the appropriate distribution of inefficiency, one should also test, or present uncertainty, as it pertains to an individual efficiency score. Each JLMS efficiency score is a prediction of inefficiency, and it is possible to calculate prediction intervals. Interestingly, few applied papers cover in depth uncertainty of estimated efficiency scores.

A prediction interval for $E[u_i|\varepsilon_i]$ was first derived by Taube (1988) and also appeared in Hjalmarsson, Kumbhakar & Heshmati (1996), Horrace & Schmidt (1996), and Bera & Sharma (1999) (see the discussion of this in Simar & Wilson 2010). The prediction interval is based on $f(u_i|\varepsilon_i)$. The lower (L_i) and upper (U_i) bounds for a $(1-\alpha)100\%$ prediction interval are

$$L_i = \mu_{*i} + \Phi^{-1}\left(1 - \left(1 - \frac{\alpha}{2}\right)\left[1 - \Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right)\right]\right)\sigma_*, \tag{2.15}$$

$$U_i = \mu_{*i} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\left[1 - \Phi\left(-\frac{\mu_{*i}}{\sigma_*}\right)\right]\right)\sigma_*, \tag{2.16}$$

where μ_{*i} and σ_* are defined in (2.11) and (2.12), respectively and replacing them with their MLE estimates will give estimated prediction intervals for $E[u_i|\varepsilon_i]$.

Wheat, Greene & Smith (2014) derived minimum width prediction intervals noting that the confidence interval studied in Horrace & Schmidt (1996) was based on a symmetric two sided interval. Given that the distribution of u_i conditional on ε_i is truncated (at 0) normal and asymmetric, this form of interval is not minimum width. Parmeter & Kumbhakar (2014) showed that depending upon the ratio of σ_u to σ_v, the difference in relative widths of Horrace & Schmidt's (1996) and Wheat et al.'s (2014) prediction intervals can be quite substantial. It is thus recommended to use the intervals provided by Wheat et al. (2014) as these are not based on symmetry. Note that although we could predict u and construct a prediction

interval, this information is not that useful for policy purposes unless there are some variables that affect inefficiency and such variables can be changed by a specific policy.

2.4. **Do Distributional Assumptions Even Matter?** An important empirical concern when using the SFM is the choice of distributional assumptions made for v and u. The distribution of v has almost universally been accepted as being normal in both applied and theoretical work (a recent exception is Horrace & Parmeter 2014); the distribution of u is more commonly debated, but relatively little work has been devoted to discerning the impact that alternative shapes of the distribution can have. Moreover, choice of u is often driven through available statistical software to implement the method rather than an underlying theoretical link between a model of productive inefficiency and the exact shape of the corresponding distribution.

A majority of applied papers studying productivity do not rigorously check differences in estimates, or perform inference, across different distributional assumptions. Greene (1990) is often cited as one of the first analyses to compare average inefficiency levels across several distributional specifications (half-normal, truncated-normal, exponential, and gamma), and he finds little difference in average inefficiency across 123 U.S. electric generation firms. Following Greene's (1990) investigation into the choice of distribution, Kumbhakar & Lovell (2000) calculated the rank correlations amongst the JLMS scores from these same four models, producing rank correlations as low as 0.75 and as high as 0.98.[21]

The intuition underlying these findings is that one's understanding of inefficiency, as measured through the JLMS score, is robust to distributional choices, at least from a ranking perspective. The reason for this can be found in the work of (Ondrich & Ruggiero 2001, p. 438) who have shown that the JLMS efficiency scores are monotonic in ε provided that

[21]In a limited 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; Horrace & Parmeter (2014) conducted a similar set of experiments and found essentially the same results.

the distribution of v is log-concave (which the normal distribution is). The implication here is that firm rankings can be obtained via the OLS residuals without the need of distributional assumptions whatsoever (Bera & Sharma 1999). Thus, in light of these insights, the important aspect of distributional choice for u is the impact that it has on the corresponding estimates of the production function; when these estimates are robust to distributional choice, so too will be the inefficiency rankings. Thus, if interest hinges on features of the frontier, then so long as inefficiency does not depend on conditional variables (see Section 4), one can effectively ignore the choice of distribution, as this only affects (usually but not substantially) the level of the estimated technology, but not its shape - which is what influences measures such as returns to scale and elasticities of substitution.

2.5. **Finite Sample Identification of Inefficiency.** An early analysis of the finite sample performance of the normal-half-normal SFM by Olson et al. (1980) uncovered an interesting phenomena, quite regularly the corrected OLS estimator would produce an estimate of $\sigma_u^2 \leq 0$. This was deemed a 'Type I' failure of the SFM; further Olson et al. (1980, pg. 70) noted that "It is also true that, in every case of Type I failure we encountered, the MLE estimate of $[\sigma_u^2]$ also turned out to equal zero. (This makes some sense, though we cannot prove analytically that it should happen.)" Waldman (1982) provided the analytic foundation behind this result, demonstrating that a stationary point of the log-likelihood function exists, and this stationary point is a local maximum when the sign of the skewness of residuals stemming from OLS estimation of the SFM is positive. This is broadly viewed as a deficiency of the SFM as an estimate of σ_u^2 of 0 is literally interpreted as a finding of no inefficiency.

However, this is an unfortunate interpretation because it is purely a finite sample issue. If in fact u is distributed half-normal, then as shown in Uekusa & Torii (1985), Coelli (1995), and Simar & Wilson (2010), as $n \to \infty$ the likelihood of drawing a random sample which will have positive skewness decreases, and the rate of this decrease is directly related to σ_u^2/σ_v^2; the larger this ratio the faster the decrease in the probability of observing a random

sample with positive skew.[22] The observance of OLS residuals with positive skew is, by and of itself, of no concern. What is concerning is that for an applied researcher whose focus is to study the efficiency level of firms, analysis of a sample where the residuals from the SFM have positive skewness leads to conclusions of all firms being efficient and this finding might be incongruent with either preconceptions about the industry or perceived publication standards when applying these methods. This has often led to various forms of respecification: using a different data set, trying an alternative functional form for the production function, or most likely, deploying different distributional assumptions regarding inefficiency.

As noted by Simar & Wilson (2010), none of these respecification approaches are appropriate or warranted. Again, Table 1 in Simar & Wilson (2010) evinces that even when everything about the SFM is correctly specified, positively skewed OLS residuals are still a regular occurrence. Their suggestion is to use special resampling techniques based off of bootstrapping to conduct inference on either overall inefficiency of the industry under study or specific firms. The finding of OLS residuals with positive skewness is commonly denoted the 'wrong skew problem', though it is not clear where this term initially originated. It is unfortunate that this term has crept into the lexicon of productivity analysis as there really is no problem at all, except for the problem of misinterpretation and mistreatment.

One reason why respecification is troubling is that classical statistical inference assumes that model specification is selected independently of estimation. When specification searches are conducted this introduces biases into the final parameter estimates. Further, there is the concern in published research that if the researcher did encounter positive skew, that this information is not provided to the reader. It is worth mentioning that not all SFMs are plagued by this issue. In fact, some distributional combinations will lead to identification of inefficiency regardless of the sign of the skewness of the OLS residuals. Examples include

[22]Note that the estimator of the skewness coefficient is distributed asymptotically standard normal so it is feasible to have either negative or positive skewness in any finite sample.

the normal-uniform SFM of Li (1996), the normal-Weibull SFM of Tsionas (2007), the normal-binomial SFM of Carree (2002), and the normal-doubly truncated SFM of Almanidis et al. (2014). Even more recently, Horrace & Parmeter (2014) demonstrated, in the style of Waldman (1982) that the log-likelihood function of the Laplace-exponential SFM is not dependent upon the sign of the skewness of the OLS residuals. This mainly stems from the fact that as $\sigma_u^2 \to 0$ that this model converges to a regression model with error term distributed Laplace, for which the MLE is the least absolute deviations (LAD) estimator.

Despite the history behind the impact of the sign of the skewness of the OLS residuals on the SFM, interest still abounds surrounding this issue. Recently, Hafner, Manner & Simar (2016) presented a generalized method which always ensures that the SFM can be identified, and that this model will converge to the traditional SFM model as $n \to \infty$ if the traditional SFM is correctly specified. Bonanno, De Giovanni & Domma (2017) introduced a generalized SFM which allows v to be distributed as a Type 1 generalized logistic which introduces asymmetry in v, coupled with allowing dependence between u and v. These two additional assumptions, similarly to Hafner et al. (2016), allow the parameters of the SFM to be identified regardless of the sign of the OLS residuals. Feng, Horrace & Wu (2015) describe a constrained MLE that uses the traditional normal-half-normal distributional pair, but imposes a penalty in estimation to combat the potential for positive skewness of the OLS residuals to lead to an estimate of σ_u^2 of 0. Finally, Horrace & Wright (2016) generalize the theory of Waldman (1982) by studying the SFM without explicit distributional assumptions. All told, this issue is one that still generates a substantial amount of interest in the academic community and it is one that is not likely to fade any time soon (see the discussion in Almanidis & Sickles 2011).

3. Handling Endogeneity in the SFM

A common assumption in the SFM is that $\boldsymbol{x}$ is either exogenous or independent of both u_i and v_i. If either of these conditions are violated then the MLE will be biased and most

likely inconsistent. Yet, it is not difficult to think of settings where endogeneity is likely to exist. For example, if shocks are observed before inputs are chosen, then producers may respond to good or bad shocks by adjusting inputs, leading to correlation between $\boldsymbol{x}$ and v. Alternatively, if managers know they are inefficient, they may use this information to guide their level of inputs, again, producing endogeneity. In a regression model, dealing with endogeneity is well understood. However, in the composed error setting, these methods cannot be simply transferred over, but require care in how they are implemented (Amsler, Prokhorov & Schmidt 2016).

To incorporate endogeneity into the SFM in (2.1), we set $m(\boldsymbol{x}_i;\boldsymbol{\beta}) = \beta_0 + \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2$ where $\boldsymbol{x}_1$ are our exogenous inputs, and $\boldsymbol{x}_2$ are the endogenous inputs, where endogeneity may arise through correlation of $\boldsymbol{x}_2$ with u, v or both. To deal with endogeneity we require instruments, $\boldsymbol{w}$, and identification necessitates that the dimension of $\boldsymbol{w}$ is at least as large as the dimension of $\boldsymbol{x}_2$. The natural assumption for valid instrumentation is that $\boldsymbol{w}$ is independent of both u and v. Our following discussion here will center on the distributional assumptions of ALS.

Why worry about endogeneity? Economic endogeneity means that the inputs in question are choice variables and chosen to optimize some objective function such as cost minimization or profit maximization. Statistical endogeneity arises from simultaneity, omitted variables, and measurement errors. For example if the omitted variable is managerial ability, which is part of inefficiency, inefficiency is likely to be correlated with inputs because managerial ability affects inputs. This is the Mundlak argument for why omitting a management quality variable (for us inefficiency) will cause biased parameter estimates. Endogeneity can also be caused by simultaneity meaning that more than one variable in the model are jointly determined.

One way to address the problem is to look at it as pure statistically and use instrumental variables. The other solution is economic, that is address the economic issue that is causing endogeneity. We consider first the statistical solution and then the economic solution. In

many applied settings it is not clear what researchers mean when they attempt to handle endogeneity inside the SFM. An excellent introduction into the myriad influences that endogeneity can have on the estimates stemming from the SFM can be found in Mutter, Greene, Spector, Rosko & Mukamel (2013). Mutter et al. (2013) used simulations designed around data based on the California nursing home industry to understand the impact of endogeneity of nursing home quality on inefficiency measurement.

3.1. **A Corrected Two Stage Least Squares Approach.** The simplest approach to accounting for endogeneity is to use a corrected two stage least squares (C2SLS) approach, similar to the common COLS approach that has been used to estimate the SFM. This method estimates the SFM using standard 2SLS with instruments $\boldsymbol{w}$. This produces consistent estimators for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ but not β_0, as this is obscured by the presence of $E[u]$ (to ensure that the residuals have mean zero). The second and third moments of the 2SLS residuals are then used to recover estimators of σ_v^2 and σ_u^2. Once $\widehat{\sigma}_u^2$ is determined, the intercept can be corrected by adding $\sqrt{\frac{2}{\pi}}\hat{\sigma}_u$.

This represents a simple avenue to account for endogeneity, and it does not require specifying how endogeneity enters the model, i.e. through correlation with v, with u or both. However, as with other corrected procedures based off of calculation of the second and third moments of the residuals, from Olson et al. (1980) and Waldman (1982), if the initial 2SLS residuals have positive skew (instead of negative), then σ_u^2 cannot be identified and its estimator is 0. Further, the standard errors from this approach need to be modified for the estimator of the intercept to account for the step-wise nature of the estimation.

3.2. **A Likelihood Approach.** The SFM with endogeneity has recently been studied by Kutlu (2010), Karakplan & Kutlu (2013), Tran & Tsionas (2013), and Amsler et al. (2016). Here we describe maximum likelihood estimation of the SFM under endogeneity. Our discussion here follows Amsler et al. (2016) as their derivation of the likelihood relies on a simple conditioning argument as opposed to the earlier work relying on the Cholesky decomposition.

While both approaches lead to the same likelihood function, the conditioning idea of Amsler et al. (2016) is simpler and more intuitive.

Consider the stochastic frontier system:

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i \tag{3.1}$$

$$\boldsymbol{x}_{2i} = \boldsymbol{w}_i\boldsymbol{\Gamma} + \boldsymbol{\eta}_i \tag{3.2}$$

where $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, $\boldsymbol{w}_i = (\boldsymbol{x}_{1i},\ \boldsymbol{q}_i)$ is the vector of instruments, $\boldsymbol{\eta}_i$ is uncorrelated with $\boldsymbol{w}_i$ and endogeneity of $\boldsymbol{x}_{2i}$ arises through $cov(\varepsilon_i, \boldsymbol{\eta}_i) \neq 0$. Here simultaneity bias (and the resulting inconsistency) exists because $\boldsymbol{\eta}_i$ is correlated with either v_i, u_i or both.

The following assumptions are used by Amsler et al. (2016): $u_i \sim N_+(0, \sigma_u^2)$, $m(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \beta_0 + \boldsymbol{x}_{1i}\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}\boldsymbol{\beta}_2$, and conditional on $\boldsymbol{w}_i$, $\psi_i = (v_i, \eta_i)' \sim N(\mathbf{0}, \Omega)$, where

$$\Omega = \begin{bmatrix} \sigma_v^2 & \Sigma_{v\eta} \\ \Sigma_{\eta v} & \Sigma_{\eta\eta}\boldsymbol{\eta}_i \end{bmatrix}.$$

Amsler et al. (2016) focused on the setting where u_i is independent of $\psi_i = \begin{bmatrix} v_i \\ \boldsymbol{\eta}_i \end{bmatrix}$. To derive the likelihood function, Amsler et al. (2016) condition on the instruments, $\boldsymbol{w}$. Doing this yields $f(y, \boldsymbol{x}_2|\boldsymbol{w}) = f(y|\boldsymbol{x}_2, \boldsymbol{w}) \cdot f(\boldsymbol{x}_2|\boldsymbol{w})$. With the density in this form, the log-likelihood follows suite: $\ln \mathcal{L} = \ln \mathcal{L}_1 + \ln \mathcal{L}_2$, where $\ln \mathcal{L}_1$ corresponds to $f(y|\boldsymbol{x}_2, \boldsymbol{w})$ and $\ln \mathcal{L}_2$ corresponds to $f(\boldsymbol{x}_2|\boldsymbol{w})$. These two components can be written as

$$\ln \mathcal{L}_1 = -(n/2)\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\tilde{\varepsilon}_i^2 + \sum_{i=1}^{n}\ln\left[\Phi\left(-\lambda_c\tilde{\varepsilon}_i/\sigma\right)\right]$$

$$\ln \mathcal{L}_2 = -(n/2)\ln|\Sigma_{\eta\eta}| - 0.5\sum_{i=1}^{n}\boldsymbol{\eta}_i'\Sigma_{\eta\eta}^{-1}\boldsymbol{\eta}_i,$$

where $\tilde{\varepsilon}_i = y_i - \beta_0 - \boldsymbol{x}_i\boldsymbol{\beta} - \mu_{ci}$, $\mu_{ci} = \Sigma_{v\eta}\Sigma_{\eta\eta}^{-1}\boldsymbol{\eta}_i$, $\sigma^2 = \sigma_v^2 + \sigma_u^2$, $\lambda_c = \sigma_u/\sigma_c$ and $\sigma_c^2 = \sigma_v^2 - \Sigma_{v\eta}\Sigma_{\eta\eta}^{-1}\Sigma_{\eta v}$. The subtraction of μ_{ci} in $\ln \mathcal{L}_1$ is an endogeneity correction while it should

be noted that $\ln \mathcal{L}_2$ is nothing more than the standard likelihood function of a multivariate normal regression model (as in (3.1)). Estimates of the model parameters $(\boldsymbol{\beta}, \sigma_v^2, \sigma_u^2, \boldsymbol{\Gamma}, \Sigma_{v\eta})$ and $\Sigma_{\eta\eta}$ can be obtained by maximizing the likelihood function $\ln \mathcal{L}$.

While direct estimation of the likelihood function is possible, a two-step approach is also available (Kutlu 2010). However, as pointed out by both Kutlu (2010) and Amsler et al. (2016), this two-step approach will have incorrect standard errors. Even though the two-step approach might be computationally simpler, it is, in general, different from full optimization of the likelihood function of Amsler et al. (2016). This is due to the fact that the two-step approach ignores the information provided by $\boldsymbol{\Gamma}$ and $\Sigma_{\eta\eta}$ in $\ln \mathcal{L}_1$. In general full optimization of the likelihood function is recommended as the standard errors (obtained in a usual manner from the inverse of the Fisher information matrix) are valid.[23]

3.3. **A Method of Moments Approach.** In insightful avenue to deal with endogeneity in the SFM that differs from the traditional corrected methods or maximum likelihood is proposed by Amsler et al. (2016), who used the work of Hansen, McDonald & Newey (2010). The idea is to use the first order conditions for maximization of the likelihood function under exogeneity:

$$E\left[\varepsilon_2^2/\sigma^2 - 1\right] = 0 \tag{3.3}$$

$$E\left[\frac{\varepsilon_i \phi_i}{1 - \Psi_i}\right] = 0 \tag{3.4}$$

$$E\left[\boldsymbol{x}_i \varepsilon_i/\sigma + \lambda \boldsymbol{x}_i \frac{\phi_i}{1 - \Phi_i}\right] = 0, \tag{3.5}$$

[23]Typically the standard errors can be obtained either through use of the outer product of gradients (OPG) or direct estimation of the Hessian matrix of the log-likelihood function. Given the nascency of these methods it has yet to be determined which of these two methods is more reliable in practice, though in other settings both tend to work well. One caveat for promoting the use of the OPG is that since this only requires calculation of the first derivatives, it can be more stable (and more likely to be invertible) than calculation of the Hessian. Also note that in finite samples, the different estimators of covariance of MLE estimator can give different numerical estimates, even suggesting different implications on the inference (reject or do not reject the null hypothesis). So, for small samples, it is often advised to check all feasible estimates whenever there is suspicion of ambiguity in the conclusions (e.g., when a hypothesis is rejected only at say around the 10% of significance level).

where $\phi_i = \phi(\frac{\lambda\varepsilon_i}{\sigma})$ and $\Phi_i = \Phi(\frac{\lambda\varepsilon_i}{\sigma})$. Note that these expectations are taken over $\boldsymbol{x}_i$ and y_i (and by default, ε_i) and solved for the parameters of the SFM.

The key here is that these first order conditions (one for σ^2, one for λ and the vector for $\boldsymbol{\beta}$) are valid under exogeneity and this implies that the maximum likelihood estimator is the generalized methods of moments estimator. Under endogeneity however, this relationship does not hold directly. But the seminal idea of Amsler et al. (2016) is that the first order conditions (3.3) and (3.4) are based on the distributional assumptions on v and u, not on the relationship of $\boldsymbol{x}$ with v and/or u. Thus, these moment conditions are valid whether $\boldsymbol{x}$ contains endogenous components or not. The only moment condition that needs to be adjusted is (3.5). In this case the first order needs to be taken with respect to $\boldsymbol{w}$, the exogenous variable, not $\boldsymbol{x}$. Doing so results in the following amended first order condition:

$$E\left[\boldsymbol{w}_i\varepsilon_i/\sigma + \lambda\boldsymbol{w}_i\frac{\phi_i}{1-\Phi_i}\right] = 0, \tag{3.6}$$

where ϕ_i and Φ_i are identical to those in (3.5). It is important to acknowledge that this moment condition is valid when ε_i and $\boldsymbol{w}_i$ are independent. This is a more stringent requirement than the typical regression setup with $E[\varepsilon_i|\boldsymbol{w}_i] = 0$. As with the C2SLS approach, the source of endogeneity for $\boldsymbol{x}_2$ does not need to be specified (through v and/or u).

3.4. **Estimation of Individual Inefficiency.** An interesting, and important finding from Amsler et al. (2016) is that when there is endogeneity, one can potentially improve estimation of inefficiency through the JLMS estimator. The traditional predictor of Jondrow et al. (1982) is $E(u_i|\varepsilon_i)$. However, more information is available when endogeneity is present, namely via $\boldsymbol{\eta}_i$. This calls for a modified JLMS estimator, $E(u_i|\varepsilon_i, \boldsymbol{\eta}_i)$. Note that even though it is assumed that u_i is independent from $\boldsymbol{\eta}_i$ (as in Amsler et al. 2016), because $\boldsymbol{\eta}_i$ is correlated with v_i, there is information that can be used to help predict u_i even after conditioning on ε_i.

Amsler et al. (2016) showed that $\boldsymbol{\eta}_i$ is independent of $(u_i, \tilde{\varepsilon}_i)$:

$$E(u_i|\varepsilon_i, \boldsymbol{\eta}_i) = E(u_i|\tilde{\varepsilon}_i, \boldsymbol{\eta}_i) = E(u_i|\tilde{\varepsilon}_i).$$

and that the distribution of u_i conditional on $\tilde{\varepsilon}_i = y_i - \beta_0 - \boldsymbol{x}_i\boldsymbol{\beta} - \mu_{ci}$ is $N_+(\mu_*, \sigma_*^2)$ with $\mu_* = -\sigma_u^2\tilde{\varepsilon}_i/\sigma^2$ and $\sigma_*^2 = \sigma_u^2\sigma_c^2/\sigma_2$, which is identical to the original JLMS estimator, except that σ_v^2 is replaced with σ_c^2 and $\tilde{\varepsilon}_i$ taking the place of ε_i. The modified JLMS estimator in the presence of endogeneity becomes $E(u_i|\varepsilon_i, \boldsymbol{\eta}_i) = \sigma_* \left(\frac{\phi(\xi_i)}{1-\Phi(\xi_i)} - \xi_i \right)$ with $\xi_i = \lambda\tilde{\varepsilon}_i/\sigma$. Note that $E(u_i|\varepsilon_i, \boldsymbol{\eta}_i)$ is a better predictor than $E(u_i|\varepsilon_i)$ because $\sigma_c^2 < \sigma_v^2$. The improvement in prediction follows from the textbook identity for variances, where for any random vector (X, Z), where X and Z are random sub-vectors, we have

$$var(X) = \underbrace{var[E(X|Z)]}_{Explained} + \underbrace{E(var[X|Z])}_{Unexplained}.$$

In this case, by conditioning on both ε_i and $\boldsymbol{\eta}_i$ the conditioning set is larger for and so it must hold that the unexplained portion of $E(u_i|\varepsilon_i, \boldsymbol{\eta}_i)$ is smaller than that of $E(u_i|\varepsilon_i)$. It then holds that there is less variation in $E(u_i|\varepsilon_i, \boldsymbol{\eta}_i)$ as a predictor than $E(u_i|\varepsilon_i)$, which is a good thing. While it is not obvious at first glance, one benefit of endogeneity is that researchers may be able to more accurately predict firm level inefficiency, though it comes at the expense of having to deal with endogeneity. This improvement in prediction may also be accompanied by narrower prediction intervals, however, this is not known as Amsler et al. (2016) did not study the prediction intervals.

3.5. **An Economic Approach to Deal with Endogeneity.** An alternative to developing valid instruments and correcting for endogeneity is to use what is known as a primal system approach, when inputs are endogenous (Kumbhakar et al. 2015, chapt. 8). This setup estimates the traditional SFM but appends the first order conditions stemming from cost minimization (one could alternatively attach profit maximization or return to the outlay conditions instead if this was a more representative behavior for the industry under study).

That is, if a producer minimizes costs[24]

$$\min \boldsymbol{p}'\boldsymbol{x}, \text{ s.t. } y = m(\boldsymbol{x};\boldsymbol{\beta}) + v - u, \tag{3.7}$$

for input prices $\boldsymbol{p}$, the first order conditions in this case are

$$\frac{m_j(\boldsymbol{x};\boldsymbol{\beta})}{m_1(\boldsymbol{x};\boldsymbol{\beta})} = \frac{p_j}{p_1}, \quad j = 2, \ldots, J, \tag{3.8}$$

where $m_j(\boldsymbol{x};\boldsymbol{\beta})$ is the partial derivative of $m(\boldsymbol{x};\boldsymbol{\beta})$ with respect to x_j. These first order conditions are exact, which usually does not arise in practice, rather, a stochastic term is added, which is designed to capture allocative inefficiency. That is, our empirical first order conditions are $\frac{m_j(\boldsymbol{x};\boldsymbol{\beta})}{m_1(\boldsymbol{x};\boldsymbol{\beta})} = \frac{p_j}{p_1}e^{\xi_j}$ for $j = 2, \ldots, J$ where e^{ξ_j} captures allocative inefficiency for the j^{th} input relative to input 1 (the choice of input to compare to is without loss of generality). The idea behind allocative inefficiency is that firms could be fully technically efficient, and still have room for improvement due to over or under use of inputs, relative to another input, given the price ratio. In general if firms are cost minimizers and one estimates a production function, the inputs will be endogenous as these are choice variables to the firm. Hence, a different approach is needed.

The primal system approach estimates the SFM as in (2.1) but also incorporates the information in the $J-1$ conditions in (3.8) with allocative inefficiency built in. Shephard's lemma in microeconomics dictates that the first order conditions are actually cost share information, when the logarithm of the production function is taken, the first derivatives represent the cost shares of the corresponding inputs,

$$\frac{m_j(\boldsymbol{x};\boldsymbol{\beta})}{m_1(\boldsymbol{x};\boldsymbol{\beta})} = \frac{\frac{\partial \ln m}{\partial \ln x_j}}{\frac{\partial \ln m}{\partial \ln x_1}} = \frac{s_j/x_j}{s_1/x_1}. \tag{3.9}$$

[24]It is possible to treat a subset of $\boldsymbol{x}$ as endogenous; i.e., $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$, where $\boldsymbol{x}_1$ is endogenous and $\boldsymbol{x}_2$ is exogenous.

When these are equated to the ratio of input prices, one obtains $\frac{s_j/x_j}{s_1/x_1} = \frac{p_j}{p_1}e^{\xi_j}$, which can be rearranged to yield $\frac{s_j}{s_1} = \frac{p_j x_j}{p_1 x_1}e^{\xi_j}$. Taking logarithms produces

$$\ln(s_j) - \ln(s_1) - \ln(p_j x_j) + \ln(p_1 x_1) = \xi_j. \tag{3.10}$$

If distributional assumptions are imposed on v, u and $\boldsymbol{\xi}$, the parameters of the production function can be estimated along with technical and allocative efficiency. An unfortunate consequence of the primal system approach is that only for quite specific assumptions on the production function are the input demand and cost functions analytically tractable (Cobb-Douglas being one). In these cases a more complicated process is required to determine the impact of technical and allocative inefficiency on costs (Kumbhakar & Wang 2006). See Kumbhakar (2011, 2013) for more detailed discussion of these types of primal system approaches to handle economic endogeneity across a ranges of settings.

4. Modeling Determinants of Inefficiency

Use of the SFM is exciting for productivity analysis because a prediction of firm level efficiency can be obtained. However, in the benchmark SFM, u_i is treated as completely random, and so nothing connects the level of inefficiency to variables which might serve as an explanation for the existence and the level of inefficiency. As the SFM has gained popularity in applied productivity analysis, it has become common to introduce variables outside the main production structure which influence output through their effect on inefficiency.[25]

As a concrete example, consider the study of productivity within the banking industry. A researcher may want to know whether a bank's level of efficiency 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, corporate governance practices, etc. Similarly, the

[25]Reifschneider & Stevenson (1991) used the term 'inefficiency explanatory variables', while others call them 'environmental variables', but it is now common to refer to these variables as 'determinants of inefficiency.' A variety of approaches have been proposed to model the determinants of inefficiency with the first pertaining to panel data models (Kumbhakar 1987, Battese & Coelli 1992) (see Section 5).

government might be interested in whether regulations (such as allowing banks to merge) improve banks' performance. To answer these questions, the relationship between efficiency and its potential determinants needs to be modeled and estimated.

Consider estimating what influences firm level inefficiency in the benchmark SFM. This model assumes that both v_i and u_i are homoskedastic. In a traditional linear regression, heteroskedasticity has no impact on the bias/consistency of the OLS estimator. However, if we were to allow σ_u^2 to depend on determinants of inefficiency, $\boldsymbol{z}$, then ignoring this will lead to, except in special settings, a biased and inconsistent estimator of the parameters of the SFM. Both Kumbhakar & Lovell (2000, Section 3.4) and Wang & Schmidt (2002) provide detailed accounts of the consequences of ignoring the presence of determinants of inefficiency in the SFM.

Recall from Section 2, that $E[u] = \sqrt{2/\pi}\sigma_u$. Now imagine ignoring the composed structure of ε and estimating the SFM via OLS. If it is the case that determinants of inefficiency are present, so that $\sigma_u^2 = \sigma_u^2(\boldsymbol{z})$, this omission leads to biased parameter estimates of the SFM given that the assumed model is

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

with $\varepsilon_i^* = \varepsilon_i - \sqrt{2/\pi}\sigma_u$, whereas the true model is

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

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. The reason that this issue presents itself is the fact that the mean of u, due to the truncation at 0, must depend on the variance. Thus, it is not possible to allow u to be heteroskedastic without the mean of u being a function of $\boldsymbol{z}$ as well. Notice here that we have specifically separated the impacts of $\boldsymbol{x}$ and $\boldsymbol{z}$ on output, with $\boldsymbol{x}$ capturing pure production and $\boldsymbol{z}$ capturing inefficiency. This is commonly known as the separability assumption. In some

settings this assumption does not have to be made, but in other settings it is a necessity for identification. See Parmeter & Zelenyuk (2016) for a more detailed discussion of the separability assumption. Our use of it here is more for expositional clarity.

Exactly how to model the influence of $\boldsymbol{z}$ on inefficiency is unknown and at various points in time practitioners have deployed a simpler, two step analysis to account for the presence of determinants of inefficiency. This approach constructs JLMS predictions in the first step, and then regresses these inefficiency estimates on $\boldsymbol{z}$ in the second step. Pitt & Lee (1981) were the first to implement this type of approach (in a panel data setting) and many others followed this two-step approach blindly (Ali & Flinn 1989, Kalirajan 1990, Bravo-Ureta & Rieger 1991). However, this route to modeling determinants of inefficiency has been met with criticism repeatedly, and for good reason.

As explained in Battese & Coelli (1995), the first stage model is misspecified if $\boldsymbol{z}$ is ignored. Further, Wang & Schmidt (2002) note that if $\boldsymbol{x}$ and $\boldsymbol{z}$ are correlated then an omitted variable bias exists in the the first step rendering the second step ineffectual. Even in the special case where $\boldsymbol{x}$ and $\boldsymbol{z}$ are uncorrelated, ignoring the dependence of u on $\boldsymbol{z}$ will lead to the estimated JLMS predictions in the first stage to have too little variation (see also Schmidt 2011) and, subsequently, the estimator in the second stage regression will be biased downward. Caudill & Ford (1993) provide Monte Carlo evidence on the impact that ignoring $\boldsymbol{z}$ on u has on the estimator of the parameters of the SFM while Wang & Schmidt (2002) provide a detailed analysis of the bias of the second stage parameter estimators.

As should be clear, the two stage approach to account for determinants of inefficiency in the SFM has no statistical foundation and is widely agreed upon to yield poor insights on the actual behavior of inefficiency, as such this approach should be strictly avoided; even with these criticisms of the two-step approach, one will occasionally happen across research that adopts this flawed two-step methodology.

While the two stage approach has undesirable statistical properties this does not mean that determinants of inefficiency cannot be accounted for. Quite the contrary. The preferred

approach to studying the exogenous influences on efficiency is a single-step procedure that explicitly accounts for $\boldsymbol{z}$.

4.1. **Proper Modeling of the Determinants of Inefficiency.** The first proper proposals to model $\boldsymbol{z}$ in the SFM are Kumbhakar, Ghosh & McGuckin (1991) and Reifschneider & Stevenson (1991), who used the normal truncated-normal SFM as the basis for estimation.[26] While their focus was on the normal truncated-normal SFM, the key insights hold for the normal-half-normal SFM, which is what we will base our discussion on here. The main idea is to specify σ_u^2 as a parametric function of $\boldsymbol{z}$.[27] Formally, their parameterization of σ_u^2 is

$$\sigma_u^2 = e^{\boldsymbol{z}'\boldsymbol{\delta}}, \tag{4.1}$$

The log-likelihood function of the heteroskedastic model is the same as in (2.4), except that we replace σ_u^2 with (4.1).[28] Here all of the model parameters are estimated simultaneously and once they are found, technical inefficiency can be computed using (2.13) or (2.14) with the appropriate form of σ_u^2 substituted into the expressions.

If u follows a half-normal distribution, with the σ_u^2 function depending upon $\boldsymbol{z}$ then the mean of u_i is

$$E[u_i|\boldsymbol{z}_i] = \sqrt{2/\pi}e^{\boldsymbol{z}_i'\boldsymbol{\delta}} = e^{\frac{1}{2}\ln(2/\pi)+\boldsymbol{z}_i'\boldsymbol{\delta}}. \tag{4.2}$$

Note that the $\frac{1}{2}\ln(2/\pi)$ term can be absorbed by the constant term in $\boldsymbol{z}_i'\boldsymbol{\delta}$. Therefore, by parameterizing σ_u^2, we allow $\boldsymbol{z}$ to affect the expected value of inefficiency. More importantly, however, is that the parameterization (4.1) produces maximum likelihood estimates of $\boldsymbol{\delta}$ which may not be very informative. This is because $E[u_i|\boldsymbol{z}_i]$ is nonlinear in $\boldsymbol{z}$, and therefore

[26]Caudill & Ford (1993), Huang & Liu (1994), Battese & Coelli (1995), Caudill, Ford & Gropper (1995), Hadri (1999), and Wang (2002) present alternative specifications as well.

[27]It is also possible to model σ_v^2 as a function of variables, but this poses fewer problems and we omit the details here. See Parmeter & Kumbhakar (2014) and Simar, Van Keilegom & Zelenyuk (2017) for more discussion.

[28]Actually, given the reparameterization of the log-likelihood function, the specification for σ_u implies a particular specification for both λ and σ.

the slope coefficients $\boldsymbol{\delta}$ are *not* the marginal effects of $\boldsymbol{z}$. For instance, assume the jth variable in $\boldsymbol{z}$ has an estimated coefficient of 0.5. This number itself tells us very little about the magnitude of the jth variable's (marginal) effect on the inefficiency, though it does tell us the direction of the effect on inefficiency. Also, the nonlinearity of the conditional mean of u, implies that for different levels of $\boldsymbol{z}$, there will be different expected levels of u. In these instances the marginal effect of $\boldsymbol{z}$ may be useful for empirical purposes.

For the given parameterization of the normal-half-normal SFM, the marginal effect of the jth variable of $\boldsymbol{z}_i$, z_{ji} on $E[u_i|\boldsymbol{z}_i]$ is

$$\frac{\partial E[u_i|\boldsymbol{z}_i]}{\partial z_j} = \delta_j \sqrt{2/\pi}\sigma_{u,i} \tag{4.3}$$

where $\sqrt{2/\pi}$ is approximately 0.80. It is clear that (4.3) also implies

$$\text{sign}\left(\frac{\partial E[u_i|\boldsymbol{z}_i]}{\partial z_j}\right) = \text{sign}(\delta_k) \tag{4.4}$$

so that the sign of the coefficient reveals the direction of impact of z_{ji} on $E[u_i|\boldsymbol{z}_i]$. This property does not always hold across distributional assumptions, for example in the normal-truncated-normal SFM the sign of the coefficient cannot be interpreted directly (Parmeter & Kumbhakar 2014). In general, only in one parameter families for the pdf of u (exponential, half-normal, etc.) does this correspondence hold; this suggests caution in directly interpreting the impact that a particular variable z_j has on inefficiency based purely on the sign of δ_j.

The nonlinear nature of the relationship of $E[u|\boldsymbol{z}]$ with $\boldsymbol{z}$ implies that for a sample of n observations we have n marginal effects for each variable. A concise statistic to present is the average partial effect (APE) on inefficiency or the partial effect of the average (PEA):

$$APE(z_j) = (\delta_j^u \sqrt{2/\pi}) \left(n^{-1} \sum_{i=1}^{n} \sigma_{u,i}\right) \tag{4.5}$$

$$PEA(z_j) = \delta_j^u \sqrt{2/\pi} e^{\bar{\boldsymbol{z}}_u' \boldsymbol{\delta}}. \tag{4.6}$$

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, but rather on average and ceteris paribus, i.e., holding other covariates fixed; for example, it could be that half of the sample has a very negative effect that is balanced by positive effects in the other half of the sample, thus getting nearly zero on average, which might misrepresent the phenomenon. It is also possible to standardize further by using elasticities, which will cancel out the $\sqrt{2/\pi}$ term, this occurs when the variables are measured in logarithms. It could also prove useful to present the estimates of these at either quartiles or at particular points of interest suggested by a particular empirical context (for example a specific regulation output target).

4.2. **Incorporating Determinants when u is Truncated-Normal.** As we have discussed earlier, the truncated-normal distribution offers greater flexibility to model an array of shapes of the true, but unknown, distribution of u. When determinants of inefficiency are present and one elects to assume the truncated-normal distribution, several additional modeling choices become available to the researcher. These additional choices are important because, as with the choice of distributional assumption, there is typically little guidance on how best to incorporate the determinants.

What do we mean? Consider again the truncated-normal density that would be assumed for u, when determinants of inefficiency are present:

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma_u(\boldsymbol{z};\boldsymbol{\delta}_1)\Phi\left(\mu(\boldsymbol{z};\boldsymbol{\delta}_2)/\sigma_u(\boldsymbol{z};\boldsymbol{\delta}_1)\right)} e^{-\frac{(u-\mu(\boldsymbol{z};\boldsymbol{\delta}_2))^2}{2\sigma_u(\boldsymbol{z};\boldsymbol{\delta}_1)^2}}, \quad u \geq 0. \tag{4.7}$$

In this case the impact of $\boldsymbol{z}$ on u can be modeled through the pre-truncation mean, μ and the pre-truncation standard deviation, σ_u. The issue with where to assume that $\boldsymbol{z}$ influences u is that modelling either parameter as a function of $\boldsymbol{z}$ impacts all of the moments of u, due to the truncation. Consider the conditional (on $\boldsymbol{z}$) mean of a truncated normal random

variable

$$E[u|\boldsymbol{z}] = \sigma_u(\boldsymbol{z};\boldsymbol{\delta}_1)\left[\frac{\mu(\boldsymbol{z};\boldsymbol{\delta}_2)}{\sigma_u(\boldsymbol{z};\boldsymbol{\delta}_1)} + \frac{\phi(\frac{\mu(\boldsymbol{z};\boldsymbol{\delta}_2)}{\sigma_u(\boldsymbol{z};\boldsymbol{\delta}_1)})}{\Phi(\frac{\mu(\boldsymbol{z};\boldsymbol{\delta}_2)}{\sigma_u(\boldsymbol{z};\boldsymbol{\delta}_1)})}\right]. \tag{4.8}$$

Regardless of whether σ_u or μ is constant, $\boldsymbol{z}$ still influences the mean of inefficiency unless both are constant. This is what makes the choice of where to incorporate $\boldsymbol{z}$ thorny when using the truncated normal distribution. Parametric specification of either σ_u or μ will allow for $\boldsymbol{z}$ to influence expected inefficiency, but in different manners, and, in nonlinear fashion. Given that μ can be positive or negative, it is common to model it in a linear fashion, i.e. $\mu(\boldsymbol{z};\boldsymbol{\delta}_2) = \boldsymbol{z}'\boldsymbol{\delta}_2$ and to model $\sigma_u(\boldsymbol{z};\boldsymbol{\delta}_1)$ as $e^{\boldsymbol{z}'\boldsymbol{\delta}_1}$, to ensure positivity of the pre-truncation standard deviation.

When we assume that u has the half-normal distribution our choice is easy because only a single parameter exists and it is clear where $\boldsymbol{z}$ enters. However, in the truncated-normal setup we could elect to have $\boldsymbol{z}$ enter only through the pre-truncation mean, only through the pre-truncation standard deviation, or both. In fact, various applied papers have used any of these three approaches. Kumbhakar et al. (1991) and Reifschneider & Stevenson (1991), modeled the impact of determinants of inefficiency through μ,[29] while Caudill & Ford (1993) incorporated determinants through σ_u.[30] Lastly, Wang (2002) modeled the determinants through both μ and σ_u. The benefit of modelling both pre-truncation parameters jointly as functions of $\boldsymbol{z}$ is that this leaves open little room for ambiguity and makes inference of where $\boldsymbol{z}$ belongs a viable option. The costs are that the model is more complex to estimate and may lead to identification problems, as raised in Ritter & Simar (1997). An alternative approach, which we discuss next, is to invoke a special assumption on the distribution that makes it more amenable to modelling the influence of determinants of inefficiency in the SFM.

[29] See also Huang & Liu (1994) and Battese & Coelli (1995) for early approaches following this strategy.
[30] Other early approaches that followed this route include Caudill et al. (1995) and Hadri (1999).

4.3. **The Scaling Property.** Many of the main proposals to incorporate determinants of inefficiency did so through the normal truncated-normal SFM. The two parameter nature of the truncated-normal distribution implies that determinants could influence the pre-truncation mean, μ, the pre-truncation variance, σ_u^2, or both. Further still, different variables could influence each parameter.

A popular simplification (Simar, Lovell & van den Eeckaut 1994, Wang & Schmidt 2002), which encapsulates the normal-half-normal SFM, is to assume that inefficiency behaves as

$$u_i \sim g(\boldsymbol{z}_i;\, \boldsymbol{\delta})\, u_i^*, \tag{4.9}$$

where $g(\cdot) \geq 0$ is a function of the exogenous variables while $u_i^* \geq 0$ is a random variable. This behavior is known as the scaling property. Single parameter distributions, such as the half-normal and the exponential, automatically possess this property, but more flexible distributions, such as truncated-normal or gamma, can have this property imposed. The key feature of the scaling property is that u_i^* does not depend on $\boldsymbol{z}_i$ in any fashion; u_i^* is known as base inefficiency (Wang & Schmidt 2002, Alvarez, Amsler, Orea & Schmidt 2006).

When a distribution possesses the scaling property the *shape* of the distribution of u_i is the same for all firms, which can be viewed as an attractive feature. The scaling function, $g(\cdot)$, expands or contracts the horizontal axis so that the scale of the distribution of u_i changes while preserving the underlying shape of the distribution. In comparison, the normal truncated-normal SFM models allow different scalings for each u_i, so that for some firms the distribution of inefficiency is close to a normal (if the pre-truncation mean is large), while for other firms the distribution of inefficiency is the extreme right tail of a normal with a mode of zero (if the pre-truncation mean is negative). 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 of $\boldsymbol{\delta}$ when $g(\boldsymbol{z}_i, \boldsymbol{\delta}) = e^{\boldsymbol{z}_i'\boldsymbol{\delta}}$,

$$\frac{\partial \ln E[u_i|\boldsymbol{z}]}{\partial z_j} = \delta_j. \tag{4.10}$$

That is, δ_j is the semi-elasticity (or elasticity if $\boldsymbol{z}$ is already measured on the logarithmic scale) of expected inefficiency with respect to the jth element of $\boldsymbol{z}$, and more importantly, this interpretation is distinct from any distributional assumption placed on u^*. An interpretation of this ilk is generally not available in other model specifications. Further, the sign of the elements of $\boldsymbol{\delta}$ can be directly interpreted.

The scaling property provides an attractive economic interpretation as well. u^* can be interpreted as a benchmark level of inefficiency of the firm (Alvarez et al. 2006). The scaling function then allows a firm to exploit (or fail to exploit) these talents through other variables, $\boldsymbol{z}$, which might include experience of the plant manager, the operating environment of the firm, or regulatory restrictions.

The scaling property is not a fundamental feature, rather, as with the choice of distribution on u, it is an assumption on the features of the inefficiency distribution. As such it can be tested against models that do not possess this property for the inefficiency distribution. As it currently stands, all tests of the scaling property hinge on a given distributional assumption, for example, estimating the normal truncated-normal SFM and then estimating a restricted version of the same model, but imposing the scaling property. An important avenue for future research is the development of a test (or tests) that do not require specific distributional assumptions.

4.4. **Estimation Without Imposing Distributional Assumptions.** In settings where the researcher is comfortable with imposing the scaling property on the distribution of inefficiency, the SFM can be estimated without distributional assumptions. This is perhaps

the key benefit of invoking the scaling property. To understand how it is possible to estimate the SFM without distributional assumptions we expound on the discussion of Simar et al. (1994), Wang & Schmidt (2002), and Alvarez et al. (2006). The SFM with the scaling property can be written as[31]

$$y_i = m(\boldsymbol{x}_i;\boldsymbol{\beta}) + v_i - e^{\boldsymbol{z}_i'\boldsymbol{\delta}}u_i^*. \tag{4.11}$$

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

$$E[y|\boldsymbol{x},\boldsymbol{z}] = \boldsymbol{x}'\boldsymbol{\beta} - e^{\boldsymbol{z}'\boldsymbol{\delta}}\mu^* \tag{4.12}$$

where $\mu^* = E[u^*]$ and $E[v|\boldsymbol{x},\boldsymbol{z}] = 0$. The SFM is then

$$y_i = m(\boldsymbol{x}_i;\boldsymbol{\beta}) - e^{\boldsymbol{z}_i'\boldsymbol{\delta}}\mu^* + v_i - e^{\boldsymbol{z}_i'\boldsymbol{\delta}}(u_i - \mu^*) = m(\boldsymbol{x}_i;\boldsymbol{\beta}) - e^{\boldsymbol{z}_i'\boldsymbol{\delta}}\mu^* + \varepsilon_i^*, \tag{4.13}$$

with $\varepsilon_i^* = v_i - e^{\boldsymbol{z}_i'\boldsymbol{\delta}}(u_i - \mu^*)$, which, for a given parameterization of $m(\boldsymbol{x}_i;\boldsymbol{\beta})$, can be estimated using nonlinear least squares (NLS) as

$$\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\delta}},\widehat{\mu^*}\right) = \min_{\boldsymbol{\beta},\boldsymbol{\delta},\mu^*} n^{-1}\sum_{i=1}^{n}\left[y_i - m(\boldsymbol{x}_i;\boldsymbol{\beta}) + \mu^* e^{\boldsymbol{z}_i'\boldsymbol{\delta}}\right]^2. \tag{4.14}$$

The elegance of invoking the scaling property is that the SFM can be estimated in a distribution free manner via NLS; the need for NLS stems from the fact that the scaling function must be positive and if it was specified as linear this would be inconsistent with theoretical requirements on the variance of the distribution.

Direct NLS will produce a consistent estimator of all of the terms of the SFM. However, the error term ε_i^* is heteroskedastic,

$$var(\varepsilon_i^*|\boldsymbol{x}_i,\boldsymbol{z}_i) = \sigma_v^2 + \sigma_u^{2*}e^{2\boldsymbol{z}_i'\boldsymbol{\delta}},$$

[31]Note here that we are making the implicit assumption that $\boldsymbol{z}$ is different from $\boldsymbol{x}$. The nonlinearity of the scaling function does allow $\boldsymbol{z}$ and $\boldsymbol{x}$ to overlap however.

where $\sigma_v^2 = var(v_i)$ and $\sigma_u^{2*} = var(u_i^*)$. As such, a generalized NLS estimator would be called for to produce an efficient estimator (as similar to the MLE). Unfortunately, a generalized NLS algorithm hinges on distributional assumptions to appropriately separate σ_v^2 and σ_u^{2*}. An alternative, which allows valid inference to be undertaken, is to compute heteroskedasticity robust standard errors for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ (White 1980).

An interesting extension of this idea was recently proposed by Paul & Shankar (2017) for the setting where u_i has already been converted into technical efficiency. In this case the level of inefficiency must be bound between 0 and 1. To account for this Paul & Shankar (2017) model the impact of $\boldsymbol{z}$ on the level of inefficiency through a probit function. Again, given the nonlinear nature of the probit function, this necessitates use of NLS if one wishes to eschew distributional assumptions.

With the wide range of statistical software that can quickly implement a NLS problem, it is perhaps surprising that this avenue has not been exploited in applied research. Certainly the scaling property is an assumption that requires judicious justification, but not more so than distributional assumptions imposed on the composed error structure of the SFM. It is also possible in this nonlinear setup that the calculation of expected firm efficiency can be done without requiring distributional assumptions, leading to the potential for more robust conclusions regarding observation specific inefficiency. It is also possible to estimate the SFM in (4.13) without imposing assumptions on the scaling function, an issue we will discuss in Section 6.

Currently no test of the scaling property exists without enforcing distributional assumptions. Alvarez et al. (2006) proposed standard tests of the scaling property by using the nesting structure of the normal-truncated-normal distributional pair against the normal-half-normal distributional pair. Unfortunately this testing facility requires distributional assumptions on both v_i and u_i. A test of the statistical significance of the determinants of inefficiency, using the NLS framework just described is available (Kim & Schmidt 2008).

Under $H_0 : \boldsymbol{\delta} = 0$ it follows that

$$y_i = m(\boldsymbol{x}_i; \boldsymbol{\beta}) - \mu^* e^{\boldsymbol{z}_i'\boldsymbol{\delta}} + \varepsilon_i^* = m^*(\boldsymbol{x}_i; \boldsymbol{\beta}) - \mu^*(1 - e^{\boldsymbol{z}_i'\boldsymbol{\delta}}) + \varepsilon_i^*,$$

where $m^*(\boldsymbol{x}_i; \boldsymbol{\beta}) = m(\boldsymbol{x}_i; \boldsymbol{\beta}) + c$ for a constant c. That is, one can only identify μ if at least one element of $\boldsymbol{\delta}$ is nonzero; note that under $H_0 : 1 - e^{\boldsymbol{z}_i'\boldsymbol{\delta}} = 0$, μ^* cannot be separately identified. This lack of identification creates issues for inference under the null hypothesis, and invalidates the common asymptotic behavior of Wald and likelihood ratio tests. The solution, which Kim & Schmidt (2008) proposed to avoid this problem, is to use the Lagrange Multiplier (LM) test which involves estimation imposing the null hypothesis. A novel insight of Kim & Schmidt (2008) is that the LM test they proposed has power in directions where the scaling property does not hold. This is due to the fact that the model being tested H_0 is indifferent to 'how' inefficiency enters the model. Thus, while an explicit test of the scaling property without requiring distributional assumptions would be a useful tool, the Kim & Schmidt (2008) LM test is likely to be sufficient.

The LM test is based on the derivative of the NLS criterion function in (4.14) with respect to $\boldsymbol{\delta}$, evaluated at the restricted estimates ($\boldsymbol{\delta} = 0$):

$$\frac{2}{n} \sum_{i=1}^{n} \left(y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta}) + \mu^*\right) \left(\mu^* \boldsymbol{z}_i\right). \tag{4.15}$$

The test statistic is designed to determine how close the derivative of the NLS objection 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. The reason that distributional assumptions are not needed for this test to work properly is that this test is identical to an F-test, and F-tests are invariant to the scale of the covariates (Kim & Schmidt 2008). Thus, one can simply set $\mu^* = 1$ and use NLS to regress y on $(\boldsymbol{x}, \boldsymbol{z})$ and test the significance of $\boldsymbol{\delta}$.

4.5. **Estimation When Determinants of Efficiency and Endogeneity Are Present.** Quite recently, attention has focused on estimation of the SFM when some of the determinants of inefficiency may be endogenous (Amsler, Prokhorov & Schmidt 2017, Latruffe, Bravo-Ureta, Carpentier, Desjeux & Moreira 2017). These models can be estimated using traditional instrumental variables methods. However, given that the determinants of inefficiency enter the model nonlinearly, nonlinear methods are required. To begin, we consider the model of Amsler, Prokhorov & Schmidt (2017),

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i - u_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i - u_i^* e^{\boldsymbol{z}_i'\boldsymbol{\delta}}, \tag{4.16}$$

where the scaling property has been invoked. The covariates $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are partitioned as

$$\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{x}_{2i} \end{bmatrix}, \quad \boldsymbol{z}_i = \begin{bmatrix} \boldsymbol{z}_{1i} \\ \boldsymbol{z}_{2i} \end{bmatrix},$$

where $\boldsymbol{x}_{1i}$ and $\boldsymbol{z}_{1i}$ are exogenous and $\boldsymbol{x}_{2i}$ and $\boldsymbol{z}_{2i}$ are endogenous. The set of instruments used to combat endogeneity are defined as

$$\boldsymbol{w}_i = \begin{bmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{z}_{1i} \\ \boldsymbol{q}_i \end{bmatrix},$$

where $\boldsymbol{q}_i$ are the traditional outside instruments. Identification of all the parameters requires that the dimension of $\boldsymbol{q}$ be at least as large as the dimension of $\boldsymbol{x}_2$ plus the dimension of $\boldsymbol{z}_2$ (the rank condition).

In the model of Amsler, Prokhorov & Schmidt (2017), endogeneity arises through correlation between a variable in the model ($\boldsymbol{x}_2$ and/or $\boldsymbol{z}_2$) and noise, v. That is, both $\boldsymbol{x}$ and $\boldsymbol{z}$ are assumed to be independent of basic inefficiency u^*. Given that $E[u_i]$ is not constant, the COLS approach to deal with endogeneity proposed by Amsler et al. (2016) cannot be used here. To develop an appropriate estimator, add and subtract the mean of inefficiency

to produce a composed error term that has mean 0,

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} - \mu^* e^{\boldsymbol{z}_i'\boldsymbol{\delta}} + v_i - (u_i^* - \mu^*)e^{\boldsymbol{z}_i'\boldsymbol{\delta}}. \tag{4.17}$$

Proper estimation through instrumental variables requires that the moment condition

$$E\left[v_i - (u_i^* - \mu^*)e^{\boldsymbol{z}_i'\boldsymbol{\delta}} | \boldsymbol{w}_i\right] = 0. \tag{4.18}$$

The nonlinearity of these moment conditions would necessitate use of nonlinear two stage least squares (NL2SLS) (Amemiya 1974).

Latruffe et al. (2017) have a similar setup as Amsler, Prokhorov & Schmidt (2017), using the model in (4.16), but develop a four step estimator for the parameters; additionally, only $\boldsymbol{x}_2$ is treated as endogenous. Latruffe et al.'s (2017) approach is based off of Chamberlain (1987) on the construction of efficient moment conditions. The vector of instruments proposed in Latruffe et al. (2017) is defined as

$$\boldsymbol{w}_i(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \begin{bmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{q}_i'\boldsymbol{\gamma} \\ \boldsymbol{z}_i e^{\boldsymbol{z}_i'\boldsymbol{\delta}} \end{bmatrix}, \tag{4.19}$$

where $\boldsymbol{q}_i'\boldsymbol{\gamma}$ captures the linear projection of $\boldsymbol{x}_2$ on the external instruments $\boldsymbol{q}$. The four-stage estimator is defined as

Step 1: Regress $\boldsymbol{x}_2$ on $\boldsymbol{q}$ to estimate $\boldsymbol{\gamma}$. Denote the OLS estimator of $\boldsymbol{\gamma}$ as $\widehat{\boldsymbol{\gamma}}$.

Step 2: Use NLS to estimate the SFM in (4.16). Denote the NLS estimates of $(\boldsymbol{\beta}, \boldsymbol{\delta})$ as $(\ddot{\boldsymbol{\beta}}, \ddot{\boldsymbol{\delta}})$. Use the NLS estimate of $\boldsymbol{\delta}$ and the OLS estimate of $\boldsymbol{\gamma}$ in Step 1 to construct the instruments $\boldsymbol{w}_i(\widehat{\boldsymbol{\gamma}}, \ddot{\boldsymbol{\delta}})$.

Step 3: Using the estimated instrument vector $\boldsymbol{w}_i(\widehat{\boldsymbol{\gamma}}, \ddot{\boldsymbol{\delta}})$, calculate the NL2SLS estimator of $(\boldsymbol{\beta}, \boldsymbol{\delta})$ as $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\delta}})$. Use the NL2SLS estimate of $\boldsymbol{\delta}$ and the OLS estimate of $\boldsymbol{\gamma}$ in Step 1 to construct the instruments $\boldsymbol{w}_i(\widehat{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\delta}})$.

Step 4: Using the estimated instrument vector $\boldsymbol{w}_i(\widehat{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\delta}})$, calculate the NL2SLS estimator of $(\boldsymbol{\beta}, \boldsymbol{\delta})$ as $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\delta}})$.

This multi-step estimator is necessary in the context of efficient moments because the actual set of instruments is not used directly, rather $\boldsymbol{w}_i(\boldsymbol{\gamma}, \boldsymbol{\delta})$ is used, and this instrument vector requires estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. The first two steps of the algorithm are designed to construct estimates of these two unknown parameter vectors. The third step then is designed to construct a consistent estimator of $\boldsymbol{w}_i(\boldsymbol{\gamma}, \boldsymbol{\delta})$, which is not done in Step 2 as given that the endogeneity of $\boldsymbol{x}_2$ is ignored (note that NLS is used as opposed to NL2SLS). The iteration from Step 2 to Step 3 does produce a consistent estimator of $\boldsymbol{w}_i(\boldsymbol{\gamma}, \boldsymbol{\delta})$, and as such, Step 4 produces consistent estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$). While Latruffe et al. (2017) proposed a set of efficient moment conditions to handle endogeneity, the model of Amsler, Prokhorov & Schmidt (2017) is more general because it can handle endogeneity in the determinants of inefficiency as well.

5. Panel Data

Our current discussion of the SFM has focused on having access to cross-sectional data. When repeated observations of firms are available, then more useful information about inefficiency (and often with more flexibility) can be extracted and a range of panel data SFMs are available to the applied researcher. Here we highlight some of the most prominent models. The advantage of panel data is that more information on inefficiency and productivity can be parsed, and in particular, shed light on changes in efficiency or productivity, which differs from a cross-sectional setting, which can only provide a static portrayal of inefficiency.

While Pitt & Lee (1981) were the first to consider extending the cross sectional SFM to the panel data setting, it was Schmidt & Sickles (1984) who brought prominence to the use of models tailored exclusively to panel data. They raise three problems with cross-sectional models that are used to measure inefficiency and productivity: First, if the MLE is used to estimate the parameters of the SFM and inefficiency through JLMS, everything is contingent

on distributional assumptions for both noise and inefficiency; Second, technical inefficiency is assumed to be independent of the regressor(s).[32]; Third, the JLMS estimator is not a consistent estimator of u, as $E[u|\varepsilon]$ never approaches u as the number of cross-sectional units approaches infinity ($n \to \infty$). Access to panel data can, to varying degrees, mitigate all of these issues. However, with panel data comes a range of additional assumptions that the researcher needs to carefully consider before proceeding.

To begin, consider the benchmark linear panel data regression model:

$$y_{it} = m(\boldsymbol{x}_{it}; \boldsymbol{\beta}) + c_i + v_{it}. \tag{5.1}$$

Aside from the indexing of our data by individual, i, and time, t, we have the presence of firm specific heterogeneity, c_i. The common dilemma facing application of the linear panel data regression model is how to treat the relationship between c_i and $\boldsymbol{x}_{it}$. Under the fixed effects (FE) framework (Wooldridge 2010), $\boldsymbol{x}_{it}$ is allowed to be correlated with c_i and the parameters of the model can be estimated consistently using the within transformation (Baltagi 2013). Under the random effects (RE) framework, $\boldsymbol{x}_{it}$ and c_i are required to be uncorrelated, leading to OLS being a consistent estimator, but ultimately inefficient given that the variance-covariance matrix of the composed error term $c + v$ is no longer diagonal. A feasible generalized least squares approach is available to obtain asymptotically efficient estimators of the parameters of the regression model in this case.

Now, to think about where inefficiency enters the model in (5.1), we must characterize the nature of inefficiency. If inefficiency is assumed to be constant over time, then it is likely that c_i might be augmented to also capture inefficiency. If inefficiency is time-varying then we could include a second, one-sided error term to be convolved with v_{it} in (5.1), in much the same way we did in the benchmark SFM. Or, it could be that inefficiency is composed of both a time-invariant component and a time-varying component. All told, the general panel

[32]If firms maximize profit, and inefficiency is known to the firm, then this assumption is unlikely to be true as firms may adjust their inputs to account for inefficiency (e.g., see Mundlak 1961).

data SFM is

$$y_{it} = m(\boldsymbol{x}_{it};\boldsymbol{\beta}) + c_i - \eta_i + v_{it} - u_{it} = m(\boldsymbol{x}_{it};\boldsymbol{\beta}) + \alpha_i + \varepsilon_{it}, \tag{5.2}$$

where $\alpha_i = c_i - \eta_i$ with c_i capturing time-invariant heterogeneity and η_i encapsulating time-invariant inefficiency while $\varepsilon_{it} = v_{it} - u_{it}$ with u_{it} representing time-varying inefficiency. The panel data SFM looks identical to the panel data regression model in (5.1), except that, due to $u_{it} > 0$, ε_{it} no longer has mean zero, and α_i no longer solely captures individual specific heterogeneity. Early approaches that studied inefficiency in panel data settings placed restrictions on how inefficiency entered the panel data SFM. As time progressed, fewer assumptions were made, especially as more advanced econometric techniques were exploited.

5.1. **Time-invariant Technical Inefficiency Models.** When inefficiency in the panel data SFM is assumed to be time-invariant, it is possible to estimate the model without the need for distributional assumptions. To begin, we assume that u_{it} does not exist in (5.2) and all time-invariant unobserved heterogeneity is inefficiency, $\alpha_i = \eta_i$. With these restrictions, the panel data SFM is written as

$$y_{it} = m(\boldsymbol{x}_{it};\boldsymbol{\beta}) - \eta_i + v_{it}; i = 1,\ldots,n; t = 1,\ldots,T. \tag{5.3}$$

This model is termed the time-invariant SFM. Aside from the one-sided nature of η_i, this model can be estimated with standard panel data regression techniques, once an assumption on the underlying statistical relationship (either the FE or RE framework) between $\boldsymbol{x}_{it}$ and η_i is made. Which framework to deploy depends upon the relationship that one assumes exists between the covariates of the model and firm level inefficiency. Under the FE framework correlation is allowed between $\boldsymbol{x}_{it}$ and η_i, whereas under the RE framework no correlation is permitted between $\boldsymbol{x}_{it}$ and η_i. Regardless of which framework is deemed appropriate, neither requires distributional assumptions for η or v. This freedom from imposing a parametric

assumption on the distribution of η_i (i.e., we have some statistical requirements on the distribution but do not require a precise parametric form) has led to the time-invariant SFM being referred to as a distribution free approach (Schmidt & Sickles 1984). To estimate the time-invariant SFM respecting the one-sided nature of η_i, a simple transformation is needed to interpret the individual effect as time-invariant inefficiency as opposed to pure firm heterogeneity. One major limitation of the time-invariant SFM is that separate identification of inefficiency and individual heterogeneity is not considered. Additionally, the production technology is assumed to be time constant, which may be a further limitation depending upon the time dimension one has access to.

We briefly discuss how to estimate the time-invariant SFM under the FE framework, which was first proposed by Schmidt & Sickles (1984). For ease of exposition, we assume $m(\cdot)$ is linear in $\boldsymbol{x}_{it}$. The time-invariant SFM is

$$\begin{aligned} y_{it} =& \beta_0 + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - \eta_i \\ =& (\beta_0 - \eta_i) + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} \end{aligned} \tag{5.4}$$

$$= c_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} \tag{5.5}$$

where $c_i \equiv \beta_0 - \eta_i$. Under the FE framework, η_i and thus α_i, $i = 1, \ldots, n$ are allowed to have arbitrary correlation with $\boldsymbol{x}_{it}$.

Given the similarity of the time-invariant SFM and a traditional panel data regression model, Schmidt & Sickles (1984) used standard estimation methods to estimate the parameters of the model, namely, within estimation. 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 c_i. OLS can then be used to estimate the transformed model, essentially regressing transformed y on transformed $\boldsymbol{x}$. The OLS estimator with the transformed data, $\widehat{\boldsymbol{\beta}}$, is a consistent estimator for $\boldsymbol{\beta}$. An estimator of c_i, $\hat{c}_i$, is constructed from the mean of the residuals for each cross sectional

unit, i.e., $\hat{c}_i = \bar{y}_{i\cdot} - \boldsymbol{x}'_{i\cdot}\widehat{\boldsymbol{\beta}}$, but it is biased, because $\eta_i > 0 \ \forall i$. A simple transformation will produce a consistent estimator of η_i. Once $\hat{c}_i$ is determined, $\hat{\eta}_i$ is estimated as (Schmidt & Sickles 1984):

$$\hat{\eta}_i = \max_i\{\hat{c}_i\} - \hat{c}_i \geq 0, \quad i = 1, \ldots, n. \tag{5.6}$$

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

$$\widehat{TE}_i = e^{-\hat{\eta}_i}, \quad i = 1, \ldots, n. \tag{5.7}$$

Operating under the FE framework may be more appropriate for empirical applications in which inefficiency is believed to be correlated with the inputs used. A disadvantage of using the time-invariant SFM under the FE framework is that no other time-invariant variables can be included. For example, when T is short (say only a few periods), the gender of a plant manager, or ownership status of the firm (if it does not change over the time frame), and so effectively, their influence (if present in reality) will be accumulated in (and distort) the estimates of inefficiency.

In settings where time-invariant variables are expected to be relevant regressors in the production model, an alternative is to operate under the RE framework. Estimation of the model still does not require distributional assumptions on v or η, but OLS on the transformed model no longer represents an efficient estimator given that the composed error term, $v_{it} - \eta_i$ no longer has a diagonal variance-covariance matrix, besides the requirement of no correlation between inefficiency and inputs. Schmidt & Sickles (1984) discuss estimation under the RE framework through generalized least squares as well. Another alternative, if one was uncomfortable with the implications stemming from RE framework would be to make

distributional assumptions, and estimate the model via maximum likelihood. This avenue was suggested by Pitt & Lee (1981) and can allow time-invariant covariates to enter the model while still identifying time-invariant inefficiency. The cost is the use of distributional assumptions so that the likelihood function can be constructed. Following Aigner et al. (1977), Pitt & Lee (1981) assume that η_i follows a half-normal distribution and v_{it} follows a normal distribution. Kumbhakar (1987) discussed estimation of inefficiency in such a model by extending the JLMS formulation.

5.2. **Time-varying Technical Inefficiency Models.** The time-invariant SFM allows inefficiency to differ across individuals, but restricts any change over time. The implication of this is that an inefficient firm could not improve productivity over time by lessening inefficiency. This may be unrealistic in a variety of applied settings, or where T is large. We must consider models that allow both technology and inefficiency to change over time to accommodate the idea of productivity and efficiency improvement at the firm level.

A nice feature of time-varying SFMs is that the time-invariant SFM is a special case and, correspondingly, the time-invariant specification can be tested, opening up a variety of inferential opportunities for empirical analyses. To introduce the time-varying SFM, recall the model in (5.5):

$$y_{it} = c_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it}. \tag{5.8}$$

To allow c_i to be time-varying, one may impose some reasonable and tractable structure, e.g., Cornwell, Schmidt & Sickles (1990) suggested replacing c_i by c_{it} where

$$c_{it} = c_{0i} + c_{1i}t + c_{2i}t^2, \tag{5.9}$$

where t is the time trend variable. The parameterization in (5.9) allows the parameters to be firm-specific. 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 a standard OLS regression. The coefficients associated with the firm dummies and their interactions are the estimates of c_{0i}, c_{1i}, and c_{2i}. These estimated coefficients can be used to obtain estimates of c_{it}, $\tilde{c}_{it}$. Again, the within estimator can be used to consistently estimate $\boldsymbol{\beta}$ along with the $3n$ parameters from the parameterization of c_{it}. Finally, $\hat{c}_{it}$ (the estimator of relative inefficiency) is obtained from

$$\hat{c}_{it} = \hat{c}_t - \tilde{c}_{it} \quad \text{where} \quad \hat{c}_t = \max_j(\tilde{c}_{jt}) \quad \forall t. \tag{5.10}$$

In this model efficiency is calculated relative to the best firm in each year. Since the firm with the maximum $\tilde{c}_{jt}$ is likely to change over time, different firms may be fully efficient (or inefficient at different levels) in different years. An alternative would be to calculate $\hat{c}_{jt} = \max_{jt}(\tilde{c}_{jt})$, the maximum over all j and t and replace $\hat{c}_t$ with this definition in (5.10), then efficiency is relative to the firm that was the most efficient over the entire sample period.

The Cornwell et al. (1990) estimation procedure is easy to implement. It relies on the standard panel data estimator with the FE framework. Note that since t appears in the inefficiency function, it cannot also appear as a regressor in $\boldsymbol{x}_{it}$, which would be required if one were to capture technical change, i.e., a shift in the production frontier, $m(\boldsymbol{x})$. In other words, the above model cannot separate inefficiency from technical change, which is an obvious drawback of this approach. In general, if one wants to have both time-varying inefficiency and technical change, then the distribution free route of Cornwell et al. (1990) will not work. In this case distributional assumptions will be necessary to allow time (and higher powers of it) to enter the model in various places.

In a model with large n and small T the model will have too many parameters ($3n$ parameters in the c_{it} function alone). A somewhat parsimonious time-varying inefficiency model was proposed by Lee & Schmidt (1993):

$$y_{it} = m(\boldsymbol{x}_{it}; \boldsymbol{\beta}) + v_{it} - u_{it} = m(\boldsymbol{x}_{it}; \boldsymbol{\beta}) + \varepsilon_{it}. \tag{5.11}$$

where $u_{it} = u_i \ell_t$ and ℓ_t represent time specific effects to be estimated. This model is quite flexible in its ability to model time-varying inefficiency. However, the temporal pattern of inefficiency is assumed to be exactly the same for all firms (ℓ_t). Under the FE 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. Though no distributional assumptions are required by Lee & Schmidt (1993), the structure of inefficiency is similar to that assuming the scaling property discussed above. Again, given that inefficiency depends directly upon time it is difficult to model both time-varying inefficiency and technical change in 5.11.

A similar idea was used prior to Lee & Schmidt (1993) in Kumbhakar (1990) and Battese & Coelli (1992), who proposed time-varying SFMs, but made distributional assumptions on both v_{it} and u_{it} and estimated the corresponding likelihood functions. Lee & Schmidt's (1993) model is more general than either the Kumbhakar (1990) or Battese & Coelli (1992) models as both can be derived as special cases with appropriate parametric restrictions on ℓ_t. Further still, the time-invariant SFM is also a special case: $\ell_t = 1 \; \forall \; t$. Once ℓ_t and u_i are estimated, inefficiency can be estimated from

$$\hat{u}_{it} = \max_j \{\hat{u}_j \hat{\ell}_t\} - \hat{u}_i \hat{\ell}_t. \tag{5.12}$$

So far, the time-varying models that we have discussed treat inefficiency in a fully deterministic fashion, i.e. no distributional assumptions are required. In the Lee & Schmidt (1993) time-varying SFM, both u_i and ℓ_t are deterministic. This model can also be estimated treating the time component as deterministic, but the individual component as stochastic (through a distributional assumption). The deviation from the Lee & Schmidt (1993) time-varying SFM in (5.11) is that $u_{it} = G(t)u_i$ with $G(t)$ being a deterministic function of time and $u_i \sim N_+(\mu, \sigma_u^2)$ (Kumbhakar 1990, Battese & Coelli 1992). The ideas discussed pertaining to the scaling property appear here, where firms have a base level of inefficiency, and

then, through time, become more or less efficient. The stochastic component, u_i, utilizes the panel structure of the data in this model. The $G(t)$ component is common across individuals (as in, but not limited to, Lee & Schmidt 1993).

Given $u_i \geq 0$, $u_{it} \geq 0$ is ensured by having $G(t) > 0$. Undoubtedly, the most popular form of $G(t)$ is that proposed by Battese & Coelli (1992)

$$G(t) = \exp\left[\gamma(t - T)\right], \tag{5.13}$$

where T is the terminal period of the sample. The specification for $G(t)$ is a simplification of the first attempt to introduce stochasticity into the time-varying SFM by Kumbhakar (1990) that assumes a more general specification of $G(t)$ given by

$$G(t) = \left[1 + \exp(\gamma_1 t + \gamma_2 t^2)\right]^{-1}. \tag{5.14}$$

The Battese & Coelli (1992) specification essentially enforces $\gamma_2 = 0$ in the Kumbhakar (1990) time-varying SFM. The popularity of the Battese & Coelli (1992) time-varying SFM has been aided by the freely available statistical package Frontier V4.1 which implements this model at the push of a button (see Section 9 as well). Other specifications for $G(t)$ have also been proposed, see Cuesta (2000) and Kumbhakar & Wang (2005) for more recent examples. Little research has been done on comparing a variety of forms of $G(t)$. Lastly, modelling technical change in the Kumbhakar (1990) or Battese & Coelli (1992) framework is trivial because the imposition of distributional assumptions allows inclusion of t (as a deterministic time-trend, e.g., linear, quadratic, etc.) as a component of $\boldsymbol{x}_{it}$.

5.3. **Models that Separate Firm Heterogeneity from Inefficiency.** While the time-invariant SFM is a standard panel data model where c_i is the unobservable individual effect, a notable drawback of this approach is that inefficiency is indistinguishable from individual heterogeneity. All time-invariant heterogeneity is confounded with inefficiency, and therefore $\hat{c}_i$ will capture heterogeneity in addition to, or even instead of, inefficiency (Greene

2005*b*). An important question for practitioners using the time-invariant SFM is how to view the time-invariant component. Should it be thought of as persistent inefficiency (as per Kumbhakar 1991, Kumbhakar & Hjalmarsson 1993, Kumbhakar & Heshmati 1995, Kumbhakar & Hjalmarsson 1998) or is it more appropriate to think of it as individual heterogeneity, capturing the effects of unobserved time-invariant covariates? If it is the latter, then the insights from the time-invariant panel data SFMs are incorrect. A less rigid perspective is that the truth lies somewhere in the middle; inefficiency may be decomposed into two components: one that is persistent over time and one that varies over time.

Unless persistent inefficiency is disentangled from time-invariant individual heterogeneity, practitioners need to choose between either the case in which c_i represents persistent inefficiency or c_i represents an individual-specific effect (heterogeneity). Here, we will discuss 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 model we will describe is

$$y_{it} = c_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - u_{it}. \tag{5.15}$$

Compared to a standard panel data model, we have the additional time-varying inefficiency term, $-u_{it}$, in (5.15). If one treats c_i, $i = 1, \ldots, n$ as a random variable that may be correlated with $\boldsymbol{x}_{it}$ but does not capture inefficiency, then the above model becomes what has been termed the 'true fixed-effects' panel SFM (Greene 2005*a*). The model is labeled as the 'true random-effects' SFM when c_i is treated as uncorrelated with $\boldsymbol{x}_{it}$. Note that these specifications are of the same nature as the models proposed by Kumbhakar (1991), Kumbhakar & Hjalmarsson (1993), Kumbhakar & Heshmati (1995), and Kumbhakar & Hjalmarsson (1998). The difference is in the interpretation of the 'time-invariant term', c_i.

Estimation of the model in (5.15) is not straightforward. When c_i, $i = 1, \ldots, n$, are embedded in the FE 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 c_i in (5.15). In this situation, consistency of the parameter estimates is not guaranteed even if $n \to \infty$ because the number of c_i increases with n. Therefore, usual asymptotic results may not apply. In addition to this specific statistical problem, another technical issue in estimating (5.15) is that the number of parameters to be estimated can be prohibitively large for large nT.

For a standard linear panel data model (i.e., one that does not have $-u_{it}$ in (5.15)), the literature has developed estimation methods to deal with this problem. These methods involve transforming the model so that c_i is removed before estimation. Without c_i in the transformed model, the incidental parameters problem no longer exists and the number of parameters to be estimated no longer increases with the number of individuals. Methods of transformation include conditioning the model on c_i's sufficient statistic[33] 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). For the basic panel data SFM, this could be done by transforming the error term if assumptions on v_{it} and u_{it} are such that the composed error term's distribution is closed-skew normal (i.e., the normal-half-normal distributional pair).

These standard methods, however, are usually not applicable to (5.15). For the conditional MLE of (5.15), Greene (2005*b*) showed that there is no sufficient statistic for c_i. For the marginal MLE, the resulting model after the within or first-difference transformation usually does not have a closed form likelihood function, if one uses standard procedures.[34] 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

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

[34] Colombi, Martini & Vittadini (2011) showed that the likelihood function has a closed form expression. Chen et al. (2014) considered a special case of Colombi et al. (2011) and derived a closed form expression.

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 c_i, $i = 1, \ldots, n$ 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 relatively large (e.g., $T \geq 10$). 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 FE framework. They showed that the likelihood function of the within transformed and the first-difference model have closed form expressions using results in Domínguez-Molina, González-Farías & Ramos-Quiroga (2003). The same theorem in Domínguez-Molina et al. (2003) is used by Colombi, Kumbhakar, Martini & Vittadini (2014) to derive the log-likelihood function in the RE framework.

Using a different approach, Wang & Ho (2010) solve the problem classified in Greene (2005*b*) by proposing a class of SFMs in which the within and first-difference transformations 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 c_is are removed from the model in (5.15), 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 c_is reduces the number of parameters to be estimated to a manageable number. The catch is in the specification of inefficiency which is the product of an i.i.d non-negative random component and a deterministic function of $\boldsymbol{z}_{it}$ (determinants of inefficiency). Formally, the Wang & Ho (2010) model is:

$$y_{it} = c_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \tag{5.16}$$

where $\varepsilon_{it} = v_{it} - u_{it}$ with $v_{it} \sim N(0, \sigma_v^2)$ and $u_{it} = g_{it}u_i^*$ with $u_i^* \sim N_+(\mu, \sigma_u^2)$, which is 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 $g_{it} = g(\boldsymbol{z}_{it}'\boldsymbol{\delta})$. What allows the model transformation to be applied is the scaling property; the within and first-difference transformations leave this stochastic term intact as u_i^* does not change with time. As Wang & Ho (2010) showed that the within-transformed and the first-differenced models are algebraically identical we have only provided discussion on the first-differenced model. 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. Lastly, as with the models of Kumbhakar (1990) or Battese & Coelli (1992), the use of distributional assumptions allows both time-varying inefficiency and technical change to be modeled in (5.16).

5.4. **Models that Separate Persistent and Time-varying Inefficiency.** Although several 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. It is important to quantify persistent inefficiency, especially in short panels, as it captures the effects of inputs like management quality (Mundlak 1961). Unless there is a change in something that affects management practices at the firm (for example new government regulations or a change in ownership), it is unlikely that persistent inefficiency will change. The importance of persistent inefficiency contrasts with time-varying as this can change over time without requiring structural changes which impact 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 argued 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[35]

$$y_{it} = \beta_0 + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} = \beta_0 + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} - (\eta_i + u_{it}) \tag{5.17}$$

Technical inefficiency is represented as $\eta_i + u_{it}$ where η_i is the persistent, firm-specific component (for example, time-invariant ownership or geographic location) and u_{it} is the time-varying component of technical inefficiency which is firm and time specific. Model (5.17) generalizes the previously discussed models because it allows for firm heterogeneity, time-invariant and time-varying inefficiency all at once.

Such a decomposition is desirable because, since η_i does not change over time, for a firm to improve efficiency a structural change in policy or management would need to arise. Additionally, η_i does not fully capture firm level inefficiency because it does not account for learning over time since it is time-invariant; the time-varying component, u_{it} can capture this aspect. In (5.17) the level of overall firm inefficiency, as well as the components, are important to know because they convey different types of information. Thus, for example, it may be argued that if residual inefficiency for a firm is relatively large in a particular year this is due to an event which is unlikely to occur in the following next year. Alternatively, if persistent inefficiency is large, then a firm is expected to operate with a relatively high

[35]This is the model proposed by Kumbhakar & Hjalmarsson (1993), Kumbhakar & Heshmati (1995), and Kumbhakar & Hjalmarsson (1998), among others.

level of inefficiency over time, unless some changes in policy and/or management occur. Therefore, a large value of η_i is more concerning in the long run given its persistent nature than is a high value of u_{it}.

The specification in (5.17) offers that advantage of testing for the presence of the persistent nature of technical inefficiency without the imposition of a specific parametric form of time-dependence. Furthermore, by including time in the $\boldsymbol{x}_{it}$ vector, (5.17) has the ability to separate exogenous technical change from technical inefficiency.

To estimate the model we rewrite (5.17) as

$$y_{it} = \alpha_i + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \omega_{it} = (\beta_0 - \eta_i - E[u_{it}]) + \boldsymbol{x}'_{it}\boldsymbol{\beta} + v_{it} - (u_{it} - E[u_{it}]). \tag{5.18}$$

The error, ω_{it}, has zero mean and constant variance. Model (5.18) is a standard panel data model with firm-specific heterogeneity (one-way error component model), and can be estimated either by the within transformation (under the FE framework) or by generalized least-squares (under the RE framework).

The SFM in (5.18) can be estimated under the FE framework using a step-wise procedure.

Step 1: The standard within transformation can be performed on (5.18) 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 (5.18) to obtain consistent estimates of $\boldsymbol{\beta}$.

Step 2: Given the estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, from Step 1, construct the pseudo-residuals $r_{it} = y_{it} - \boldsymbol{x}'_{it}\hat{\boldsymbol{\beta}}$, which contain information on $\alpha_i + \omega_{it}$. Using these, we first estimate α_i from the mean of r_{it} for each i. Then, we can estimate α_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_{it} over time for a firm. The above formula gives an estimate of α_i relative to the best firm in the sample.

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

$$\ln \mathcal{L} = -N \ln \sigma + \sum_{i=1}^{n}\sum_{t=1}^{T} \ln \Phi(-e_{it}\lambda/\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{t=1}^{T} e_{it}^2 \tag{5.19}$$

Note that the parameters to be estimated here are β_0, σ_ν^2 and σ_τ^2. Once these parameters have been estimated a JLMS conditional mean or median technique can be used to estimate u_{it} for each observation.

To summarize estimation under the FE framework, we estimate (5.18) using standard FE panel data tools to obtain consistent estimates of $\boldsymbol{\beta}$ in Step 1. Step 2 estimates persistent technical inefficiency, η_i. Lastly, Step 3 involves estimation of β_0 and the parameters associated with the distributional assumptions imposed on the random components, v_{it} and u_{it}. One can then use the JLMS formula to estimate the time-varying (residual) component of inefficiency, u_{it}. Note that no distributional assumptions are used in the first two steps. Without further assumptions, residual inefficiency cannot be identified and hence, distributional assumptions are needed in the last step. This model can also be estimated under the RE framework (see also Colombi et al. 2014).

5.5. **Models that Separate Firm Effects, Persistent Inefficiency and Time-varying Inefficiency.** All of the panel data SFMs introduced so far have departed from the general model introduced in (5.2) in some aspect pertaining to the four separate error components. This is due to the fact, that until recently, it was not clear how to estimate the full panel data SFM represented by (5.2). The models of Kumbhakar, Lien & Hardaker (2014) and Colombi et al. (2014) overcome the limitations of the previous models by embracing the

nature of the four component structure inherent in the general panel data SFM. In the SFM represented in (5.2) the four components take into account different factors affecting output, given the inputs. As in Greene (2005*b*, 2005*a*) the first component captures firms' latent heterogeneity, which needs to be extricated from inefficiency; the second component captures time-varying inefficiency, the third component captures time-invariant inefficiency as in Kumbhakar & Hjalmarsson (1993), Kumbhakar & Heshmati (1995), and Kumbhakar & Hjalmarsson (1998) while the fourth component captures stochastic shocks beyond control of the firm.

The ability to estimate model (5.2) allows improvement over the previous models in several ways. To begin, while some of the time-varying inefficiency models just described can accommodate firm effects, these models fail to acknowledge the potential for factors that might have time-invariant effects on firm inefficiency. Second, SFMs which allow time-varying inefficiency commonly assume that the inefficiency level of the firm at time t is independent of its previous level of inefficiency; it is more reasonable to assume that a firm may eliminate some of its inefficiency by mitigating short-run rigidities, while other sources of inefficiency may remain over time. The former is captured by the time-invariant component, η_i, and the latter by the time-varying component, u_{it}. Finally, many panel SFMs do consider time-invariant inefficiency, but do not simultaneously account for the presence of unobserved firm heterogeneity. In doing so, these models confound time-invariant inefficiency with firm effects (heterogeneity). The 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 firm-specific time-varying inefficiency term; a firm-specific effect capturing latent heterogeneity; and a time- and firm-varying 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 inefficiency. As before, latent heterogeneity is confounded with long-run inefficiency.

Estimation of the panel data SFM in (5.2) can be undertaken in a single stage MLE method based on distributional assumptions on the four components (Colombi et al. 2011). We first describe a simpler, multi-step procedure (Kumbhakar et al. 2014). For this, we rewrite the model in (5.2) as

$$y_{it} = \beta_0^* + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \tag{5.20}$$

where $\beta_0^* = \beta_0 - E[\eta_i] - E[u_{it}]$; $\alpha_i = c_i - \eta_i + E[\eta_i]$; and $\varepsilon_{it} = v_{it} - u_{it} + E[u_{it}]$. With this specification both α_i and ε_{it} are zero mean and constant variance random variables. (5.20) is estimated in three steps.

Step 1: Standard random effect panel regression is used to estimate $\hat{\beta}$ (since (5.20) is a common panel data model). Predicted values of α_i and ε_{it}, denoted by $\hat{\alpha}_i$ and $\hat{\varepsilon}_{it}$ are also available after estimating (5.20).

Step 2: Time-varying technical inefficiency, u_{it}, is estimated using $\hat{\varepsilon}_{it}$ from Step 1. Since

$$\varepsilon_{it} = v_{it} - u_{it} + E[u_{it}], \tag{5.21}$$

by assuming v_{it} is i.i.d. $N(0, \sigma_v^2)$ and u_{it} is i.i.d. $N_+(0, \sigma_u^2)$, which yields $E[u_{it}] = \sqrt{2/\pi}\,\sigma_u$, and ignoring the difference between the true and predicted values[36] of ε_{it}, we can estimate (5.21) using standard SFA techniques. Doing so produces predictions of the time-varying technical inefficiency component u_{it}, $E\left[e^{-u_{it}}|\varepsilon_{it}\right]$, (i.e., Battese & Coelli 1988), which we call relenting technical efficiency (RTE).

Step 3: Estimate η_i following a similar strategy as in Step 2. For this we use $\hat{\alpha}_i$ from Step 1. Since

$$\alpha_i = c_i - \eta_i + E[\eta_i], \tag{5.22}$$

[36]Which is the standard practice in any two- or multi-step procedure.

by assuming c_i is i.i.d. $N(0, \sigma_\mu^2)$, η_i is i.i.d. $N_+(0, \sigma_\eta^2)$, where $E[\eta_i] = \sqrt{2/\pi}\,\sigma_\eta$, estimate (5.22) using the standard normal-half-normal cross-section SFM and obtain estimates of the persistent technical inefficiency component, η_i following JLMS. Persistent technical efficiency (PTE) can then be estimated as PTE $= e^{-\eta_i}$, where $\hat{\eta}_i$ is the JLMS estimator of η_i. Overall technical efficiency (OTE) is then constructed as the product of PTE and RTE, i.e., OTE = PTE×RTE.

It is possible to extend this model (in steps 2 and 3) to include PTE and RTE that is distributed as truncated-normal or exponential as opposed to half-normal.

While the multi-step approach of Kumbhakar et al. (2014) is straightforward to implement, it is inefficient relative to full MLE. However, given the structure of the four separate errors, deriving the likelihood function was previously seen as infeasible. However, using insights related to the closed-skew normal distribution, as in Colombi et al. (2014), a tractable likelihood function turned out to be easily obtainable.

Colombi et al. (2014) made skew normal distributional assumptions for both $c_i - \eta_i$ and $v_{it} - u_{it}$ in (5.20).[37] Assuming v_{it} is i.i.d 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 c_i and η_i. Thus, model (5.2)'s likelihood, can be derived. Even though the log-likelihood for (5.2) can be determined based on skew normal assumptions for the time-varying and time-invariant error components, it can be daunting to implement. Greene & Fillipini (2014) recently proposed a simulation based optimization routine which circumvents many of the challenges associated with direct optimization. They used a trick suggested by Butler & Moffitt (1982), conditioning on c_i and η_i. This conditioning eliminates many of the computational hurdles that direct optimization of the likelihood function presents.

5.6. **The Four Component Panel Data SFM with Determinants of Inefficiency.** A further generalization of the four component model in (5.2) involves the inclusion of

[37]The skew normal distribution is a more general distribution than the normal distribution, allowing for asymmetry (Azzalini 1985).

determinants of inefficiency, either for the time-varying or the time invariant components. An estimator for this model was recently proposed in Badunenko & Kumbhakar (2017),

$$y_{it} = m(\boldsymbol{x}_{it};\boldsymbol{\beta}) + c_i - \eta_i + v_{it} - u_{it}, \tag{5.23}$$

where $\eta_i \sim N_+(0,\sigma^2_{\eta,i})$, $u_{it} \sim N_+(0,\sigma^2_{u,it})$, $c_i \sim N(0,\sigma^2_{c,i})$, and $v_{it} \sim N(0,\sigma^2_{v,it})$. These distributional assumptions are imposed so that the time invariant composed error and the time-varying composed error both follow the closed skew normal distribution. Each of the variance parameters of the four components is dependent upon a set of covariates and specified as an exponential function: $\sigma^2_{\eta,i} = \sigma^2_{\eta} e^{\boldsymbol{z}'_{\eta,i}\boldsymbol{\delta}_\eta}$, $\sigma^2_{c,i} = \sigma^2_{c} e^{\boldsymbol{z}'_{c,i}\boldsymbol{\delta}_c}$, $\sigma^2_{u,it} = \sigma^2_{u} e^{\boldsymbol{z}'_{u,it}\boldsymbol{\delta}_u}$, and $\sigma^2_{v,it} = \sigma^2_{v} e^{\boldsymbol{z}'_{v,it}\boldsymbol{\delta}_v}$. The time-constant and time-varying $\boldsymbol{z}$ vectors can overlap due to the assumed distributional assumptions, that is $\boldsymbol{z}_{c,i}$ can share elements with $\boldsymbol{z}_{\eta,i}$ and $\boldsymbol{z}_{u,it}$ can share elements with $\boldsymbol{z}_{v,it}$.

To estimate this four component model Badunenko & Kumbhakar (2017) used the insights of Greene & Fillipini (2014) and deployed simulated maximum likelihood techniques. The benefit of this approach is that rather than having T integrals to evaluate, by conditioning on $c_i - \eta_i$, the likelihood function can be written as the product of T univariate integrals. Simulation methods are required to construct draws of $c_i - \eta_i$ inside the convolution density. The final log-likelihood function is

$$\mathcal{L} = \sum_{i=1}^{n} \log\left(R^{-1} \sum_{r=1}^{R} \left[\prod_{t=1}^{T} \frac{2}{\sigma_{it}} \phi\left(\frac{\varepsilon_{itr}}{\sigma_{it}}\right) \Phi\left(\frac{\varepsilon_{itr}\lambda_{it}}{\sigma_{it}}\right) \right] \right), \tag{5.24}$$

where $\sigma_{it} = \sqrt{e^{\boldsymbol{z}'_{u,it}\boldsymbol{\delta}_u} + e^{\boldsymbol{z}'_{v,it}\boldsymbol{\delta}_v}}$, $\lambda_{it} = \sqrt{e^{\boldsymbol{z}'_{u,it}\boldsymbol{\delta}_u - \boldsymbol{z}'_{v,it}\boldsymbol{\delta}_v}}$, $\varepsilon_{itr} = \epsilon_{it} - \left(\sqrt{e^{\boldsymbol{z}'_{c,i}\boldsymbol{\delta}_c}} V_{ir} - \sqrt{e^{\boldsymbol{z}'_{\eta,i}\boldsymbol{\delta}_\eta}} |U_{ir}|\right)$ and $\epsilon_{it} = y_{it} - m(\boldsymbol{x}_{it};\boldsymbol{\beta})$. R is the number of draws over which to numerically evaluate the integral (larger R increases accuracy but slows down the routine, smaller R leads to faster computation but decreases accuracy). Lastly, both V_{ir} and U_{ir} are random draws from a standard normal distribution. Implementation of this routine is straightforward if one has access to a standard normal random number generator (which is typically available in any

general statistical software). Once draws for V_{ir} and U_{ir} have been constructed, the likelihood is evaluated for the current set of parameters $(\boldsymbol{\beta}, \boldsymbol{\delta}_u, \boldsymbol{\delta}_v, \boldsymbol{\delta}_\eta, \boldsymbol{\delta}_c)$. This process is then iterated over different sets of parameter values. Naturally, one can impose constancy at various parts of the error components by restricting $\boldsymbol{\delta}_\ell = 0$ for $\ell \in \{u, v, c, \eta\}$.

5.7. **Inference Across the Panel Data SFM.** The most general SFM in the panel context is the model which allows for firm specific heterogeneity, persistent technical efficiency, relenting technical inefficiency and individual-time specific idiosyncratic shocks. Colombi et al. (2014) denote this model as TTT (True for having firm specific heterogeneity, True for having time constant inefficiency and True for having time-varying inefficiency). The majority of all panel data models that have appeared in the literature are special cases of TTT. For example, the widely used true RE model of 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 model which best fits the data at hand. One benefit of nearly all of the panel data SFM discussed here is that standard panel data type tests (coefficient significance, fixed versus random effects framework, serial correlation, etc.) are easily implemented. This is similar to the benefits of the cross-sectional SFM that we discussed earlier.

What is less straightforward is to test the most general TTT model against more restricted versions. 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 of the parameters of interest lie on the boundary of the parameter space. Under reasonable assumptions the asymptotic distribution of the log-likelihood ratio test statistic is $\bar{\chi}^2$, as discussed in Section 2.3.1. For example, the model of Pitt & Lee (1981) could be tested against the TTT model with the log-likelihood ratio test statistic but using the $\bar{\chi}^2$ to determine the p-value, see Table 1.

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 efficiency yields different policy recommendations and so working with models that document their presence, or lack of one, are important for proper analysis.

6. Nonparametric Estimation of the SFM

6.1. **Early Attempts.** In a nutshell, the semiparametric and nonparametric approaches to SFA typically use the benchmark SFM of Aigner et al. (1977) as the stepping-stone, generalizing it in different ways by relaxing all or some parametric assumptions by utilizing existing semiparametric and nonparametric statistical methods, such as the Nadaraya-Watson estimator, the local polynomial estimator or the likelihood (pseudo or local) estimators.

To facilitate further and more precise discussion recall that the benchmark SFM for a sample of n DMUs is given by:

$$y_i = m(\boldsymbol{x}_i) + v_i - u_i = m(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{6.1}$$

where $m(\cdot)$ is the frontier of the production technology that can be used to transform vector of inputs $\boldsymbol{x} \in \mathbb{R}^q_+$ into scalar output y_i, perturbed by some statistical noise v_i and adjusted by technical inefficiency u_i. As we discussed in Section 2, traditional parametric estimation of the model begins by assuming a particular functional form for the production technology, most commonly a Cobb-Douglas or a Translog, besides making distributional assumptions on both v_i and u_i, which help to identify and estimate the unknown parameters via, say, the maximum likelihood approach. All the asymptotic results (consistency, asymptotic normality) are conditional on these assumptions and if they happen to be incorrect then, strictly speaking, all these results may be invalid. In such cases, the parametric MLE will be inconsistent or converging in probability not to the truth (e.g., true elasticities) but to some other numbers, which can even be very far from the truth if the parametric assumption made on a function is far from the true one.

The early attempts to estimating SFM nonparametrically or semiparametrically go back to at least Banker & Maindiratta (1992), Fan, Li & Weersink (1996) and Kneip & Simar (1996). Specifically, Banker & Maindiratta (1992) proposed a nonparametric approach in the spirit of the DEA estimator but embedded in a maximum likelihood framework, similar to parametric SFA, and thus allow for modeling both the noise and the inefficiency. A few years later, Fan et al. (1996) proposed estimating the production frontier in another flexible manner, using nonparametric kernel regression methods embedded into the parametric maximum likelihood. About the same time, Kneip & Simar (1996) suggested using the kernel regression estimator (Nadaraya-Watson in particular) for the panel data SFM.

Importantly, note that the estimated conditional mean $E[y_i|\boldsymbol{x}]$ of the production frontier is a biased estimator when ignoring the inefficiency term. Indeed, a critical assumption for consistent estimation of the production frontier in a regression setting is $E[\varepsilon_i|\boldsymbol{x}] = 0$ and due to the one-sided nature of u_i, this assumption is not satisfied, because $E[\varepsilon_i|\boldsymbol{x}] = \mu_u \neq 0$ in the simplest case when inefficiency is independent of the inputs, or more generally, $E[\varepsilon_i|\boldsymbol{x}] = \mu_u(\boldsymbol{x}) \neq 0 \ \forall \boldsymbol{x}$. Therefore, the production frontier cannot be identified in the regression setup, where one would estimate

$$y_i = m(\boldsymbol{x}_i) + \varepsilon_i = m(\boldsymbol{x}_i) + \mu_u + (\varepsilon_i - \mu_u) \equiv m^*(\boldsymbol{x}_i) + \varepsilon_i^*. \tag{6.2}$$

Realizing this, Fan et al. (1996) proposed correcting the estimation bias of $m(\boldsymbol{x})$ via a three-stage semiparametric pseudo-likelihood estimation of the SFM. In this approach, at the first stage, one estimates (6.2) non-parametrically.[38] Results from this first stage are then fed into the second stage, involving parametric MLE with particular assumptions on the distribution of the noise and inefficiency that help identifying and disentangling the two.[39] Once the parameters of this symbiosis of MLE and kernel-regression are estimated,

[38]They used a local constant (Nadaraya-Watson) regression, although other consistent nonparametric estimators can be used there too.

[39]In their work, the normal-half-normal assumption was used, but other assumptions we discussed above can be used there too. Note however that for some alternative distributional assumptions on u, for example

the estimated conditional mean can then, in the third stage, be corrected for the bias by the estimated mean of inefficiency (as in COLS), $\hat{\mu}_u(\boldsymbol{x}_i)$ to get a consistent estimator $m(\boldsymbol{x}_i)$ given by

$$\widehat{m}(\boldsymbol{x}_i) = \widehat{m}^*(\boldsymbol{x}_i) - \hat{\mu}_u(\boldsymbol{x}_i), \tag{6.3}$$

Kneip & Simar (1996) also proposed a similar strategy for correcting for the bias occurring in estimating (6.2) nonparametrically, but avoided using MLE due to possibility to disentangle the noise from inefficiency without distributional assumptions, by utilizing the panel-data SFA framework.

The approaches of Fan et al. (1996) and Kneip & Simar (1996) provided a useful framework and formed a foundation on which many other approaches have been built.[40] For example, more recent approaches of Kuosmanen & Kortelainen (2012) and Parmeter & Racine (2012) share some essence of Fan et al. (1996) except that they required the estimated production frontier to obey traditional axioms of production, such as monotonicity and concavity, something that Fan et al. (1996) did not accommodate in their approach. Specifically, Parmeter & Racine (2012) employ the framework of Fan et al. (1996) but combine it with constraint weighted bootstrapping (Hall & Huang 2001, Du, Parmeter & Racine 2013) to ensure that monotonicity and concavity are enforced during estimation. More recently, Noh (2014) made improvements to the approach of Parmeter & Racine (2012), which resulted in small sample performance gains. On the other hand, Kuosmanen & Kortelainen (2012) used an entirely different estimation approach, concave nonparametric least-squares (CNLS), to impose monotonicity and concavity. Lastly, Martins-Filho & Yao (2015) showed that while the estimator of Fan et al. (1996) is consistent, the parametric estimator for the parameters of the density of the convolved error yields an asymptotic bias (when normalized by $\sqrt{n}$)

exponential or truncated-normal, a concentrated version of the log-likelihood function may not exist, causing identification problems.

[40]See Parmeter & Zelenyuk (2016) for a more comprehensive review of this topic.

and proposed an alternative estimator that estimates the distributional parameters and the unknown frontier jointly.

6.2. **Local Likelihood Methods.** Local likelihood approach (Tibshirani & Hastie 1987) is known to be a natural alternative to the semi-parametric pseudo-likelihood, and it was first proposed in the SFA context by Kumbhakar, Park, Simar & Tsionas (2007). This approach closely resembles the parametric likelihood approach with the only (yet fundamental) difference being the kernel-based weights (instead of the equal weights) used to weigh each individual contribution to the likelihood, which help localizing the estimation in the direction of each continuous variable through the bandwidths. Specifically, for a given regression error density, $f_\varepsilon(\varepsilon, \theta)$, we have the local log-likelihood function

$$\check{\mathcal{L}}_n(\theta(\boldsymbol{x}), m_{\boldsymbol{x}}) = (n|h|)^{-1} \sum_{i=1}^{n} \ln f_\varepsilon(y_i - m(\boldsymbol{x}_i); \theta(\boldsymbol{x})) K_{i\boldsymbol{x}}, \tag{6.4}$$

where $m_{\boldsymbol{x}}$ captures the conditional mean of y given $\boldsymbol{x}$ (a $q \times 1$ vector of covariates) and θ is the vector of remaining parameters of f_ε, $K_{i\boldsymbol{x}} = \prod_{s=1}^{q} h_s^{-1} k\left(\frac{\boldsymbol{x}_{is} - \boldsymbol{x}_s}{h_s}\right)$ is the standard product kernel where $k(\cdot)$ is any second order univariate kernel (Epanechnikov, Gaussian, e.g.), h_s is the smoothing parameter for the s^{th} covariate (and is the s^{th} element of vector h), while $|h| = h_1 h_2 \cdots h_q$.

Kumbhakar et al. (2007) used a local-linear approximation for the unknown production function $m(\boldsymbol{x}_i)$ combined with the assumption of a normal, half-normal convolved error term, where parameters are also modeled as unknown functions of the covariates,

$$\check{\mathcal{L}}_n = (n|h|)^{-1} \sum_{i=1}^{n} \left[-0.5\ddot{\sigma}^2_{\boldsymbol{x}}(\boldsymbol{x}_i) - 0.5\ddot{\varepsilon}_i^2 e^{-\ddot{\sigma}^2_{\boldsymbol{x}}(\boldsymbol{x}_i)} + \ln \Phi\left(-\ddot{\varepsilon}_i e^{\ddot{\lambda}_{\boldsymbol{x}}(\boldsymbol{x}_i) - 0.5\ddot{\sigma}^2_{\boldsymbol{x}}(\boldsymbol{x}_i)} \right) \right] K_{i\boldsymbol{x}} \tag{6.5}$$

where $\ddot{\varepsilon}_i = y_i - \ddot{m}_{\boldsymbol{x}}(\boldsymbol{x}_i)$, $\ddot{m}_{\boldsymbol{x}}(\boldsymbol{x}_i) = \ddot{m}_0 - \ddot{m}_1'(\boldsymbol{x}_i - \boldsymbol{x})$, $\ddot{\sigma}^2_{\boldsymbol{x}}(\boldsymbol{x}_i) = \ddot{\sigma}^2_0 + \ddot{\sigma}^{2\prime}_1(\boldsymbol{x}_i - \boldsymbol{x})$, and $\ddot{\lambda}_{\boldsymbol{x}}(\boldsymbol{x}_i) = \ddot{\lambda}_0 + \ddot{\lambda}_1'(\boldsymbol{x}_i - \boldsymbol{x})$.[41]

[41] One could also use a quadratic approximation, but note that even in this local-linear case, there are already $3 + 3q$ parameters to estimate (i.e., optimize over) at each point of interest $\boldsymbol{x}$: these are the three functional estimates, $\ddot{m}_0$, $\ddot{\sigma}^2_0$ and $\ddot{\lambda}_0$ and the $3q$ derivative estimates of the functions, $\ddot{m}_1$, $\ddot{\sigma}^2_1$ and $\ddot{\lambda}_1$.

Noting that often the main focus of interest is related to σ_u, Park, Simar & Zelenyuk (2015) suggested directly parameterizing the local likelihood function in terms of $\ln \sigma_v^2$ and $\ln \sigma_u^2$ which also impose positivity of σ_v^2 and σ_u^2 throughout the estimation, making it more stable computationally. Park et al. (2015) also outlined an asymptotic theory for modeling discrete variables in the context of the local-likelihood approach, which can be imperative for many applications, since many covariates researchers have access to are categorical in nature (regulated vs. non-regulated firms or industries, private vs. publicly owned companies, male vs. female managers, etc.). The local-likelihood function in this case would be

$$\begin{aligned}\check{\mathcal{L}}(\theta(\boldsymbol{x}^c,\boldsymbol{x}^d),m_{\boldsymbol{x}^c,\boldsymbol{x}^d}) =&(n|h|)^{-1}\sum_{i=1}^{n}\Big[-0.5\ln\left(e^{\ddot{\sigma}_v^2(\boldsymbol{x}_i^c,\boldsymbol{x}_i^d)}+e^{\ddot{\sigma}_u^2(\boldsymbol{x}_i^c,\boldsymbol{x}_i^d)}\right)-0.5\ddot{\varepsilon}_i^2/\left(e^{\ddot{\sigma}_v^2(\boldsymbol{x}_i^c,\boldsymbol{x}_i^d)}+e^{\ddot{\sigma}_u^2(\boldsymbol{x}_i^c,\boldsymbol{x}_i^d)}\right)\\ &+\ln\Phi\left(-\ddot{\varepsilon}_i e^{\ddot{\sigma}_u^2(\boldsymbol{x}_i^c,\boldsymbol{x}_i^d)/2-\ddot{\sigma}_v^2(\boldsymbol{x}_i^c,\boldsymbol{x}_i^d)/2}/\sqrt{e^{\ddot{\sigma}_v^2(\boldsymbol{x}_i^c,\boldsymbol{x}_i^d)}+e^{\ddot{\sigma}_u^2(\boldsymbol{x}_i^c,\boldsymbol{x}_i^d)}}\right)\Big]K_{i\boldsymbol{x}^c}W^i(\boldsymbol{x}_i^d).\end{aligned} \tag{6.6}$$

where $\boldsymbol{x}_i^c$ is a vector of continuous regressors while $\boldsymbol{x}_i^d$ is a vector of discrete regressors and $W^i(\boldsymbol{x}_i^d)$ is an appropriate discrete kernel, e.g., the one proposed by Aitchison & Aitken (1976) or its variations. The theory in Park et al. (2015) is derived for the case of kernel from Racine & Li (2004), given by $W^i(\boldsymbol{x}^d)=\prod_{j=1}^{k}\omega_j^{I(x_{ij}^d\neq x_j^d)}$, which is a standardized version of Aitchison-Aitken kernel, standardized so that the bandwidths for a jth discrete variable, here denoted as ω_j, are always between 0 and 1, regardless of the number of categories. However, this theory also extends (with some modifications) to cases with other discrete kernels. For example, one might prefer the so-called discrete Epanechnikov kernels, which are particularly useful and can be superior to others in case of sparse data (e.g., see Chu, Henderson & Parmeter (2017) and the references cited therein). One can also use more adaptive bandwidths, e.g., allow for bandwidths of some or all continuous regressors to differ across categories of some or all discrete variables (e.g., see Li, Simar & Zelenyuk (2016) for related discussion).

Standard optimization algorithms can be used here, but as with any nonlinear optimization, careful choice of starting values is imperative, especially in selecting the bandwidths.

For example, Kumbhakar et al. (2007) suggested starting with the local-linear least-squares estimates for $\ddot{m}_0$ and $\ddot{m}_1$ and the global, parametric maximum likelihood estimates for σ^2 and λ (from Aigner et al. 1977) so that $\ddot{m}_0$ is properly corrected (as in Fan et al. 1996).

Selection of the bandwidths is a very important step here (as is true in general for kernel-based methods) and many interesting general selection methods can be adapted to the current context. One of the most popular approach is cross-validation.[42] Kumbhakar et al. (2007) outlined how to use least-squares cross-validation (LSCV) for their approach. Meanwhile, Park et al. (2015) suggested using maximum likelihood cross-validation (MLCV), which is more natural for the local-likelihood approach, although it may be more demanding in computation. For the starting values in numerical optimization of LSCV or MLCV for selecting optimal bandwidths, one could use the so-called rules-of-thumb bandwidths that reflect the rates of convergence required for the asymptotic theory, e.g., for a continuous variable x_s^c, use $h_0(x_s^c) = 1.06 \times n^{-1/(4+q)}\hat{\sigma}_{x_s^c}$, where $\hat{\sigma}_{x_s^c}$ is estimated standard deviation of x_s^c, and $\omega_0 = n^{-2/(q+4)}$ for the discrete bandwidths.

Kneip, Simar & Van Keilegom (2015) provide an update of the Kumbhakar et al. (2007) estimator whereby the distributional assumption on the inefficiency term can be dropped. The only parametric assumption required in Kneip et al. (2015) is that the two-sided error term is normal, which allows them to rely on penalized likelihood, where the unknown density is constructed non-parametrically via a histogram over the support of the covariate space and the penalty term is included to ensure appropriate smoothness of the resulting density. Both the theory and simulated evidence appearing in Kneip et al. (2015) suggest that this estimator works quite well in a variety of settings. To date, no application of this method has appeared to our knowledge and so it represents an exciting opportunity moving forward.

6.3. **Local Least-Squares Approaches.** In spite of the appealing theoretical advantages of the likelihood-based approaches they involve numerical optimization of the local likelihood

[42]For more discussions on the pros and cons, as well as references on this approach in general, see Henderson & Parmeter (2015*a*).

function over many parameters at each point of interest, which can be computationally complex, especially if bootstrap methods are needed to conduct inference. An attractive alternative that is much simpler to compute is provided by adopting the local least-squares methods; because these methods do not require nonlinear optimization (given closed form solutions), only basic matrix operations are required, marking dramatic improvements in computation time.

Recently, Simar et al. (2017) (SVKZ hereafter) proposed what can be viewed as a nonparametric and semiparametric generalizations of COLS (Olson et al. 1980)[43], which also allow for modeling determinants of inefficiency. Specifically, they considered a generalization of (6.1) given by

$$y_i = m(\boldsymbol{x}_i, \boldsymbol{z}_i) + v_i - u_i = m(\boldsymbol{x}_i, \boldsymbol{z}_i) + \varepsilon_i. \tag{6.7}$$

where $m(\boldsymbol{x}_i, \boldsymbol{z}_i)$ is the production frontier evaluated at $\boldsymbol{x}_i$, the realizations of inputs for observation i, and at $\boldsymbol{z}_i$, the realization of the so-called environmental factors faced by the observation i, and disturbed by the realizations of statistical noise v_i and inefficiency u_i. In general, they required fairly general and mild conditions on the model, e.g., $(u_i|\boldsymbol{x}_i = \boldsymbol{x}, \boldsymbol{z}_i = \boldsymbol{z}) \sim D^+(\mu_u(\boldsymbol{x}, \boldsymbol{z}), \sigma_u^2(\boldsymbol{x}, \boldsymbol{z}))$ with $D^+(\cdot, \cdot)$ being a non-negative random variable with mean $\mu_u(\cdot, \cdot)$ and finite positive variance $\sigma_u^2(\cdot, \cdot)$, while $(v_i|\boldsymbol{x}_i = \boldsymbol{x}, \boldsymbol{z}_i = \boldsymbol{z}) \sim D(0, \sigma_v^2(\boldsymbol{x}, \boldsymbol{z}))$ with $D(0, \cdot)$ being a random variable with mean zero and finite positive variance $\sigma_v^2(\cdot, \cdot)$. They also assumed that, conditional on $(\boldsymbol{x}_i, \boldsymbol{z}_i)$, u_i and v_i are independent random variables. Further, given that v_i has a symmetric distribution around zero, while u_i is a positive random variable from a skewed distribution $E[\varepsilon_i|\boldsymbol{x}, \boldsymbol{z}] = -E[u_i|\boldsymbol{x}, \boldsymbol{z}] \neq 0$. Therefore, after recentering, we have

$$y_i = m(\boldsymbol{x}_i, \boldsymbol{z}_i) + v_i - u_i + E[u_i|\boldsymbol{x}, \boldsymbol{z}] - E[u_i|\boldsymbol{x}, \boldsymbol{z}] = m^*(\boldsymbol{x}_i, \boldsymbol{z}_i) + \varepsilon_i^* \tag{6.8}$$

[43]As with our earlier discussion, SVKZ referred to this approach as nonparametric MOLS, but cite Olson et al. (1980), who used the term COLS and so we refer to it as COLS here.

where $m^*(\boldsymbol{x}_i, \boldsymbol{z}_i) = m(\boldsymbol{x}_i, \boldsymbol{z}_i) - E[u_i|\boldsymbol{x}, \boldsymbol{z}]$ and $\varepsilon_i^* = \varepsilon_i + E[u_i|\boldsymbol{x}, \boldsymbol{z}]$. Adapting the strategy of COLS from Olson et al. (1980), SVKZ proposed in the first stage the estimator of $m^*(\boldsymbol{x}, \boldsymbol{z})$, $\widehat{m}^*(\boldsymbol{x}, \boldsymbol{z})$ using local-polynomial least-squares, noting that under mild regularity conditions and appropriate choice of the bandwidths, such estimators have desirable statistical properties (consistency, asymptotic normality, etc.; see Fan & Gijbels 1996, Li & Racine 2007, Henderson & Parmeter 2015*a*). Then, in the second stage, they utilized the moment conditions implied by the assumptions on u_i and v_i, namely

$$\begin{aligned} E[\varepsilon^*|\boldsymbol{x}, \boldsymbol{z}] =& 0, \\ E[(\varepsilon^*)^2|\boldsymbol{x}, \boldsymbol{z}] =& \sigma_u^2(\boldsymbol{x}, \boldsymbol{z}) + \sigma_v^2(\boldsymbol{x}, \boldsymbol{z}), \\ E[(\varepsilon^*)^3|\boldsymbol{x}, \boldsymbol{z}] =& -E\left[(u - E[u|\boldsymbol{x}, \boldsymbol{z}])^3 |\boldsymbol{x}, \boldsymbol{z}\right], \end{aligned}$$

and estimate the second and third moments of ε^* using local-polynomial methods with the residuals $\widehat{\varepsilon}_i^* = y_i - \widehat{m}^*(\boldsymbol{x}_i, \boldsymbol{z}_i)$ from the first stage, i.e.,

$$\widehat{m}_2(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n} A_i(\boldsymbol{x}, \boldsymbol{z}) \widehat{\varepsilon}_i^{*2} \tag{6.9}$$

and

$$\widehat{m}_3(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n} A_i(\boldsymbol{x}, \boldsymbol{z}) \widehat{\varepsilon}_i^{*3}, \tag{6.10}$$

where $A_j(\boldsymbol{x}, \boldsymbol{z})$ would vary depending upon the local smoothing method used. If one desires to estimate the level of the frontier in SVKZ's setup, then (local) parametric distributional assumptions for u_i is needed, although the ranking of output would be independent of this distributional choice. Importantly, note that if the moments of u_i depend on $\boldsymbol{x}$ or $\boldsymbol{z}$, then the frontier correction will also depend on $\boldsymbol{x}$ and $\boldsymbol{z}$ implying that any features of the production frontiers, such as returns to scale, may depend on the distribution of u_i. One therefore needs to either make some type of distributional assumption or to assume a type of separability

assumption, such as $E[u|\boldsymbol{x},\boldsymbol{z}] = E[u|\boldsymbol{z}]$. With the normal-half-normal framework, SVKZ showed (adapting Olson et al. 1980), that

$$\widehat{\sigma}_u(\boldsymbol{x},\boldsymbol{z}) = \max\left\{0, \left[\sqrt{\frac{\pi}{2}}\left(\frac{\pi}{\pi-4}\right)\widehat{m}_3(\boldsymbol{x},\boldsymbol{z})\right]^{1/3}\right\} \tag{6.11}$$

$$\widehat{\sigma}_v^2(\boldsymbol{x},\boldsymbol{z}) = \widehat{m}_2(\boldsymbol{x},\boldsymbol{z}) - \widehat{\sigma}_u^2(\boldsymbol{x},\boldsymbol{z})\left(\frac{\pi-2}{\pi}\right), \tag{6.12}$$

These estimates can then be used to obtain the estimates of the efficiency scores for each observation, in the spirit of Jondrow et al. (1982), generalized to the heteroskedastic case involving $E[u_i|\varepsilon_i,\boldsymbol{x}_i,\boldsymbol{z}_i]$ instead of $E[u_i|\varepsilon_i]$. However, as mentioned in the parametric context above, one should be careful interpreting these estimates of efficiency scores, as they are "predicted values" conditional on unobserved ε_i, replaced with its estimate for the specific realization i, and as such the prediction intervals tend to be quite wide (see Simar & Wilson (2010), for related discussion). In turn, the conditional mean of inefficiency can be consistently estimated as

$$\widehat{\mu}_u(\boldsymbol{x},\boldsymbol{z}) = \sqrt{\frac{2}{\pi}}\widehat{\sigma}_u(\boldsymbol{x},\boldsymbol{z}). \tag{6.13}$$

and then use it at any point of interest $(\boldsymbol{x},\boldsymbol{z})$ to form a consistent estimate of the level of frontier, $m(\boldsymbol{x},\boldsymbol{z})$, using

$$\widehat{m}(\boldsymbol{x},\boldsymbol{z}) = \widehat{m}^*(\boldsymbol{x},\boldsymbol{z}) + \widehat{\mu}_u(\boldsymbol{x},\boldsymbol{z}). \tag{6.14}$$

SVKZ also derived the asymptotic properties of these estimators, generalizing earlier results from Fan & Yao (1998) and Chen, Cheng & Peng (2009).

Finally, and perhaps most interestingly, SVKZ pointed out that if one is only interested in the influence of $\boldsymbol{z}$ or $\boldsymbol{x}$ on the (conditional mean) efficiency, or as a special case to test if $E[u|\boldsymbol{x},\boldsymbol{z}]$ is a constant, then no parametric distributional specification is required for u_i, only a condition that it belongs to the one parameter scale family of distributions. Specifically,

they showed that the elasticity measure of $E[u|\boldsymbol{x},\boldsymbol{z}]$ w.r.t. some ψ_ℓ that is an element of $(\boldsymbol{x},\boldsymbol{z})$, defined as

$$\xi_{\psi_\ell}(\boldsymbol{x},\boldsymbol{z}) = \frac{\partial \mu_u(\boldsymbol{x},\boldsymbol{z})}{\partial \psi_\ell}\frac{\psi_\ell}{\mu_u(\boldsymbol{x},\boldsymbol{z})} \tag{6.15}$$

assuming that $\mu_u(\boldsymbol{x},\boldsymbol{z}) \neq 0$, can be estimated as

$$\widehat{\xi}_{\psi_\ell}(\boldsymbol{x},\boldsymbol{z}) = \frac{1}{3}\frac{\widehat{\partial m}_3(\boldsymbol{x},\boldsymbol{z})}{\partial \psi_\ell}\frac{\psi_\ell}{\widehat{m}_3(\boldsymbol{x},\boldsymbol{z})} \tag{6.16}$$

where $\widehat{m}_3(\boldsymbol{x},\boldsymbol{z})$ and $\widehat{\partial m}_3(\boldsymbol{x},\boldsymbol{z})/\partial\psi_\ell$, are the estimates from the local polynomial estimator and provided that $\widehat{m}_3(\boldsymbol{x},\boldsymbol{z}) \neq 0$ for the particular combination of interest $(\boldsymbol{x},\boldsymbol{z})$. Importantly, SVKZ also derived the asymptotic law for this elasticity estimator, showing that

$$(nh^{p+d+2})^{1/2}\big(\widehat{\xi}_{\psi_\ell}(\boldsymbol{x},\boldsymbol{z}) - \xi_{\psi_\ell}(\boldsymbol{x},\boldsymbol{z})\big) \longrightarrow N(0, s^2_{\xi_\ell}(\boldsymbol{x},\boldsymbol{z})), \tag{6.17}$$

In turn, these asymptotic results can be used for statistical testing about influence of elements in $(\boldsymbol{x},\boldsymbol{z})$ onto expected inefficiency.

A practical limitation of SVKZ is that the estimated production technology may not satisfy axioms of production. One might be tempted to follow Kuosmanen & Kortelainen (2012) or Parmeter & Racine (2012), imposing the desired constraints first, and then recovering $\widehat{E}[u|\boldsymbol{x},\boldsymbol{z}]$. However, as we noted earlier, the methods of Kuosmanen & Kortelainen (2012) and Parmeter & Racine (2012) work when the distribution of inefficiency is independent of $\boldsymbol{x}$ and $\boldsymbol{z}$, i.e. when u is homoskedastic. The issue the applied researcher faces here is much more subtle. When heteroskedasticity is present in u, one must recognize that what is being estimated in the first stage is a conditional mean, and **not** a production frontier. Thus, it is not necessarily the case that the axioms of production should be expected to hold when estimating the conditional mean.

Consider the case of a monotonic production function. The conditional mean of output could be non-monotonic if $\widehat{E}[u|\boldsymbol{x},\boldsymbol{z}]$ was non-monotonic, even though the production function is monotonic. Further, it is well known that adding two concave functions might not produce a concave function, so even if $\widehat{E}[u|\boldsymbol{x},\boldsymbol{z}]$ was concave, adding it to the production frontier may not produce a concave production function. And therein lies the danger of imposing constraints when estimating the conditional mean, it is not necessarily the case that they should be satisfied. This might seem innocuous except for the fact that imposing constraints on a conditional mean that are incorrect will not produce a consistent estimator and typically, consistent estimates in the first stage are needed for the second stage (recovering inefficiency) to produce valid estimates.

Take for example the discussion in Kuosmanen, Johnson & Saastamoinen (2015, pg. 233), who consider estimation of a production frontier nonparametrically, while also allowing u to depend on $\boldsymbol{x}$. In this case they stated (in our notation) "...Note that the shape of function g can differ from that of frontier m because $E(u_i|\boldsymbol{x}_i)$ is a function of inputs $\boldsymbol{x}$ ...It is also worth noting that function g is not necessarily monotonic increasing and concave even if the production function m satisfies these axioms because $-E(u_i|\boldsymbol{x}_i)$ can be a non-monotonic and non-concave function of inputs ...To apply CNLS in step 1, we need to assume that the curvature of the production function m dominates and that function g is monotonic increasing and concave (at least by approximation)." Unless the **conditional mean** of output satisfies the axioms of production, it is recommended the axiomatic restrictions be enforced **after** consistent, unrestricted estimation of the conditional mean as this will ensure that the first stage estimator of the conditional mean is consistent. How exactly to do this is a relatively unexplored area in stochastic frontier analysis and is a fruitful avenue for future research.

Figures 10.1-10.3 illustrate the pitfalls of enforcing constraints *ex ante* on the conditional mean of y (given $\boldsymbol{x}$). We have a single input, x, and our production frontier is logarithmic, which is naturally monotonic and concave. When inefficiency is homoskedastic we see that

the conditional mean is just a shift down of the production frontier, and remains both monotone and concave. However, if we allow heteroskedasticity of inefficiency, e.g. through a quadratic relationship, then, depending on the nature of heteroskedasticity, we can violate monotonicity, Figure 10.2, or concavity, Figure 10.3 of $E[y|x]$. This quadratic relationship is not beyond the pale, even in the parametric setting.[44]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

6.4. **Avoiding Distributional and (Some) Parametric Assumptions When Determinants of Inefficiency Are Present.** Here we discuss the approach of Tran & Tsionas (2009) and Parmeter, Wang & Kumbhakar (2017). Let the SFM be:

$$y_i = m(\boldsymbol{x}_i) + v_i - u_i = m(\boldsymbol{x}_i) + v_i - u_i + E[u_i|\boldsymbol{z}_i] - E[u_i|\boldsymbol{z}_i] = m^*(\boldsymbol{x}_i, \boldsymbol{z}_i) + \varepsilon_i^*. \tag{6.18}$$

where $m^*(\boldsymbol{x}_i, \boldsymbol{z}_i) = m(\boldsymbol{x}_i) + g(\boldsymbol{z}_i)$, $(u_i|\boldsymbol{z}_i = z) \sim D^+(\mu_u(x,z), \sigma_u^2(x,z))$, while $(v_i|\boldsymbol{x}_i, \boldsymbol{z}_i) \sim D(0, \sigma_v^2)$. This model is a special case of SVKZ's model. Now, if we specify our production technology as $m(\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$ and $E[u_i|\boldsymbol{z}_i] = g(\boldsymbol{z}_i)$, then if $\boldsymbol{\beta}$ were known, $g(\boldsymbol{z}_i)$ could be identified as the conditional mean of $\tilde{\varepsilon}_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$ given $\boldsymbol{z}_i$.

However, $\boldsymbol{\beta}$ is unknown and must be estimated. It can be estimated as follows. Conditioning only on $\boldsymbol{z}_i$ in equation (6.18) we have

$$E[y_i|\boldsymbol{z}_i] = E[\boldsymbol{x}_i|\boldsymbol{z}_i]'\boldsymbol{\beta} - g(\boldsymbol{z}_i). \tag{6.19}$$

Subtracting (6.19) from (6.18) yields

$$y_i - E[y_i|\boldsymbol{z}_i] = (\boldsymbol{x}_i - E[\boldsymbol{x}_i|\boldsymbol{z}_i])'\boldsymbol{\beta} + \varepsilon_i. \tag{6.20}$$

[44] Wang (2002) documents non-monotonic efficiency effects in a panel of Philippine rice farmers based on the age of the farmer.

If $E[y_i|\boldsymbol{z}_i]$ and $E[\boldsymbol{x}_i|\boldsymbol{z}_i]$ were known, $\boldsymbol{\beta}$ could be estimated via OLS from (6.20). The idea is to replace the unknown conditional means with their nonparametric estimates (Robinson 1988).

To estimate both $\boldsymbol{\beta}$ and $g(\boldsymbol{z}_i)$ we replace $E[y_i|\boldsymbol{z}_i]$ and $E[\boldsymbol{x}_i|\boldsymbol{z}_i]$ in (6.20) with

$$\hat{E}[y|\boldsymbol{z}_i] = \sum_{j=1}^{n} A_j(\boldsymbol{z}_i) y_j$$

$$\hat{E}[\boldsymbol{x}_s|\boldsymbol{z}_i] = \sum_{j=1}^{n} A_j(\boldsymbol{z}_i) \boldsymbol{x}_{sj},$$

For a given bandwidth, the conditional expectations for y and each element of $\boldsymbol{x}$ can be estimated and OLS can then be used to obtain a consistent estimator of $\boldsymbol{\beta}$. That is, instead of the usual regression of y on $\boldsymbol{x}$, one performs the modified OLS regression of $\tilde{y}$ on $\tilde{\boldsymbol{x}}$, where we have used the notation $\tilde{w} = w - \hat{E}[w|\boldsymbol{z}]$ to denote a random variable that has been conditionally demeaned. The estimates for $\boldsymbol{\beta}$ can then be used to obtain a consistent estimator of the conditional mean of inefficiency via standard nonparametric regression techniques.

Let $\check{\varepsilon}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is our estimate from the OLS regression of $\tilde{y}$ on $\tilde{\boldsymbol{x}}$. We then estimate $g(\boldsymbol{z}_i)$ nonparametrically via local-polynomial least-squares as

$$\widehat{g}(\boldsymbol{z}_i) = \sum_{j=1}^{n} A_j(\boldsymbol{z}_i) \check{\varepsilon}_j. \tag{6.21}$$

In the cross-sectional regression setting, without assuming some structure on the distributions of the error components, it is not possible to identify the impact that any given variable has on output directly, i.e. through the frontier, indirectly though inefficiency or both.[45] One way to achieve identification is through invocation of the separability assumption. This assumption, described in exceeding detail in Simar & Wilson (2007), essentially requires two distinct sets of variables: those which influence the frontier and those which solely influence inefficiency. In the context of a model for which two-sided noise does not exist (the standard

[45]Hall & Simar (2002) discussed nonparametric identification of the mean of inefficiency subject to the variance of the noise distribution diminishing as $n \to \infty$. Horrace & Parmeter (2011) showed how to nonparametrically identify the full distribution of inefficiency if one assumes that v is distributed normal.

DEA framework), when this assumption is satisfied, a two-step approach is available which can produce consistent estimators of both the frontier function and the inefficiency of a firm (Simar & Wilson 2007, Banker & Natarajan 2008, Simar & Wilson 2011).

In general it is recommended that if variables which influence inefficiency exist, that this information should be used directly, with a single stage estimator, such as maximum likelihood. When the separability assumption holds, then the partly linear model of Tran & Tsionas (2009) and Parmeter et al. (2017) could be deployed (albeit with some parametric assumptions imposed) or the additive model previously described can be used.[46]

Importantly, the separability assumption can be tested in the stochastic frontier context, including the fully nonparametric or semiparametric frameworks. We can compare the estimates from the additively separable SFM, with that from a fully nonparametric model to determine if there are statistical differences. Fortunately, this type of setup is conducive to inference through either a residual sum of squares test or a conditional moment test. See the discussion in chapter 6 of Henderson & Parmeter (2015*a*).

6.5. **Future Directions in Semi- and Nonparametric Estimation and Inference of the SFM.** One of the future directions of research within non- and semiparametric SFA is, naturally, related to statistical inference. The asymptotic results developed in the above mentioned papers as well as various testing procedures developed in the general statistics community make a solid foundation for this to happen, with careful adaptation and extensive

[46]The approach of SVKZ allows for both $\boldsymbol{x}$ and $\boldsymbol{z}$ to influence both the frontier and inefficiency and as such the separability assumption is not required. Yet, one may say that there is also a kind of 'separability' structure involved implicitly: $(\boldsymbol{x}, \boldsymbol{z})$ is assumed to influence the frontier via the first moment, while for the inefficiency term, u, the same $(\boldsymbol{x}, \boldsymbol{z})$ is modeled through the skedastic function defining the second moment. Besides helping with statistical identification, such structure can be viewed as quite natural to the context of measurement. Indeed, one often thinks of the frontier as the level, and so using the (conditional) first moment, measuring the (conditional) average level of outputs, would be very natural. Meanwhile, the inefficiency is often understood as the deviation from the frontier, so it would be a more natural way to model it with the second moment. In addition, one could also think of the inefficiency as a reflection of the uncertainty and related 'risk' to produce less than the potential and beyond the usual (and symmetric) noise, and it is very common to model risk through the second moment.

Monte Carlo evidence supporting the theory. Additionally, few of the methods discussed here have been fully developed in the panel data setting.

It is worth noting that neither Kumbhakar et al. (2007), nor Martins-Filho & Yao (2015), nor Park et al. (2015), nor SVKZ imposed any axioms of production on the frontier, e.g., monotonicity (i.e., require $\nabla m_{\boldsymbol{x}} \geq 0 \ \forall \boldsymbol{x}$), although some of them have brief discussions about possible extensions to do so. Specifically, to impose the desired constraints, one could adapt ideas from Daouia & Simar (2005) and Daouia & Park (2013), or use DEA or FDH on the fitted values from these methods (thus using the stochastic DEA or stochastic FDH approaches of Simar & Zelenyuk 2011), or to employ the constraint weighted bootstrapping (Hall & Huang 2001, Du et al. 2013), as was adapted to the baseline SFM by Parmeter & Racine (2012).

7. Quantile Estimation of the SFM

A recent development in the estimation of the SFM has been to embrace the use of quantile methods (Bernini, Freo & Gardini 2004, Know, Blankmeyer & Stutzman 2007, Liu, Laporte & Ferguson 2008, Behr 2010). Quantile regression is known to provide a more complete picture of a conditional distribution (Koenker & Hallock 2001, Koenker 2005) and provides a robust alternative to ordinary least squares. Whereas the ordinary least squares estimator stems from minimization of the sum of squared errors, the conditional quantile estimator is determined through minimization of the "check" function (Koenker & Bassett 1978) defined for a particular quantile, the median say.

The conditional quantile function $Q_y(\tau|\boldsymbol{x})$ for a random variable y with conditional CDF $F(y|\boldsymbol{x})$ is defined as $F^{-1}(\tau|\boldsymbol{x}) = \inf\{y : F(y|\boldsymbol{x}) \geq \tau\}$ where τ is the τ^{th} conditional quantile of the random variable y. Rather than directly inverting of the conditional distribution function, the conditional quantile can be determined through the loss function

$$\rho_\tau(\epsilon) = \epsilon\left(\tau - 1\{\epsilon < 0\}\right). \tag{7.1}$$

$\rho_\tau(\epsilon)$ is known as the check function. For a traditional linear in parameters framework, $Q_y(\tau|\boldsymbol{x}) = \boldsymbol{x}_i'\beta(\tau)$, the quantile estimator is found by minimizing

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau \left(y_i - \boldsymbol{x}_i'\beta(\tau)\right), \tag{7.2}$$

for a given τ. When the error terms are i.i.d., the conditional quantiles represent vertical shifts of the conditional median function by the appropriate quantile of the error distribution. However, when heteroskedasticity is present, the conditional quantiles are no longer vertical shifts of the conditional median, but will have varying slopes; moreover, the quantiles will become nonlinear.

The use of conditional quantile estimation to recover the frontier is appealing because in general a frontier can be thought of as a quantile in the distribution of output. At issue is the appropriate quantile, τ. For example, Bernini et al. (2004, pg. 379) estimate the frontier with the conditional quantile estimator using $\tau = 0.5$, 0.9 and 0.975. $\tau = 0.5$ corresponds to the median and is equivalent to the conditional mean in the case that $\sigma_u^2 = 0$ (see the discussion in Horrace & Parmeter 2014). Know et al. (2007, pg. 79) estimate conditional quantiles for $\tau = 0.85$, 0.9 and 0.95, while Liu et al. (2008, pg. 1080) consider $\tau = 0.5$ and 0.8. Lastly, Behr (2010, pg. 572) recommended use of $\tau = 0.95$ for estimation of production frontiers and $\tau = 0.05$ for estimation of cost frontiers.

What is lost in the recommendations of this earlier research is how one estimates (or predicts) individual efficiency once the frontier has been estimated. Currently the standard practice is to treat any firm whose output lies above the frontier as fully efficient, and any firm whose output is below the frontier as inefficient, with inefficiency defined as the difference between the estimated frontier and observed output. However, both of these recommendations ignore the fact that the composed error term represents inefficiency and noise. There does not exist at present an approach that separates inefficiency from noise in a manner similar to Jondrow et al. (1982). One idea could be to use the conditional mode

as proposed in Materov (1981). This estimator can be interpreted as a maximum likelihood estimator for the distribution of the joint density of v and u, and more importantly, for positive residuals, it is always 0, which is akin to how inefficiency is currently calculated using conditional quantile estimation. Unfortunately, as with the conditional mean, the conditional mode estimator requires distributional assumptions for it to be operational.

Lastly, we mention two important caveats with quantile estimation of frontiers. First, heteroskedasticity in either v or u has, to present, not been accounted for. This is a severe limitation as heteroskedasticity is commonly seen as present in v in applied efficiency studies, and researchers typically have access to an array of determinants of inefficiency, which induce heteroskedasticity in the inefficiency term. Moreover, unlike estimation of a conditional mean, when conditional heteroskedasticity is present, this can affect consistent estimation of the conditional quantile. Second, estimation of the conditional quantile for a specific value of τ is an implicit assumption on the ratio of signal to noise between σ_u^2 and σ_v^2. To see this, more clearly, Figures 10.4-10.7 presents the results of quantile estimation for $\tau = 0.5$, 0.8, 0.85, 0.9, and 0.95 for 1,000 observations drawn from the model

$$y_i = x_i^{0.4} e^{v_i - u_i}, \tag{7.3}$$

with $v_i \sim N(0,1)$ and $u_i \sim N_+(0, \sigma_u^2)$. In Figure 10.4 the inefficiency draws are taken with $\sigma_u^2 = 0.01$, in Figure 10.5 we have $\sigma_u^2 = 0.25$, in Figure 10.6 $\sigma_u^2 = 1$, and in Figure 10.7 $\sigma_u^2 = 4$. In the case where $\sigma_u^2 = 4$, this corresponds to a $\lambda = \sigma_u/\sigma_v = 2$ which is of decent size for an applied efficiency study. In this case the true frontier is approximately equal to the 85th quantile. It is clear that interpreting the frontier for a given quantile as the benchmark for a firm being efficient or inefficient is implicitly a statement on the ratio between the variance of the noise and the inefficiency for the sample. In Figure 10.4, where $\lambda = 0.01$, the setting where there is almost no inefficiency, the frontier is nearly equivalent to the median, which is the least absolute deviation estimator that Horrace & Parmeter (2014) discussed.

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

While the quantile estimator marks an interesting and robust alternative to traditional stochastic frontier analysis, it should be clear that more work needs to be done. We direct the reader to the earlier referenced papers for more details and additional insights on how best to use conditional quantile methods at present for conducting efficiency analysis. Furthermore, panel estimation of quantiles, as well as semi and non-parametric estimation of quantiles, is still in its infancy in this area and extensions to the SFM have as yet to appear in the literature.

8. Additional Approaches/Extensions of the SFM

As with any review or summary article, there is never enough space to cover all topics equally or broadly enough. The SFM has been studied and used for 40 years now and even though we have covered a range of approaches and insights, there are still many topics which we did not cover. These include finite mixture models (Caudill 2003, Orea & Kumbhakar 2004, Greene 2005*b*), the zero-inefficiency SFM (Kumbhakar, Parmeter & Tsionas 2013), the meta-frontier (Battese & Rao 2002, Battese, Rao & O'Donnell 2004), total factor productivity change and its individual components (Hulten 2001), the two-tier frontier (Polachek & Yoon 1987, Polachek & Yoon 1996), sample selection in the SFM (Kumbhakar, Tsionas & Sipiläinen 2009, Greene 2010), and directional distance function estimation (Atkinson & Tsionas 2016). Parmeter & Kumbhakar (2014) cover broadly estimation and inference of finite mixture models, the zero-inefficiency SFM and issues pertaining to sample selection. Full details on the measurement of total factor productivity and separation into distinct components can be found in Kumbhakar et al. (2015,

chapt. 11). Both the two-tier frontier (Kumbhakar & Parmeter 2009, Kumbhakar & Parmeter 2010, Papadopoulos 2015) and meta-frontier (O'Donnell, Rao & Battese 2008, Amsler, O'Donnell & Schmidt 2017) have started to receive more attention recently, but as of yet no broad review of either exists. Regarding the estimation of directional distance functions, we refer interested readers to Färe, Martins-Filho & Vardanyan (2010) for a thorough treatment.

9. Available Software to Estimate SFMs

Despite the popularity of the SFM, only the most basic implementations of it are available across a wide array of statistical platforms. For example, in the `R` programming environment the `frontier` (Coelli & Henningsen 2013) package allows for cross-sectional estimation of the SFM assuming either the half-normal or truncated-normal distribution for u_i and the Battese & Coelli (1992) and Battese & Coelli (1995) panel data estimators of the SFM are implemented.[47] There are similar estimators available in LIMDEP through the NLOGIT module but also include the normal-gamma specification as well as the true fixed and true random effects estimators along with the latent class stochastic frontier estimator. There are also several modules in the STATA software as described in Kumbhakar et al. (2015) which implement several other panel data estimators as described earlier. Additionally, many authors provide their own personal codes.

However, there does not yet exist a singular software that implements all of the available estimators described here. This should not be surprising. As with any applied field, as statistical improvements are made, there is a lag with available software and the array of options makes it infeasible to include all discussed models in a singular package. Researchers interested in the newest methods can invest in programming these methods and disseminating

[47]The `frontier` package accesses the Frontier V4.1 Fortran codes originally developed by Tim Coelli, which is also freely available (at http://www.uq.edu.au/economics/cepa/frontier.php), although fairly outdated by now.

them to the field, or can collaborate with the authors of the original models to develop software that can be made widely available, and we strongly encourage researchers to do so.

10. Conclusions

This review was meant to highlight some of the most important econometric developments over the past 40 years to improve the estimation of measurements of productivity and efficiency. We covered the workhorse SFM, and discussed avenues to include determinants of inefficiency and productivity, dealing with endogeneity, what to do when one has panel data, quantile estimation, and robust methods involving nonparametric regression and local-likelihood to place as few restrictions as possible on the frontier or the behavior of inefficiency. All told, a variety of methods and models exist for the practitioner and our hope is that this review will encourage applied researchers to move off of some of the basic SFMs in search of more robust and insightful conclusions.

While much has been covered, much remains unsaid. Important areas that are still being developed include modeling dependence between statistical noise and inefficiency, selection of firm technology, handling heterogeneous technology in a sample of firms, and how to allow a subset of firms to be fully efficient. While our discussion was couched in terms of the single equation stochastic production frontier, system based approaches surrounding cost, profit, or revenue frontiers are also available and, as the other methods that we mentioned without any details, they deserve attention and separate reviews.

References

Afriat, S. N. (1972), 'Efficiency estimation of production functions', *International Economic Review* **13**(3), 568–598.

Ahmad, I. A. & Li, Q. (1997), 'Testing symmetry of an unknown density function by kernel method', *Journal of Nonparametric Statistics* **7**, 279–293.

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.

Aitchison, J. & Aitken, C. (1976), 'Multivariate binary discrimination by the kernel method', *Biometrika* **63**, 413–420.

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.

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.

Amemiya, T. (1974), 'The nonlinear two-stage least-squares estimator', *Journal of Econometrics* **2**, 105–111.

Amsler, C., O'Donnell, C. J. & Schmidt, P. (2017), 'Stochastic metafrontiers', *Econometric Reviews* **36**, 1007–1020.

Amsler, C., Prokhorov, A. & Schmidt, P. (2016), 'Endogeneity in stochastic frontier models', *Journal of Econometrics* **190**, 280–288.

Amsler, C., Prokhorov, A. & Schmidt, P. (2017), 'Endogeneity environmental variables in stochastic frontier models', *Journal of Econometrics* **199**, 131–140.

Atkinson, S. E. & Tsionas, E. G. (2016), 'Direcitonal disance functions: Optimal endogenous directions', *Journal of Econometrics* **190**, 301–314.

Azzalini, A. (1985), 'A class of distributions which includes the normal ones', *Scandinavian Journal of Statistics* **12**(2), 171–178.

Badunenko, O. & Kumbhakar, S. C. (2017), 'Economies of scale, technical change and persistent and time-varying cost efficiency in Indian banking: Do ownership, regulation and heterogeneity matter?', *European Journal of Operational Research* **260**, 789–803.

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.

Banker, R. D. & Natarajan, R. (2008), 'Evaluating contextual variables affecting productivity using data envelopment analysis', *Operations Research* **56**(1), 48–58.

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.

Battese, G. E. & Rao, D. S. P. (2002), 'Technology gap, efficiency and a stochastic metafrontier function', *International Journal of Business and Economics* **1**, 1–7.

Battese, G. E., Rao, D. S. P. & O'Donnell, C. J. (2004), 'A metafrontier production function for estimation of technical efficiencies and technology gaps for firms operating under different technologies', *Journal of Productivity Analysis* **21**, 91–103.

Behr, A. (2010), 'Quantile regression for robust bank efficiency score estimation', *European Journal of Operational Research* **200**, 568–581.

Benabou, R. & Tirole, J. (2016), 'Mindful economics: The production, consumption, and value of beliefs', *Journal of Economic Perspectives* **30**(3), 141–164.

Bera, A. K. & Sharma, S. C. (1999), 'Estimating production uncertainty in stochastic frontier production function models', *Journal of Productivity Analysis* **12**(2), 187–210.

Bernini, C., Freo, M. & Gardini, A. (2004), 'Quantile estimation of frontier production function', *Empirical Economics* **29**, 373–381.

Bloom, N., Lemos, R., Sadun, R., Scur, D. & Van Reenen, J. (2016), 'International data on measuring management practices', *American Economic Review* **106**(5), 152–156.

Bonanno, G., De Giovanni, D. & Domma, F. (2017), 'The 'wrong skewness' problem: a re-specification of stochastic frontiers', *Journal of Productivity Analysis* **47**(1), 49–64.

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.

Case, B., Ferrari, A. & Zhao, T. (2013), 'Regulatory reform and productivity change in indian banking', *The Review of Economics and Statistics* **95**(3), 1066–1077.

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 measures in the presence of heteroskedasticity', *Journal of Business & Economic Statistics* **13**(1), 105–111.

Chamberlain, G. (1987), 'Asymptotic efficiency in estimation with conditional moment restrictions', *Journal of Econometrics* **34**(2), 305–334.

Chen, L.-H., Cheng, M.-Y. & Peng, L. (2009), 'Conditional variance estimation in heteroscedastic regression models', *Journal of Statistical Planning and Inference* **139**(2), 236–245.

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.

Chu, C.-Y., Henderson, D. J. & Parmeter, C. F. (2017), 'On discrete Epanechnikov kernels', *Computational Statistics and Data Analysis* . forthcoming.

Coelli, T. & Henningsen, A. (2013), *frontier: Stochastic Frontier Analysis*. R package version 1.1-0.
URL: *http://CRAN.R-Project.org/package=frontier*

Coelli, T. J. (1995), 'Estimators and hypothesis tests for a stochastic frontier function: A Monte Carlo analysis', *Journal of Productivity Analysis* **6**(4), 247–268.

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.

Daouia, A. & Park, B. U. (2013), 'On projection-type estimators of multivariate isotonic functions', *Scandinavian Journal of Statistics* **40**, 363–386.

Daouia, A. & Simar, L. (2005), 'Robust nonparametric estimators of monotone boundaries', *Journal of Multivariate Analysis* **96**(2), 311–331.

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. & Gijbels, I. (1996), *Local Polynomial Modelling and its Application*, Chapman and Hall.

Fan, J. & Yao, Q. (1998), 'Efficient estimation of conditional variance functions in stochastic regression', *Biometrika* **85**, 645–660.

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., Martins-Filho, C. & Vardanyan, M. (2010), 'On functional form representation of multi-output production technologies', *Journal of Productivity Analysis* **33**(1), 81–96.

Feng, Q., Horrace, W. C. & Wu, G. L. (2015), Wrong skewness and finite sample correction in parametric stochastic frontier models. Center for Policy Research - The Maxwell School, working paper N. 154.

Gabrielsen, A. (1975), On estimating efficient production functions. Working Paper No. A-85, Chr. Michelsen Institute, Department of Humanities and Social Sciences, Bergen, Norway.

Gagnepain, P. & Ivaldi, M. (2002), 'Stochastic frontiers and asymmetric information models', *Journal of Productivity Analysis* **18**(2), 145–159.

Greene, W. (2004), 'Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems', *Health Economics* **13**(9), 959–980.

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. (2008), The econometric approach to efficiency analysis, *in* C. A. K. L. H. O. Fried & S. S. Schmidt, eds, 'The Measurement of Productive Efficiency and Productivity Change', Oxford University Press, Oxford, United Kingdom, chapter 2.

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., Manner, H. & Simar, L. (2016), 'The "wrong skewness" problem in stochastic frontier model: a new approach', *Econometric Reviews* . forthcoming.

Hall, P. & Huang, H. (2001), 'Nonparametric kernel regression subject to monotonicity constraints', *The Annals of Statistics* **29**(3), 624–647.

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.

Hansen, C., McDonald, J. B. & Newey, W. K. (2010), 'Instrumental variables estimation with flexible distributions', *Journal of Business and Economic Statistics* **28**, 13–25.

Hattori, T. (2002), 'Relative performance of U.S. and Japanese electricity distribution: An application of stochastic frontier analysis', *Journal of Productivity Analysis* **18**(3), 269–284.

Henderson, D. J. & Parmeter, C. F. (2015*a*), *Applied Nonparametric Econometrics*, Cambridge University Press, Cambridge, Great Britain.

Henderson, D. J. & Parmeter, C. F. (2015*b*), 'A consistent bootstrap procedure for nonparametric symmetry tests', *Economics Letters* **131**, 78–82.

Hjalmarsson, L., Kumbhakar, S. C. & Heshmati, A. (1996), 'DEA, DFA, and SFA: A comparison', *Journal of Productivity Analysis* **7**(2), 303–327.

Hollingsworth, B. (2008), 'The measurement of efficiency and productivity of health care delivery', *Health Economics* **17**(10), 1107–1128.

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. & Wright, I. A. (2016), Stationary points for parametric stochastic frontier models. Center for Policy Research - The Maxwell School, working paper N. 196.

Huang, C. J. & Liu, J.-T. (1994), 'Estimation of a non-neutral stochastic frontier production function', *Journal of Productivity Analysis* **5**(1), 171–180.

Hulten, C. R. (2001), Total factor productivity. a short biography, *in* C. R. Hulten, E. R. Dean & M. J. Harper, eds, 'New Developments in Productivity Analysis', University of Chicago Press, Chicago, IL, pp. 1–54.

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.

Karakplan, M. U. & Kutlu, L. (2013), Handling endogeneity in stochastic frontier analysis. Unpublished manuscript.

Kim, M. & Schmidt, P. (2008), 'Valid test of whether technical inefficiency depends on firm characteristics', *Journal of Econometrics* **144**(2), 409–427.

Kneip, A. & Simar, L. (1996), 'A general framework for frontier estimation with panel data', *Journal of Productivity Analysis* **7**(2), 187–212.

Kneip, A., Simar, L. & Van Keilegom, I. (2015), 'Frontier estimation in the presence of measurement error with unknown variance', *Journal of Econometrics* **184**, 379–393.

Knittel, C. R. (2002), 'Alternative regulatory methods and firm efficiency: Stochastic frontier evidence form the U.S. electricity industry', *The Review of Economics and Statistics* **84**(3), 530–540.

Know, K. J., Blankmeyer, E. C. & Stutzman, J. R. (2007), 'Technical efficiency in Texan nursing facilities: a stochastic production frontier approach', *Journal of Economics and Finance* **31**(1), 75–86.

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press.

Koenker, R. & Bassett, G. (1978), 'Regression quantiles', *Econometrica* **46**(1), 33–50.

Koenker, R. & Hallock, K. (2001), 'Quantile regression', *Journal of Economic Perspectives* **15**, 143–156.

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. (2011), 'Estimation of production technology when the objective is to maximize return to the outlay', *European Journal of Operational Research* **208**, 170–176.

Kumbhakar, S. C. (2013), 'Specification and estimation of multiple output technologies: A primal approach', *European Journal of Operational Research* **231**, 465–473.

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. (2009), 'The effects of match uncertainty and bargaining on labor market outcomes: evidence from firm and worker specific estimates', *Journal of Productivity Analysis* **31**(1), 1–14.

Kumbhakar, S. C. & Parmeter, C. F. (2010), 'Estimation of hedonic price functions with incomplete information', *Empirical Economics* **39**(1), 1–25.

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., 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. (2015), *A Practitioners Guide to Stochastic Frontier Analysis Using Stata*, Cambridge University Press, Cambridge, United Kingdom.

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. & Saastamoinen, A. (2015), Stochastic nonparametric approach to efficiency analysis: A unified framework, *in* J. Zhu, ed., 'Data Envelopment Analysis', International Series in Operations Research & Management Science, Springer Science, New York., chapter 7, pp. 191–244.

Kuosmanen, T. & Kortelainen, M. (2012), 'Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints', *Journal of Productivity Analysis* **38**(1), 11–28.

Kutlu, L. (2010), 'Battese-Coelli estimator with endogenous regressors', *Economics Letters* **109**, 79–81.

Latruffe, L., Bravo-Ureta, B. E., Carpentier, A., Desjeux, Y. & Moreira, V. H. (2017), 'Subsidies and technical efficiency in agriculture: Evidence from European dairy farms', *American Journal of Agricultural Economics* **99**, 783–799.

Lee, L. (1983), 'A test for distributional assumptions for the stochastic frontier function', *Journal of Econometrics* **22**(2), 245–267.

Lee, L.-F. & Tyler, W. G. (1978), 'The stochastic frontier production function and average efficiency: An empirical analysis', *Journal of Econometrics* **7**, 385–389.

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.

Li, D., Simar, L. & Zelenyuk, V. (2016), 'Generalized nonparametric smoothing with mixed discrete and continuous data', *Computational Statistics and Data Analysis* **100**, 424–444.

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.

Lien, G., Kumbhakar, S. C. & Hardaker, J. B. (2017), 'Accounting for risk in productivity analysis: an application to Norwegian dairy farming', *Journal of Productivity Analysis* **47**(3), 247–257.

Liu, C., Laporte, A. & Ferguson, B. S. (2008), 'The quantile regression approach to efficiency measurement: Insights from Monte Carlo simulations', *Health Economics* **17**, 1073–1087.

Lovell, C. A. K. (1993), Production frontiers and productive efficiency, *in* C. A. K. L. H. O. Fried & S. S. Schmidt, eds, 'The Measurement of Productive Efficiency', Oxford University Press, Oxford, United Kingdom, chapter 1.

Martins-Filho, C. B. & Yao, F. (2015), 'Semiparametric stochastic frontier estimation via profile likelihood', *Econometric Reviews* **34**(4), 413–451.

Materov, I. S. (1981), 'On full identification of the stochastic production frontier model (in Russian)', *Ekonomika i Matematicheskie Metody* **17**, 784–788.

McFadden, D. (1989), 'A method of simulated moments for estimation of discrete response models without numerical integration', *Econometrica* **57**(5), 995–1026.

Meeusen, W. & van den Broeck, J. (1977*a*), 'Efficiency estimation from Cobb-Douglas production functions with composed error', *International Economic Review* **18**(2), 435–444.

Meeusen, W. & van den Broeck, J. (1977*b*), 'Technical efficiency and dimension of the firm: Some results on the use of frontier production functions', *Empirical Economics* **2**(2), 109–122.

Mundlak, Y. (1961), 'Empirical production function free of management bias', *Journal of Farm Economics* **43**(1), 44–56.

Mutter, R. L., Greene, W. H., Spector, W., Rosko, M. D. & Mukamel, D. B. (2013), 'Investigating the impact of endogeneity on inefficiency estimates in the application of stochastic frontier analysis to nursing homes', *Journal of Productivity Analysis* **39**(1), 101–110.

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.

Noh, H. (2014), 'Frontier estimation using kernel smoothing estimators with data transformation', *Journal of the Korean Statistical Society* **43**, 503–512.

O'Donnell, C. J., Rao, D. S. P. & Battese, G. E. (2008), 'Metafrontier frameworks for the study of firm-level efficiencies and technology ratios', *Empirical Economics* **34**, 231–255.

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.

Papadopoulos, A. (2015), 'The half-normal specification for the two-tier stochastic frontier model', *Journal of Productivity Analysis* **43**(2), 225–230.

Park, B. U., Simar, L. & Zelenyuk, V. (2015), 'Categorical data in local maximum likelihood: theory and applications to productivity analysis', *Journal of Productivity Analysis* **43**(1), 199–214.

Parmeter, C. F. & Kumbhakar, S. C. (2014), 'Efficiency Analysis: A Primer on Recent Advances', *Foundations and Trends in Econometrics* **7**(3-4), 191–385.

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., Wang, H.-J. & Kumbhakar, S. C. (2017), 'Nonparametric estimation of the determinants of inefficiency', *Journal of Productivity Analysis* **47**(3), 205–221.

Parmeter, C. F. & Zelenyuk, V. (2016), A bridge too far? the state of the art in combining the virtues of stochastic frontier analysis and data envelopment analysis. University of Miami Working Paper 2016-10.

Paul, S. & Shankar, S. (2017), An alternative specification for technical efficiency effects in a stochastic frontier production function. Crawford School Working Paper 1703.

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.

Polachek, S. W. & Yoon, B. J. (1987), 'A two-tiered earnings frontier estimation of employer and employee information in the labor market', *The Review of Economics and Statistics* **69**(2), 296–302.

Polachek, S. W. & Yoon, B. J. (1996), 'Panel estimates of a two-tiered earnings frontier', *Journal of Applied Econometrics* **11**(2), 169–178.

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.

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.

Schmidt, P. (1976), 'On the statistical estimation of parametric frontier production functions', *The Review of Economics and Statistics* **58**(2), 238–239.

Schmidt, P. (2011), 'One-step and two-step estimation in SFA models', *Journal of Productivity Analysis* **36**(2), 201–203.

Schmidt, P. & Sickles, R. C. (1984), 'Production frontiers and panel data', *Journal of Business & Economic Statistics* **2**(2), 367–374.

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. (2017), 'Nonparametric least squares methods for stochastic frontier models', *Journal of Productivity Analysis* **47**(3), 189–204.

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. (2011), 'Two-stage DEA: Caveat emptor', *Journal of Productivity Analysis* **36**(2), 205–218.

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.

Simar, L. & Wilson, P. W. (2015), 'Statistical approaches for nonparametric frontier models: A guided tour', *International Statistical Review* **83**(1), 77–110.

Simar, L. & Zelenyuk, V. (2011), 'Stochastic FDH/DEA estimators for frontier analysis', *Journal of Productivity Analysis* **36**(1), 1–20.

Solow, R. (1957), 'Technical change and the aggregate production function', *The Review of Economics and Statistics* **39**(3), 312–320.

Stevenson, R. (1980), 'Likelihood functions for generalized stochastic frontier estimation', *Journal of Econometrics* **13**(1), 58–66.

Stiglitz, J. E. & Greenwald, B. C. (1986), 'Externalities in economies with imperfect information and incomplete markets', *Quarterly Journal of Economics* **101**(2), 229–264.

Taube, R. (1988), Möglichkeiten der effizienzmess ung von öffentlichen verwaltungen. Duncker & Humbolt GmbH, Berlin.

Tibshirani, R. & Hastie, T. (1987), 'Local likelihood estimation', *Journal of the American Statistical Association* **82**, 559–568.

Timmer, C. P. (1971), 'Using a probabilistic frontier production function to measure technical efficiency', *The Journal of Political Economy* **79**(4), 776–794.

Tran, K. C. & Tsionas, E. G. (2009), 'Estimation of nonparametric inefficiency effects stochastic frontier models with an application to British manufacturing', *Economic Modelling* **26**, 904–909.

Tran, K. C. & Tsionas, E. G. (2013), 'GMM estimation of stochastic frontier models with endogenous regressors', *Economics Letters* **118**, 233–236.

Tsionas, E. G. (2007), 'Efficiency measurement with the Weibull stochastic frontier', *Oxford Bulletin of Economics and Statistics* **69**(5), 693–706.

Tsionas, E. G. (2012), 'Maximum likelihood estimation of stochastic frontier models by the Fourier transform', *Journal of Econometrics* **170**(2), 234–248.

Uekusa, M. & Torii, A. (1985), 'Stochastic production functions: An application to Japanese manufacturing industry (in Japanese)', *Keizaigaku Ronsyu (Journal of Economics)* **51**(1), 2–23.

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.

Winsten, C. B. (1957), 'Discussion on Mr. Farrell's paper', *Journal of the Royal Statistical Society Series A, General* **120**(3), 282–284.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edn, MIT Press, Cambridge, Massachusetts.

FIGURE 10.1. Concave and Monotonic Conditional Mean and Production Frontier Under Homoskedastic Inefficiency. The solid line is the production frontier while the dashed line is the conditional mean of output.

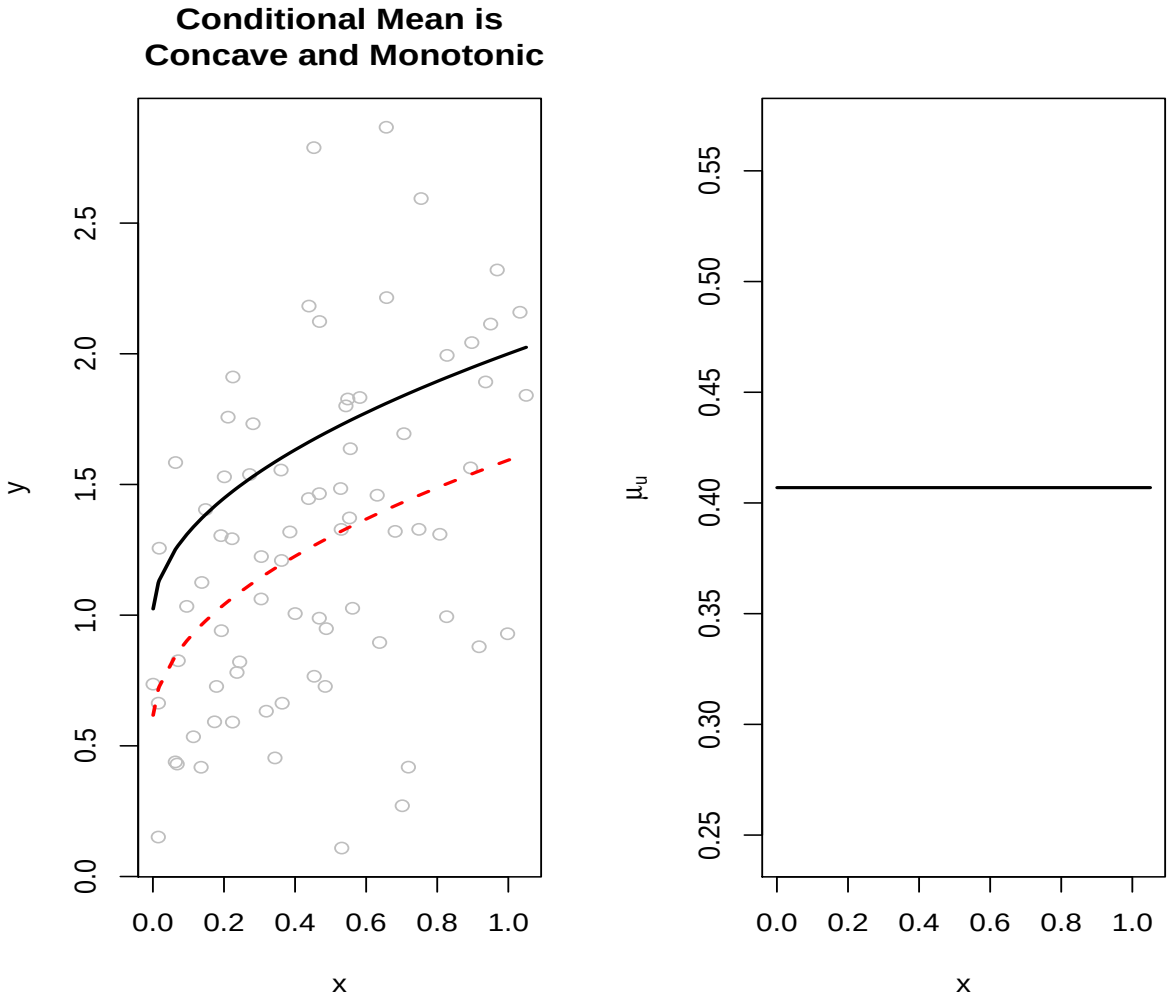

FIGURE 10.2. Concave but Non Monotonic Conditional Mean and Production Frontier Under Heteroskedastic Inefficiency. The solid line is the production frontier while the dashed line is the conditional mean of output.

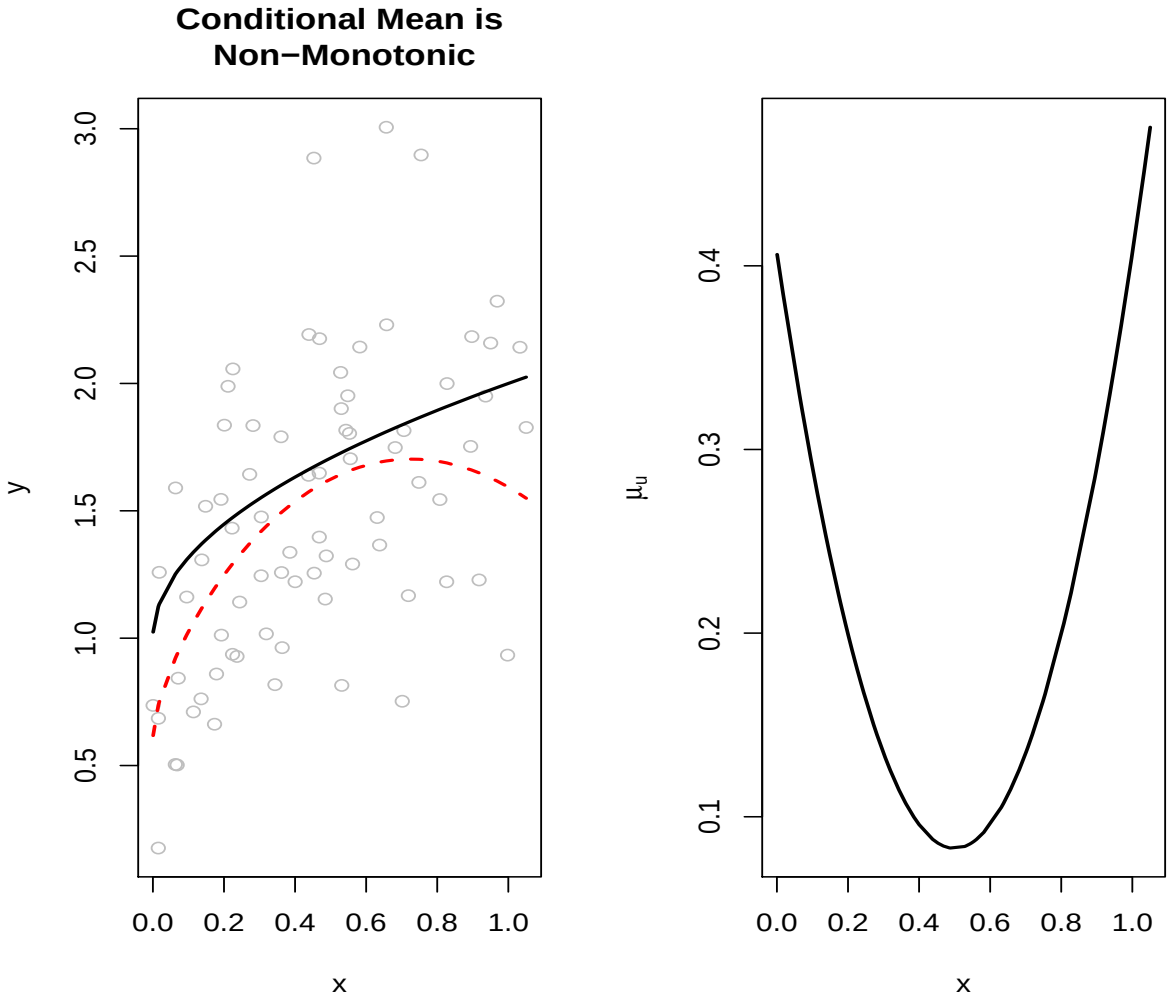

FIGURE 10.3. Monotonic but Non Concave Conditional Mean and Production Frontier Under Heteroskedastic Inefficiency. The solid line is the production frontier while the dashed line is the conditional mean of output.

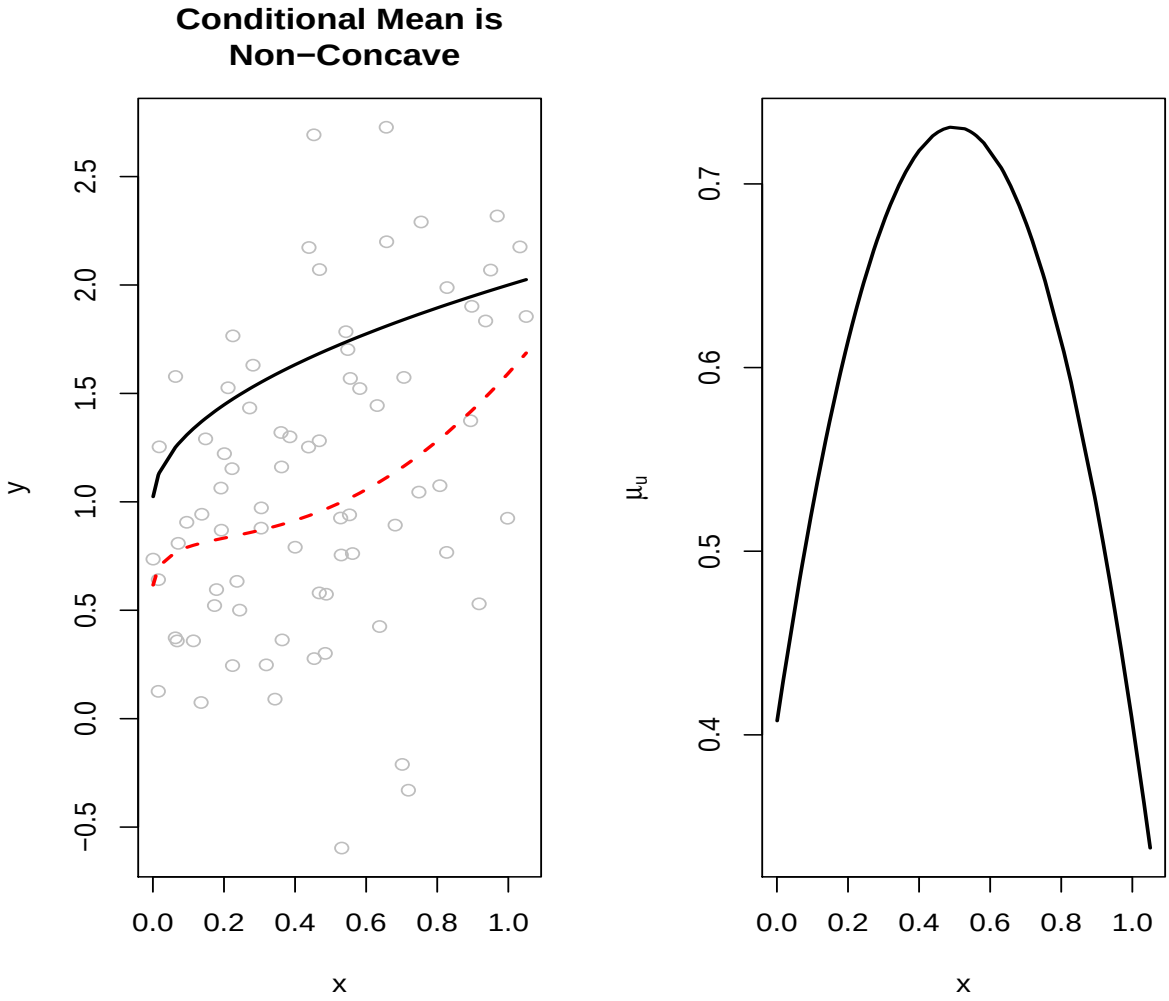

FIGURE 10.4. Conditional quantile estimation of a univariate SFM with $\sigma_u^2 = 0.01$.

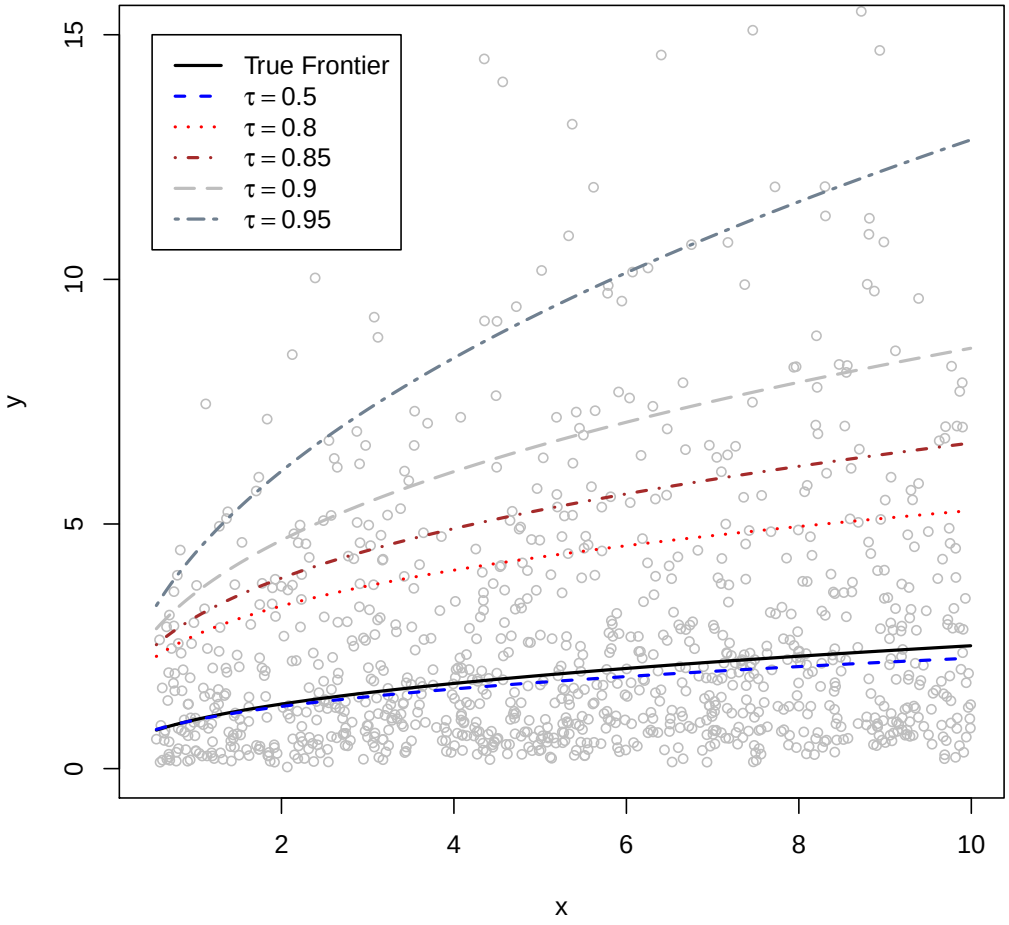

FIGURE 10.5. Conditional quantile estimation of a univariate SFM with $\sigma_u^2 = 0.25$.

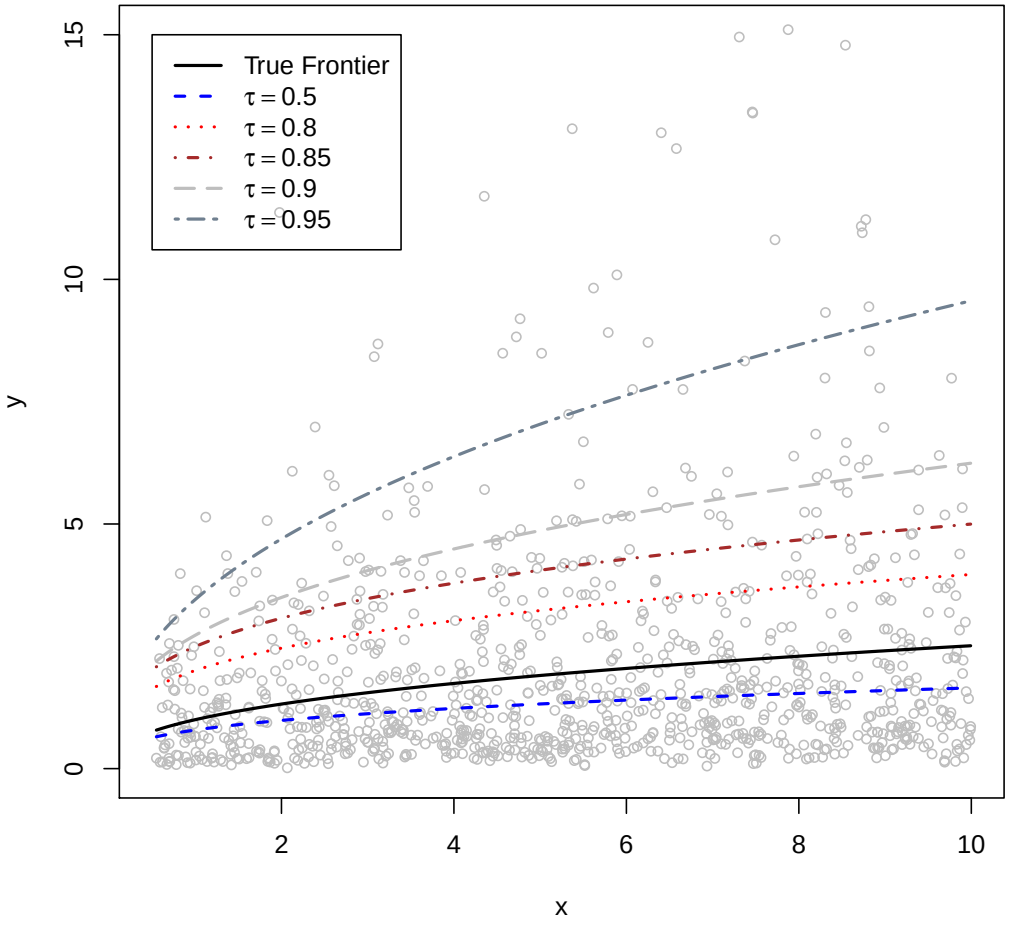

FIGURE 10.6. Conditional quantile estimation of a univariate SFM with $\sigma_u^2 = 1$.

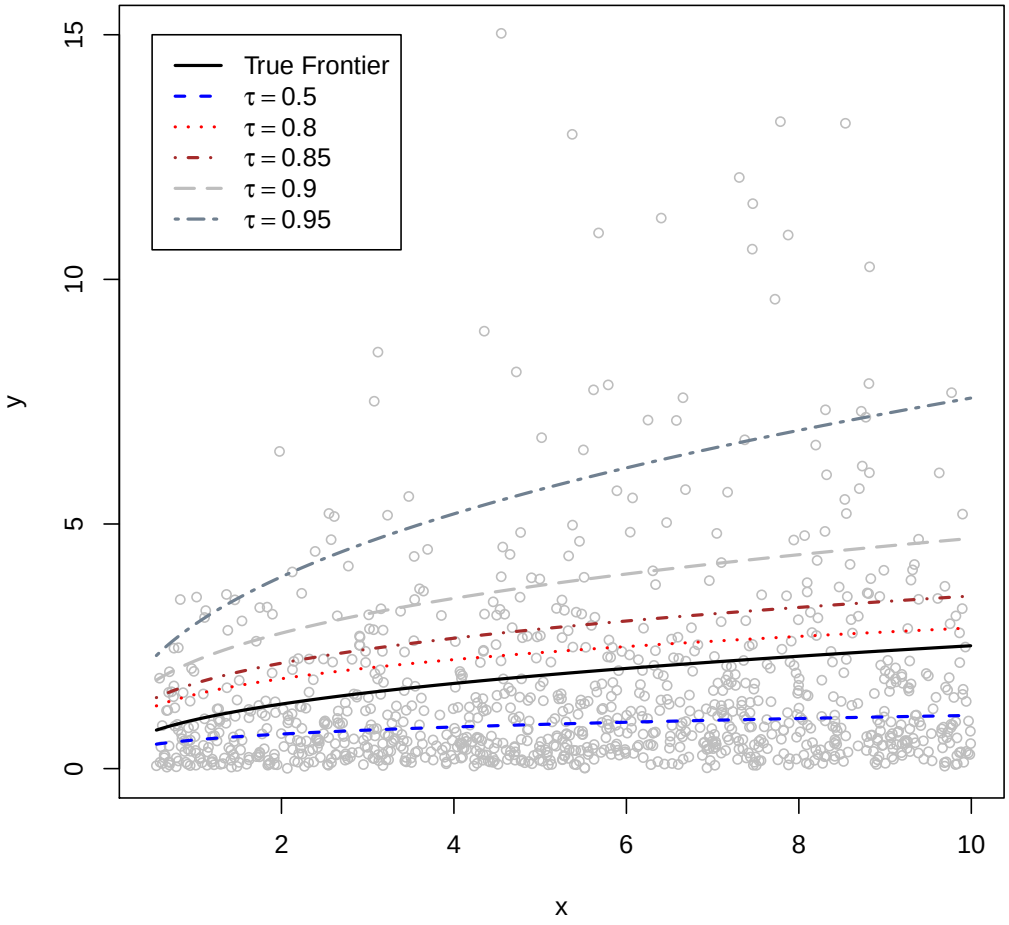

FIGURE 10.7. Conditional quantile estimation of a univariate SFM with $\sigma_u^2 = 4$.

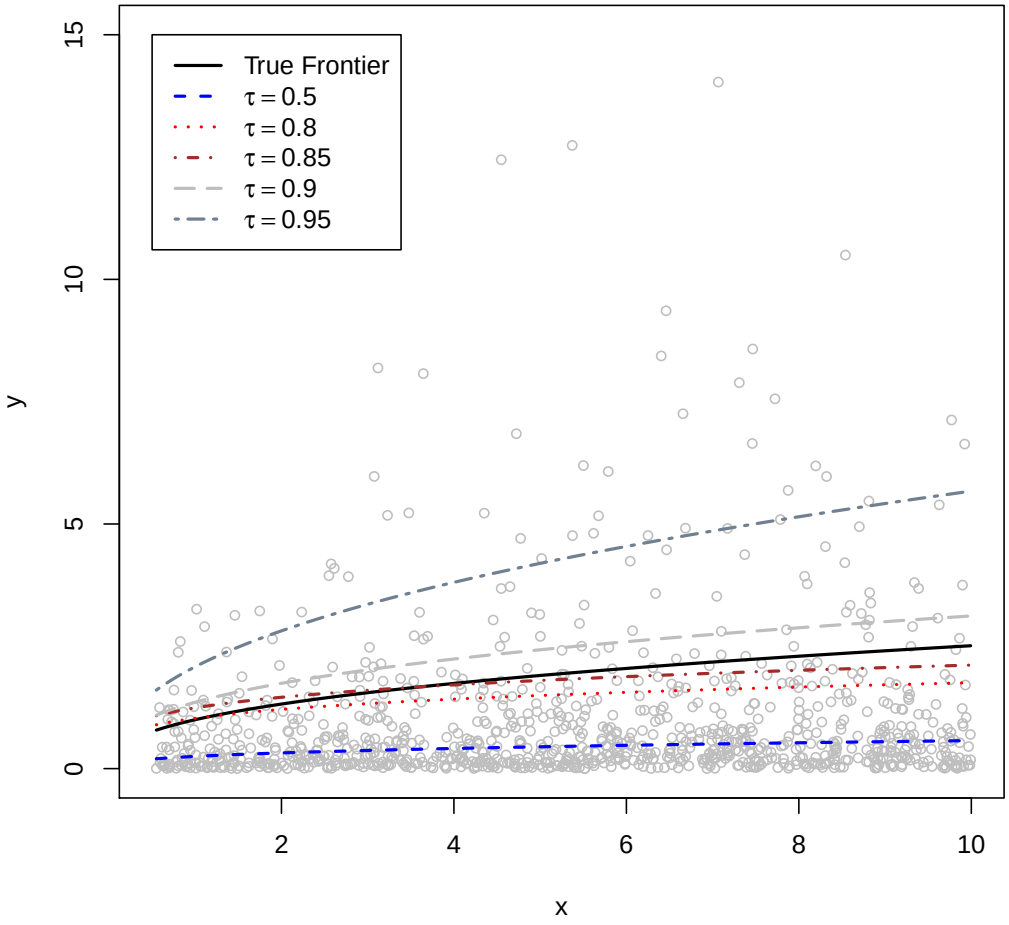

TABLE 1. Right tail critical values for both a χ^2_1 and a 50:50 mixture of a χ^2_0 and a χ^2_1, denoted as $\bar{\chi}^2$.

Significance Level	0.01	0.05	0.1	0.15	0.2	0.25
χ^2_1	6.634	3.841	2.706	2.072	1.642	1.323
$\bar{\chi}^2$	5.412	2.706	1.642	1.074	0.708	0.455