Efficiency Analysis with Stochastic Frontier Models using Popular Statistical Softwares

Bao Hoang Nguyen, Robin C. Sickles, Valentin Zelenyuk

Date: June 2021

School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia

ISSN No. 1932 - 4398
Efficiency Analysis with Stochastic Frontier Models using Popular Statistical Softwares

Bao Hoang Nguyen*  Robin C. Sickles† Valentin Zelenyuk ‡

June 15, 2021

Abstract

Our chapter provides a brief introduction to the stochastic frontier paradigm – one of the most powerful techniques for performance analysis developed over the last few decades to address various research questions for many contexts with empirical applications in a wide variety of economic sectors such as banking, healthcare, agriculture, and so on. We also document the estimation routines used to implement the classical models as well as the recent developments in this research area for practitioners, especially those who are willing to use Stata, but also with tips on sources for R and Matlab users.

Keywords: Technical efficiency, Stochastic frontier analysis, Panel data, Semi-parametric, Stata, Matlab, R.

*School of Economics, University of Queensland, Brisbane, Qld 4072, Australia
†Economics Department, Rice University, Houston, TX 77251-1892, USA
‡School of Economics and Centre for Efficiency and Productivity Analysis, University of Queensland, Brisbane, Qld 4072, Australia
1 Introduction

In this chapter, we provide practitioners, who are interested in analysing the performance of production units, with a brief introduction to the stochastic frontier paradigm – one of the most powerful techniques for performance analysis developed in the last century. Stochastic frontier analysis employs econometric models to estimate production frontiers and technical (in)efficiency with respect to these frontiers. Since its first introduction by Aigner et al. (1977) and Meensen and van den Broeck (1977), stochastic frontier analysis has been applied to study the productivity and efficiency of production units in various economic sectors, such as banking (e.g., Ferrier and Lovell (1990); Adams et al. (1999); Kumbhakar and Tsionas (2005); Malikov et al. (2016)), healthcare (e.g., Zuckerman et al. (1994); Rosko (2001); Greene (2004); Mutter et al. (2013); Comans et al. (2020)), agriculture (e.g., Battese and Coelli (1995); Battese and Broca (1997); Kumbhakar and Tsionas (2008)), to mention a few. Moreover, the methodology is also used to undertake cross-country studies on various important aspects of society such as the healthcare system (Greene, 2004) and taxation (Fenochietto and Pessino, 2013).

Our chapter also documents the estimation routines used to implement the classical models as well as the recent developments in this research area for practitioners, especially those who are willing to use Stata, but also with tips on where to find analogous programs for R and Matlab users. Interested readers can find more comprehensive overviews in Sickles and Zelenyuk (2019, Chap. 11-16), Kumbhakar et al. (2021a) and Kumbhakar et al. (2021b).

The structure of this chapter is as follows. We start our discussion with the basic stochastic frontier model. We then extend our discussion to various generalisations of the stochastic frontier paradigm, including stochastic panel data models, stochastic frontier models with determinants of inefficiency, also referred to in the literature as “environmental factors”, and the semi-parametric stochastic frontier models. To provide readers with an accessible toolkit to implement these methods, we also document available command-
s/packages in popular statistical softwares. We focus on the implementation via Stata, and also provide brief comments on the sources on analogous implementations in Matlab and R. Besides, we also provide an empirical illustration for the methods discussed in this chapter.

2 Basic Stochastic Frontier Models

The stochastic frontier paradigm can be viewed as a generalization of the classical production function approach, where the optimal allocation in production is a testable restriction rather than a prior assumption usually assumed by the neoclassical production theory (Sickles and Zelenyuk, 2019).

The distinctive feature of the stochastic frontier paradigm (compared to the canonical average production model paradigm) is its non-symmetric two-component error, composed of a regular idiosyncratic disturbance and an additional one-sided non-negative error component.\(^3\) The former accounts for factors such as measurement error, misspecification, and the randomness of the production process, whereas the latter aims to represent the technical inefficiency that reduces the actual output from its maximum feasible level.\(^4\) Assumptions in the canonical model used in stochastic frontier analysis on the conditional independence of both error terms and the regressors as well as their independence from each other have been lifted over the years in a series of refinements of the basic model. We will discuss these in turn later in our chapter.

2.1 Aigner et al. (1977) Model

The canonical model of the stochastic frontier paradigm was proposed independently by Aigner et al. (1977) (hereafter ALS) and Meeusen and van den Broeck (1977). The ALS

\(^3\)In the panel data context, which we will discuss in the next sections, the composed error can include four components.

\(^4\)In this chapter, our discussion will follow the traditional exposition based on the production function. Similar exposition (with some adaptations) applies to other characterizations of the production side, such as cost function and revenue function. Meanwhile, more elaboration is needed if one is interested in measuring profit efficiency (see Färe et al. (2019), Sickles and Zelenyuk (2019, Chapter 2) and references therein).
model is formulated as

$$\ln y_i = \ln f (x_i | \beta) + \varepsilon_i, \quad i = 1, \ldots, n,$$

$$\varepsilon_i = v_i - u_i,$$

$$v_i \sim \text{iid } N \left(0, \sigma^2_v\right),$$

$$u_i \sim \text{iid } N^+ \left(0, \sigma^2_u\right),$$

where $$y_i \in \mathbb{R}_+^1$$ is the output, $$x_i \in \mathbb{R}_+^p$$ is a vector of $$p$$ inputs and $$\beta$$ is a vector of the parameters corresponding to $$x_i$$. The error term $$\varepsilon_i$$ is composed of a normally distributed disturbance, $$v_i$$, representing the measurement and specification error, and a positive disturbance $$u_i$$ (following the half-normal distribution), representing technical inefficiency. Furthermore, $$v_i$$ and $$u_i$$ are assumed to be statistically independent from each other and from $$x_i$$. With the distribution assumptions on $$u_i$$ and $$v_i$$, the likelihood function for the model is constructed and the model is then estimated using the maximum likelihood estimator.

Once the parameters of the model have been estimated, one can obtain the expected level of technical inefficiency by estimating

$$E[u] = \sqrt{\frac{2}{\pi}} \sigma_u,$$

and the expected level of efficiency by using the following approximation:

$$E[\exp (-u)] \approx 1 - E[u].$$

If one is interested in the estimates of individual (in)efficiency of a specific production unit, more elaboration is needed. The most popular approach in the literature is to follow

---

*The formulation here is a convenient representation of a production relationship, where actual output is decomposed into the maximum output (with noise) and inefficiency, i.e., $$y_i = f (x_i | \beta) \exp (\varepsilon_i) = f (x_i | \beta) \exp (v_i) \exp (-u_i)$$. After log-transformation, we have a linear relationship as shown in (1).*

*Multiple outputs also can be considered. For example, this can be done by employing a distance function instead of the production function or by looking at the estimation of the cost frontier or by converting outputs into polar coordinates (e.g., see Simar and Zelenyuk (2011)). One can also use dimension reduction techniques to reduce the dimension outputs or inputs into smaller dimensions, e.g., via Principle Component Analysis, or using economic or price-based aggregation (e.g., see related discussion in Zelenyuk (2020) and an application in Nguyen and Zelenyuk (2021)). The latter approach can be especially useful in the case of very large dimensions (sometimes called ‘big wide data’ cases), e.g., as is done for measuring the total output of countries (e.g., GDP), industries or firms (total revenue) or for some inputs (e.g., capital). Due to space limitation we will focus here on the single output case, as was also considered in ALS and many other studies.*

*Other distributional assumptions such as exponential, truncated normal, gamma, etc., can be used for the inefficiency term (e.g., see Meeusen and van den Broeck (1977), Greene (1980a,b), Stevenson (1980), Greene (1990), Almanidis and Sickles (2012), Almanidis et al. (2014)).*

*The exact expression of the expected level of efficiency is given by $$E[\exp (-u)] = 2\Phi (-\sigma_u) \exp \left(\frac{\sigma^2_u}{2}\right).$$*
Jondrow et al. (1982) (hereafter JLMS), where the inefficiency of a production unit can be estimated or predicted using the expected value of $u_i$ conditional on the realisation of the composed error of the model, i.e., $E(u_i|\varepsilon_i)$ given by

$$E(u_i|\varepsilon_i) = \frac{\sigma_u \phi \left( \frac{\mu_i}{\sigma_u} \right)}{\Phi \left( \frac{\mu_i}{\sigma_u} \right)} + \mu_i,$$

where

$$\mu_i = \frac{-\sigma_u^2 \varepsilon_i}{\sigma_v^2 + \sigma_u^2},$$

and

$$\sigma_i^2 = \frac{\sigma_u^2 \sigma_v^2}{\sigma_v^2 + \sigma_u^2},$$

while $\phi(\cdot)$ and $\Phi(\cdot)$ are pdf and cdf of the standard normal distribution, respectively.

It is worth noting that while being originally developed for ALS, the JLMS-type procedure can be extended to predict (in)efficiency of a specific firm in the other models estimated by the maximum likelihood estimator (see more discussion in Kumbhakar, 1987).

### 2.2 Implementation of ALS Model

There are several options to estimate the basic stochastic frontier model in Stata. One can use the official Stata command `frontier` or utilise the command `sfcross` written by Belotti et al. (2013) or even set up the likelihood function using the `sfmodel` command then estimating the model using the official Stata routine for the maximum likelihood, `ml max`, as described in the handbook of Kumbhakar et al. (2015). These commands generate equivalent results for the basic stochastic frontier models and differ only in the formatting and listing of results and the options available for the different treatments of individual inefficiency (see more discussion in Jondrow et al., 1982).

One also can estimate the efficiency of a production unit by using the relationship $E[\exp(-u_i)|\varepsilon_i] \approx 1 - E[u_i|\varepsilon_i]$ or utilising the exact expression $E[\exp(-u_i)|\varepsilon_i] = \exp(-\mu_i + \frac{1}{2}\sigma_i^2) \frac{\phi \left( \frac{-\mu_i}{\sigma_i} \right)}{\Phi \left( \frac{-\mu_i}{\sigma_i} \right)}$ (Battese and Coelli, 1988).

The `sfcross` command (and the `sfpanel` command that we will discuss later for the panel data context) can be installed by executing the following command lines in Stata: `net install sfcross, all from (http://www.econometrics.it/stata)` and `net install sfpanel, all from (http://www.econometrics.it/stata)`.

The `sfmodel` and other user-written commands provided in the handbook of Kumbhakar et al. (2015) can be installed in Stata by executing the following command lines: `net install sfbook, from (https://sites.google.com/site/sfbook2014/home/install/)` replace and `sfbook install`, see more details in Kumbhakar et al. (2015) and its website, `https://sites.google.com/site/sfbook2014/`. 

---

9It is worth noting here that although this estimator is unbiased, it is an inconsistent estimator of individual inefficiency (see more discussion in Jondrow et al., 1982).

10One also can estimate the efficiency of a production unit by using the relationship $E[\exp(-u_i)|\varepsilon_i] \approx 1 - E[u_i|\varepsilon_i]$ or utilising the exact expression $E[\exp(-u_i)|\varepsilon_i] = \exp(-\mu_i + \frac{1}{2}\sigma_i^2) \frac{\phi \left( \frac{-\mu_i}{\sigma_i} \right)}{\Phi \left( \frac{-\mu_i}{\sigma_i} \right)}$ (Battese and Coelli, 1988).

11The `sfcross` command (and the `sfpanel` command that we will discuss later for the panel data context) can be installed by executing the following command lines in Stata: `net install sfcross, all from (http://www.econometrics.it/stata)` and `net install sfpanel, all from (http://www.econometrics.it/stata)`.

12The `sfmodel` and other user-written commands provided in the handbook of Kumbhakar et al. (2015) can be installed in Stata by executing the following command lines: `net install sfbook, from (https://sites.google.com/site/sfbook2014/home/install/)` replace and `sfbook install`, see more details in Kumbhakar et al. (2015) and its website, `https://sites.google.com/site/sfbook2014/`.
error distributions for the one-sided efficiency term and the inclusion of environmental factors.

As we progress in our chapter we consider a richer set of generalisations of the canonical stochastic frontier paradigm. Also, user-written commands provide us with more flexibility to estimate models that are not available with the current official Stata commands. Moreover, the user-written commands by Belotti et al. (2013) and Kumbhakar et al. (2015) also equip us with options to provide and refine the initial values for the maximum likelihood estimation, which can be very useful when dealing with complex likelihood functions.

After estimating the models, the estimates of technical inefficiency and efficiency can be obtained by using the postestimation routine `predict` (for the models estimated in the Stata version 16 by the official Stata command and the command written by Belotti et al. (2013)) or `sf_predict` (for the models estimated by the command written by Kumbhakar et al. (2015)). As an illustration, a snippet of Stata codes for implementing the ALS model is provided in Table 1.
R software also has several packages to implement the estimation of the basic stochastic frontier model. For example, one can use the package `frontier` written by Coelli and Henningsen (2020) or utilise the function `sfa` in the package `Benchmarking` written by Bogetoft and Otto (2019).

In order to estimate the basic stochastic frontier model, Matlab users need to set up the likelihood function and then utilise the optimisation routines, such as `fminunc` to optimise the likelihood function. Sickles and Zelenyuk (2019) provide a suite of Matlab codes to estimate a variety of stochastic frontier models on the website that accompanies their book. Although they do not include the ALS model, one can easily adapt their codes to obtain the estimates for this basic stochastic frontier model.

## 3 Early generation of stochastic panel data models

The basic stochastic frontier model discussed in the previous section is formulated in the cross-sectional setting and suffers from a number of drawbacks. As discussed in Schmidt and Sickles (1984), the three main disadvantages of the basic cross-sectional stochastic frontier model are: (i) there does not exist a consistent estimator of individual efficiency, (ii) the parametric distributional assumptions are usually required for the two error components (inefficiency and noise) to estimate the model and to predict the overall and individual (in)efficiency, and (iii) the assumption that inefficiency is independent of regressors is usually not plausible.

Over the past four decades, substantial efforts have been made to address these drawbacks of the cross-sectional stochastic frontier model. Among those, particular interest hinges on exploiting the advantages of panel data structure. Schmidt and Sickles (1984) were among the first who provided a general framework to extend the cross-sectional stochastic frontier model to the panel data setting, which also encompasses the Pitt and Lee (1981) full parametric random effect model.

---

13The package `frontier` uses the Fortran source codes of Frontier 4.1 originally developed by Tim Coelli (see more details in the manual of the package available at https://cran.r-project.org/web/packages/frontier/frontier.pdf).

14The website can be found at https://sites.google.com/site/productivityefficiency/home.

15The Matlab codes accompanying Sickles and Zelenyuk (2019) are also converted to R codes by Sickles et al. (2020), which can be accessed via the link provided on the book website or directly via https://sites.google.com/site/productivityinr.
3.1 Schmidt and Sickles (1984) Model

The model in Schmidt and Sickles (1984) can be formulated as follows:

\[ y_{it} = \beta_0 + x_{it}' \beta + v_{it} - u_i, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T, \]  

(7)

where \( y_{it} \in \mathbb{R}_+^1 \) is the output, \( x_{it} \in \mathbb{R}_+^p \) is a vector of \( p \) inputs of production unit \( i \) in time \( t \). \( v_{it} \) is the regular disturbance, while the unobserved individual heterogeneity, \( u_i \), represents technical inefficiency. Model (7) can be rewritten as

\[ y_{it} = \beta_0^* + x_{it}' \beta + v_{it} - u_i^* = c_i + x_{it}' \beta + v_{it}, \]  

(8)

where \( \beta_0^* = \beta_0 - E(u_i), \ u_i^* = u_i - E(u_i), \ E(u_i) \geq 0, \ c_i = \beta_0^* - u_i^* = \beta_0 - u_i. \)

Model (8) turns out to be a usual panel data model and can be estimated using the standard estimation methods in the panel data literature, such as the within estimator (i.e., in the fixed effect framework), the generalised least-square estimator (i.e., in the random effect framework), and the Hausman-Taylor estimator. After estimating the model, one can obtain the estimate \( \hat{c}_i \) of \( c_i \) and follow Schmidt and Sickles (1984) to construct a consistent estimator of technical inefficiency

\[ \hat{u}_i = \max (\hat{c}_i) - \hat{c}_i \geq 0, \quad i = 1, \ldots, n. \]  

(9)

The estimated inefficiency in (9) is measured with respect to the best practice production unit in the sample, which is implicitly assumed to be 100% efficient.

3.2 Implementation of Schmidt and Sickles (1984) Model

One can estimate the Schmidt and Sickles (1984) model using standard routines in Stata. Specifically, the official Stata command \texttt{xtreg} can be utilised to estimate the standard panel data model in (8) and the postestimation command \texttt{predict} can be used to obtain the estimate \( \hat{c}_i \) of \( c_i \). It is then straightforward to code formula (9) into Stata to get the estimates of technical inefficiency. Alternatively, one can use the command \texttt{sfpanel} written by Belotti et al. (2013) with the option \texttt{model(fe)} or \texttt{model(regls)} to estimate Schmidt and Sickles (1984) model in a fixed or random effects framework, respectively. As an illustration, a snippet of Stata codes for implementing the Schmidt and Sickles (1984) model (in the fixed effects framework) is provided in Table 2.

It is worth noting that model (8) and the individual inefficiency in (9) are estimated without any parametric assumptions on the distributions of composed errors. Alternatively, one can impose parametric assumptions on the distributions of the error components in model (7), e.g., a half-normal distribution for \( u_i \) and a normal distribution for...
Table 2: Illustration for implementation of the Schmidt and Sickles (1984) model in the fixed effects framework

---

```
/* The illustration here is for the fixed effects framework */
/* Note that output, inputs are in log forms and stored in global Stata variables $y, $xlist, respectively */
/* Implementation using the standard Stata commands */
xtrreg $y $xlist, fe /* Need to declare data to be panel before using xtrreg command*/
predict ci, u /* Obtain the estimate of ci */
quietly summarize ci
gen ineff_SS_1 = r(max) - ci /* Predict inefficiency*/
gen eff_SS_1 = exp(-ineff_SS_1) /* Predict efficiency*/
/* Implementation using the commands from Belotti et al. (2013) */
sfpanel $y $xlist, model(fe)
predict ineff_SS_2, u /* Predict inefficiency*/
gen eff_SS_2 = exp(-ineff_SS_2) /* Predict efficiency*/
```
Table 3: Illustration for implementation of the Pitt and Lee (1981) model in the random effects framework

<table>
<thead>
<tr>
<th>Partial Stata Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>predict ineff_PL_1, u /* Predict inefficiency, i.e., E(u</td>
</tr>
<tr>
<td>predict eff_PL_1, te /* Predict inefficiency, i.e., E(exp(-u)</td>
</tr>
<tr>
<td>sfpanel $y$ $xlist$, model(pl81)</td>
</tr>
<tr>
<td>predict ineff_PL_2, u /* Predict inefficiency, i.e., E(u</td>
</tr>
<tr>
<td>predict eff_PL_2, bc /* Predict inefficiency, i.e., E(exp(-u)</td>
</tr>
</tbody>
</table>

$v_{it}$ as discussed in Pitt and Lee (1981) and Schmidt and Sickles (1984). The model then can be estimated using the maximum likelihood estimator and the individual technical efficiency can be obtained by employing the JLMS procedure (extended to the panel data setting by Kumbhakar 1987). This model is estimated in Stata using the user-written command sfpanel from Belotti et al. (2013) with the option model(pl81). Alternatively, if one assumes that $u_i$ follows a truncated normal distribution, i.e., $u_i \sim N^+(\mu, \sigma^2)$, then the official Stata command xtfrontier with the option ti can be utilised. A snippet of Stata codes for implementing the Pitt and Lee (1981) model is provided in Table 3.

Estimation of the Schmidt and Sickles (1984) model also can be implemented in Matlab and R using the codes provided by Sickles and Zelenyuk (2019) and Sickles et al. (2020) (see the download links in footnote 14 and footnote 15).

### 3.3 Cornwell et al. (1990) Model

The technical inefficiency estimated within the Schmidt and Sickles (1984) framework is time-invariant, which may be an unrealistic restriction in many applied settings, especially in a long panel. To allow for time varying inefficiency in the Schmidt and Sickles (1984) framework, one can follow the suggestion in Cornwell et al. (1990) to replace $c_i$ by, e.g., $c_{it}$, where $c_{it}$ is a quadratic function of time trend $t$ with the parameters (coefficients)
being firm-specific, specifically
\[ c_{it} = \theta_0i + \theta_{1i}t + \theta_{2i}t^2. \]  

(10)

The parameters in (10) can be estimated by regressing the residual from the model (7) for production unit \( i \) on a constant, time, and time-squared (see more discussion in Cornwell et al. 1990). The fitted value from this model provides us with a consistent estimate (for large \( N \)) of \( c_{it} \), denoted as \( \hat{c}_{it} \). The individual technical inefficiency of production unit \( i \) at time \( t \) then can be estimated using an analogous procedure to Schmidt and Sickles (1984), specifically

\[ \hat{u}_{it} = \hat{c}_t - \hat{c}_{it}, \]

(11)

where

\[ \hat{c}_t = \max_j (\hat{c}_{jt}), \quad t = 1, \ldots, T. \]

(12)

3.4 Implementation of Cornwell et al. (1990) Model

Estimation of the Cornwell et al. (1990) model can be implemented using standard Stata routines in a set of procedures similar to those we discussed for the Schmidt and Sickles (1984) model. Alternatively, one can utilise the user-written command \texttt{sfpanel} from Belotti et al. (2013) with the option \texttt{model(fecss)}. A snippet of codes for implementing the Cornwell et al. (1990) model using \texttt{sfpanel} command is provided in Table 4.

Being similar to the Schmidt and Sickles (1984) model, one can estimate the Cornwell et al. (1990) model in Matlab and R using the codes provided by Sickles and Zelenyuk (2019) and Sickles et al. (2020) (see the download links in footnote 14 and footnote 15).

3.5 Kumbhakar (1990) and Battese and Coelli (1992) Models

If one is willing to impose distributional assumptions on the inefficiency component (as well as on the random disturbance term), the maximum likelihood estimation can be utilised to estimate time-varying efficiency models. Kumbhakar (1990) and Battese and Coelli (1992) appear to be the most popular models of this type. In the Kumbhakar (1990) model, time-varying inefficiency is modelled as

\[ u_{it} = \left(1 + \exp\left(at + bt^2\right)\right)^{-1} \tau_i, \]

\[ \tau_i \sim \text{iid} \mathcal{N}^+\left(0, \sigma^2\right), \]

(13)

Cornwell et al. (1990) outlined estimators for a general model in which any set of regressors could be drivers of efficiency change, if efficiency was interpreted as firm-specific heterogeneity. These regressors could be time varying. Thus the Cornwell et al. (1990) model was the first study about which we are aware to address the issue of environmental variables influencing efficiency levels.
while in Battese and Coelli (1992), time-varying inefficiency is specified as
\[
u_{it} = \{\exp[-\eta(t - T)]\} \tau_i,
\]
\[
\tau_i \sim_{iid} \mathcal{N}^+ (\mu, \sigma^2_\tau),
\] (14)
where \(a, b, \) and \(\eta\) are parameters to be estimated, and in both models, the random disturbance follows a normal distribution, i.e, \(v_{it} \sim_{iid} \mathcal{N} (0, \sigma^2_v)\).

Being similar to the Cornwell et al. (1990) model, the Kumbhakar (1990) and Battese and Coelli (1992) models extend the Pitt and Lee (1981) model by allowing the mean of inefficiency to vary over time, but they are more parsimonious in the sense that temporal patterns only depend on one or two parameters. The Cornwell et al. (1990) model, however, has an advantage in that it allows temporal patterns to vary across production units. Moreover, as discussed above, estimation of the Cornwell et al. (1990) model does not require parametric assumptions for inefficiency term.

### 3.6 Implementation of Kumbhakar (1990) and Battese and Coelli (1992) Models

The Battese and Coelli (1992) model, also known as a “time decay” model, can be estimated using Stata commands in its version 16 platform as well as by using user-written commands. Specifically, the estimation can be implemented by using the `xtfrontier` command with the option `tvd` or the command `sfpanel` from Belotti et al. (2013) with the option `model(bc92)`. The official Stata command `xtfrontier` cannot carry out estimation of the Kumbhakar (1990) model, which is available using the option `model(kumb90)` with the command `sfpanel` from Belotti et al. (2013). A snippet of Stata codes for
implementing Kumbhakar (1990) and Battese and Coelli (1992) models is provided in Table 5.

The estimation of the Battese and Coelli (1992) model can be implemented in R software by using the package frontier written by Coelli and Henningsen (2020). Alternatively, R users and Matlab users can utilise the codes prepared by Sickles and Zelenyuk (2019) and Sickles et al. (2020) (see the download links in footnote 14 and footnote 15).

4 Recent Advances of Stochastic Panel Data Models

The stochastic panel data models discussed so far have a major drawback in that technical inefficiency is not distinguishable from the unobserved individual heterogeneity, and thus technical inefficiency confounds with all time-invariant unobserved individual effects. Various approaches have been proposed in the literature to mitigate this (and other) issues. Here, we will focus on a few, namely Greene (2005a,b), Chen et al. (2014), Colombi...
et al. (2014), Kumbhakar et al. (2014) and Belotti and Ilardi (2018).

4.1 Greene (2005a, b) Models

Greene (2005a, b) proposed a stochastic panel data model in which unobserved individual heterogeneity separates from (transitory) technical efficiency. The model is formulated as

\[ y_{it} = c_i + x_{it}' \beta + \varepsilon_{it}, \]

\[ \varepsilon_{it} = v_{it} - u_{it}, \]  

where \( v_{it} \) as before is a regular disturbance, while \( u_{it} \geq 0 \) is the source of inefficiency.

Estimation of the model in (15) is challenging, especially in the fixed effects framework. The two main challenges are: (i) the estimation of parameters may be inconsistent due to the incidental parameters problem, and (ii) there does not exist a closed-form expression of the likelihood function of the within or first-difference transformation of the composed error if one follows standard procedures. Greene (2005a) proposed to use the maximum likelihood dummy variable estimator to estimate the model in the fixed effects framework and provided simulation evidence showing that the incidental parameters problem is not serious for relatively large \( T \).

4.2 Implementation of Greene (2005a, b) Models

One can implement estimation of the Greene (2005a, b) models in Stata by using the user-written command \texttt{sfpanel} from Belotti et al. (2013) with the option \texttt{model(tfe)} in the fixed effects framework and with the option \texttt{model(tre)} in the random effects framework.

Recently, Chen et al. (2014) derived a closed-form expression for the likelihood function of the within and first difference transformation of the model by exploiting the properties of the closed-skew normal distribution class. The model in (15) then can be estimated consistently in the fixed effects framework using the marginal maximum likelihood estimator. Belotti and Ilardi (2018) further extend the work of Chen et al. (2014) by considering the simulated marginal maximum likelihood estimator.

Chen et al. (2014) and Belotti and Ilardi (2018) estimators can be implemented in Stata using the command \texttt{sftfe} written by Belotti and Ilardi (2018) with the options \texttt{estimator(within)} and \texttt{estimator(mmsle)}, respectively. A snippet of Stata codes for implementing the Greene (2005a, b) models is provided in Table 6.

---

\footnote{Greene (2005a) also utilised the simulated maximum likelihood estimator to estimate the model in the random effect frameworks.}

\footnote{The \texttt{sftfe} command can be installed by executing the following command line in Stata: \texttt{net install sftfe, all from \url{http://www.econometrics.it/stata}.}
Table 6: Illustration for implementation of the Greene (2005a, b) models

*** Illustration for implementation of the Green (2005a, b) models ***
Partial Stata Codes

/* The illustration here is for the fixed effect framework */
/* Note that output, inputs are in log forms and stored in global Stata variables $y, $xlist, respectively */
/* Implementation using the maximum likelihood dummy variable estimator 
(the commands from Belotti et al. (2013)) */
sfpanel $y $xlist, distribution(hnormal) model(tfe)
predict ineff_G_1, u /* Predict inefficiency, i.e., E(u|e) */
predict eff_G_1, bc /* Predict efficiency, i.e., E(exp(-u)|e) */
/* Implementation using the marginal maximum likelihood estimator (the commands from Belotti and Ilardi (2018)) */
sftfe $y $xlist, distribution(hnormal) estimator(within)
predict ineff_G_2, u /* Predict inefficiency, i.e., E(u|e) */
predict eff_G_2, jlls /* Predict efficiency */
/* Implementation using the simulated marginal maximum likelihood estimator (the commands from Belotti and Ilardi (2018)) */
sftfe $y $xlist, distribution(hnormal) estimator(mmsle)
predict ineff_G_3, u /* Predict inefficiency, i.e., E(u|e) */
predict eff_G_3, jlls /* Predict efficiency */
To the best of our knowledge, routines to estimate the Greene (2005a,b) model are not yet available in R and Matlab in a public domain.

4.3 Colombi et al. (2014) and Kumbhakar et al. (2014) Models

The model specified in (15), although distinguishing between unobserved individual heterogeneity and technical inefficiency, only considers transitory inefficiency. Kumbhakar et al. (2014) and Colombi et al. (2014) further extend the model by decomposing the inefficiency into transitory and persistent components whose formulation is

\[ y_{it} = \beta_0 + x_{it}'\beta + c_i - \eta_i + v_{it} - u_{it}, \]

\[ c_i \sim_{iid} \mathcal{N}(0, \sigma_c^2), \]

\[ \eta_i \sim_{iid} \mathcal{N}^+(0, \sigma_\eta^2), \]

\[ v_{it} \sim_{iid} \mathcal{N}(0, \sigma_v^2), \]

\[ u_{it} \sim_{iid} \mathcal{N}^+(0, \sigma_u^2), \]

where \( c_i \) represents the unobserved individual heterogeneity, \( \eta_i \) represents the persistent inefficiency, \( u_{it} \) represents transitory inefficiency, and \( v_{it} \) is the regular disturbance. The model in (16) can be estimated using a single-stage maximum likelihood method (Colombi et al., 2014) or a multi-step procedure (Kumbhakar et al., 2014). The multi-step procedure although being inefficient relative to the single-stage maximum likelihood estimation, it is simpler and easier to implement. For the multi-step procedure, the model in (16) can be rewritten as

\[ y_{it} = \beta_0^* + x_{it}'\beta + \alpha_i + \varepsilon_{it}, \]

where

\[ \beta_0^* = \beta_0 - E[\eta_i] - E[u_{it}], \]

\[ \alpha_i = c_i - \eta_i + E(\eta_i), \]

\[ \varepsilon_{it} = v_{it} - u_{it} + E[u_{it}]. \]

The model in (17) turns out to be a standard panel data model and can be estimated by the usual panel data estimation methods. After estimating (17), one can obtain the predicted values of \( \alpha_i \) and \( \varepsilon_{it} \), \( \hat{\alpha}_i \) and \( \hat{\varepsilon}_{it} \), and then the persistent and transitory inefficiency components are estimated by applying standard stochastic frontier techniques to (19) and (20) with \( \alpha_i \) and \( \varepsilon_{it} \) replaced by \( \hat{\alpha}_i \) and \( \hat{\varepsilon}_{it} \).
Table 7: Illustration for implementation of the Kumbhakar et al. (2014) model

```
* Illustration for implementation of the Kumbhakar et al. (2014) models *
* Partial Stata Codes ***********************************************
/* The illustration here is for the random effect framework */
/* Note that output, inputs are in log forms and stored in global Stata
  variables $y, $xlist, respectively */
/* Implementation using the standard Stata commands */
xtreg $y $xlist, re
predict alp, u /* Obtain estimates of alpha */
predict esl, e /* Obtain estimates of the composed error */
/* Estimate equation (19) using the basic stochastic frontier model to
  obtain persistent (in)efficiency */
frontier alp, distribution(hnormal)
predict ineff_pers, u /* Predict persistent inefficiency, E(u|e) */
predict eff_pers, te /* Predict persistent efficiency, E(exp(-u)|e) */
/* Estimate equation (19) using the basic stochastic frontier model to
  obtain transitory (in)efficiency */
frontier esl, distribution(hnormal)
predict ineff_trans, u /* Predict transitory inefficiency, E(u|e) */
predict eff_trans, te /* Predict transitory efficiency, E(exp(-u)|e) */
```

4.4 Implementation of Colombi et al. (2014) and Kumbhakar et al. (2014) Models

The multi-step procedure to estimate the model specified in Colombi et al. (2014) and Kumbhakar et al. (2014) can be implemented in Stata using the command for panel data estimation, `xtreg`, together with the routines for basic stochastic frontier model estimation, e.g., `frontier` or `sfcross`. A snippet of Stata codes for implementing the Kumbhakar et al. (2014) model is provided in Table 7.

Similarly, R users can utilise panel data estimation routines (e.g., `plm` function) combined with commands for estimation of the basic stochastic frontier model discussed above (e.g., `sfa` or `frontier`) to estimate the Colombi et al. (2014) and Kumbhakar et al. (2014) models.

The implementation of the procedure in Matlab requires more effort since it is not easy (as in Stata or R) to perform panel data regression with this platform. With Matlab, one
needs to write their own code or download and install the panel data toolbox, e.g. the one written by Álvarez et al. (2017), to estimate a panel data regression model.

5 Stochastic Frontier Models with Determinants of Inefficiency

An interesting generalisation of stochastic frontier paradigm is extending the models to examine the impact of exogenous determinants on technical inefficiency. It is usually done by parameterising the parameters of inefficiency distribution, i.e., the pre-truncated mean or/and variance, as a function of exogenous variables. The approaches are applicable in both cross-sectional and panel data settings, and since it can be easily extended to panel data settings, here we focus our discussion on the cross-sectional context.

5.1 Popular Models

Cornwell et al. (1990) was the first to develop a model in which determinants of efficiency could be included in the stochastic frontier formulation. However, due to the linear way in which the determinants of efficiency were included in the regression model their fixed effect estimator could not point identify both a covariate’s effect on efficiency and its effect on the level of production. Kumbhakar et al. (1991) addressed this identification problem by specifying the efficiency determinants as a nonlinear function, parameterising the pretruncated mean of inefficiency as a function of exogenous variable, specifically

\[ u_i \sim N^+ \left( \mu_i, \sigma_u^2 \right), \]

\[ \mu_i = z_i' \delta, \]  

(21)

where \( z_i \in \mathbb{R}^k \) is a vector of \( k \) exogenous variables (including constant term) and \( \delta \) is a vector of the parameters to be estimated. Alternatively, Caudill et al. (1995) proposed specifying the variance of the inefficiency distribution as

\[ u_i \sim N^+ \left( 0, \sigma_{ui}^2 \right), \]

\[ \sigma_{ui}^2 = \exp \left( z_i' \delta \right). \]  

(22)

One can also at the same time parameterise both the pre-truncated mean and variance of inefficiency as a function of exogenous variables, i.e, combining (21) and (22), as in Wang (2002). These parametric stochastic frontier models are typically estimated using

\[ ^{19} \text{This model specification was cast in the panel data context and popularised by Battese and Coelli (1995)}. \]
the maximum likelihood estimator in much the same way as the basic stochastic frontier model.

Wang and Schmidt (2002) suggested a different specification for modelling the determinants of inefficiency based on a scaling property specifically

\[ u_i \sim g(z_i|\delta) u_i^*, \] (23)

where \( g(\cdot) \) is a positive function of the exogenous variables (the scaling function) and \( u_i^* \) is a positive random variable. With this specification, the distribution of inefficiency is the same for all production units, i.e., governed by \( u_i^* \), while the scale of the inefficiency distribution changes across production units depending on \( z_i \). The scaling property was further explored in Alvarez et al. (2006). Among others, they provided a nice economic interpretation for the scaling property in that \( u_i^* \) represents the baseline (in)efficiency of a production unit capturing things like the natural skills of its managers, meanwhile the scaling function allows (or prevents) the production unit to exploit these natural skills through other variables, \( z_i \), such as the experience and education of the managers, or the environment in which the production unit operates. Moreover, Alvarez et al. (2006) also devoted their attention to testing the hypothesis of the scaling property.

5.2 Implementation of Stochastic Frontier Models with Determinants of Inefficiency

Most of the parametric models discussed in this section can be easily implemented using Stata since the estimation routines for the basic stochastic frontier model in Stata also provide options to specify the pretruncated mean and/or variance of inefficiency as a function of the exogenous variables.

In particular, the Cornwell et al. (1990) estimator of course can be implemented using standard panel techniques and linear projections. The model specified in (23) can be estimated using nonlinear least squares without imposing any parametric assumption on the distribution of \( u_i^* \) or by the maximum likelihood based on the parametric distribution of the composed error. The maximum likelihood approach can be implemented in Stata by setting up the likelihood using the \texttt{sfmodel} command from Kumbhakar et al. (2015) with the option \texttt{hscale(\cdot)} and the log likelihood can be maximized using the standard Stata routine \texttt{ml max}.

A snippet of Stata codes for implementing stochastic frontier models with determinants of inefficiency is provided in Table 8.

\[ \text{It is worth mentioning here that although being popularised by Wang and Schmidt (2002), Simar et al. (1994) appear to be the first who analysed the scaling property in detail.} \]
Table 8: Illustration for implementation of stochastic frontier models with determinants of inefficiency

```stata
*** Illustration for implementation of stochastic frontier models with determinants of inefficiency
Partial Stata Codes

/* Note that output, inputs, and exogenous variables are stored in global Stata variables $y, $xlist, $zlist, respectively. Output and inputs are in log forms */
/* Implementation for the Kumbhakar et al. (1991) model (using the commands from Belotti et al. (2013)) */
sfcross $y $xlist, distribution(tnormal) emean($zlist)
predict ineff_KGM, u /* Predict inefficiency, i.e., E(u|e) */
predict eff_KGM, bc /* Predict inefficiency, i.e., E(exp(-u)|e) */
/* Implementation for the Caudill et al. (1995) model (using the commands from Belotti et al. (2013)) */
sfcross $y $xlist, distribution(hnormal) usigma($zlist)
predict ineff_CFG, u /* Predict inefficiency, i.e., E(u|e) */
predict eff_CFG, bc /* Predict inefficiency, i.e., E(exp(-u)|e) */
/* Implementation for the Wang (2002) model (using the commands from Belotti et al. (2013)) */
sfcross $y $xlist, distribution(tnormal) emean($zlist) usigma($zlist)
predict ineff_W, u /* Predict inefficiency, i.e., E(u|e) */
predict eff_W, bc /* Predict inefficiency, i.e., E(exp(-u)|e) */
/* Implementation for the Wang and Schmidt (2002) model (using the commands from Kumbhakar et al. (2015)) */
sfmodel $y, prod dist(t) frontier($xlist) scaling hscale($zlist) tau cu
cml max
sf_predict, jlms(ineff_WH) /* Predict inefficiency, i.e., E(u|e) */
sf_predict, bc(eff_WH) /* Predict inefficiency, i.e., E(exp(-u)|e) */
```

20
6 Semi-parametric Stochastic Frontier Models

Another generalisation of the stochastic frontier paradigm is to relax parametric assumptions imposed on the functional form of the production frontier and to some extent the parametric assumption on the distribution of inefficiency.

6.1 The variety of models

Banker and Maindiratta (1992) appear to be among the first attempting to estimate stochastic frontier models semiparametrically. They proposed a framework combining stochastic and deterministic frontier (i.e., data envelopment analysis) approaches and developed techniques for the maximum likelihood estimation with nonparametric characterisation of classes of monotone and concave production frontiers. Other early attempts belong to Fan et al. (1996) and Kneip and Simar (1996), who suggested using nonparametric kernel regression methods in the framework of parametric maximum likelihood estimation. Specifically, Fan et al. (1996) proposed a multi-stage semiparametric likelihood estimation approach, in which the Nadaraya-Watson nonparametric estimator is employed in the first stage to estimate the average production relationship and a full parametric maximum likelihood estimator is used in the next stage to back out the conditional mean of inefficiency, which is utilised in the last stage to identify the frontier. Kneip and Simar (1996) followed a similar strategy with Fan et al. (1996) but in a panel data setting.

Semi-parametric panel frontiers were also considered in a series of papers by Park et al. (1998, 2003, 2007) wherein firm inefficiency effects are endogenous. They constructed the semiparametric efficiency bounds and the corresponding semiparametric efficient estimators for such models under differing assumptions about the form of endogeneity, the serial dependence of the idiosyncratic error, and possible dynamic structures for the panel data model. They used kernel smoothers in these modeling efforts as did Adams et al. (1997, 1999), and Adams and Sickles (2007). Current Stata software for these models is in the development stage in Badunenko et al. (2021), while existing Matlab and R codes for these semi-nonparametric panel frontier models can be found on the website that accompanies Sickles and Zelenyuk (2019) (see the links in footnote 14 and footnote 15). Model averaging methods utilised in Sickles (2005), Duygun et al. (2017) and Isaksson et al. (2020) also can be found on that website and are currently being used in developing consensus productivity growth estimates for the United Nations Industrial Development Organization (UNIDO).

Another approach to estimate stochastic frontier models semiparametrically was pro-
posed by Kumbhakar et al. (2007), who suggested employing the local likelihood estimation. The key distinction between this approach and the parametric likelihood approach is that the estimation is localised in the sense that individual contribution to the likelihood is determined by the kernel-based weights instead of the equal weights. Kneip et al. (2015) extended the work of Kumbhakar et al. (2007) by relaxing the parametric assumption on the distribution of inefficiency, while Park et al. (2015) suggested an alternative parameterisation of the local likelihood and outlined a framework for allowing categorical variables in the local likelihood context.

Semi-parametric methods have also been introduced into the stochastic frontier paradigm to deal with specifications of inefficiency. Cornwell et al. (1990) utilised a second-order Taylor series in a time trend to model time varying inefficiency while Lee and Schmidt (1993) specified the time-varying and cross-sectionally varying inefficiency using a one-factor multiplicative model. Extensions to mixed models and more general factor models were pursued by Ahn et al. (2007, 2013), Kneip et al. (2004), Kneip and Sickles (2011), and Kneip et al. (2012). The latter model is programmed in Matlab and R on the software website for Sickles and Zelenyuk (2019) and Sickles et al. (2020) (see the links in footnote 14 and footnote 15 and its coding in Stata is in process in Badunenko et al. (2021). Finally, the Kneip and Sickles (2011) general cross-sectional and time-varying factor model is available in the R package from Oualid Bada and discussed at length in Bada and Liebl (2014).21

6.2 Simar et al. (2017) Model

Recently, Simar et al. (2017) suggested using the local least squares method as an alternative for the local likelihood approach to estimate the stochastic frontier models. The local least squares approach is much simpler to compute and easier to implement compared to the local likelihood, and we will focus our discussion here on this approach.

The model in Simar et al. (2017) can be formulated as follows

\[ y_i = m(x_i, z_i) + v_i - u_i, \quad i = 1, \ldots, n, \]  

(24)

where \( m(x_i, z_i) \) is the production frontier, \( y_i \in \mathbb{R}^1_+ \) is the output, \( x_i \in \mathbb{R}^p_+ \) is a vector of inputs, and \( z_i \in \mathbb{R}^k \) is a vector of \( k \) variables that can influence the production process. \( v_i \) is statistical noise, which is assumed to have zero mean, i.e., \( E(v_i|x_i, z_i) = 0 \), and positive finite variance, i.e., \( \text{VAR}(v_i|x_i, z_i) \in (0, \infty) \). Meanwhile, \( u_i \) is the inefficiency term following a one-sided distribution, with a positive mean, i.e., \( E(u_i|x_i, z_i) = \mu_u(x_i, z_i) \in (0, \infty) \).21

21Software instructions and downloadable codes are accessible at https://www.jstatsoft.org/article/view/v059i06
and positive finite variance, i.e., $VAR(u_i|x_i, z_i) \in (0, \infty)$. As in other stochastic frontier models, $u_i$ and $v_i$ are also assumed to be independent, conditionally on $(x_i, z_i)$.

Now, let us define

$$
\varepsilon_i^* = v_i - u_i + \mu_u(x_i, z_i),
$$

(25)

and

$$
r_1(x_i, z_i) = m(x_i, z_i) - \mu_u(x_i, z_i).
$$

(26)

We can rewrite (24) as

$$
y_i = r_1(x_i, z_i) + \varepsilon_i^*.
$$

(27)

Since $E(\varepsilon_i^*|x_i, z_i) = 0$, we can use standard nonparametric methods (e.g., local polynomial least squares) to estimate $r_1(x_i, z_i)$. In order to estimate the individual inefficiency, we also need to make a parametric assumption on the distribution of inefficiency, e.g.,

$$
u_i|x_i, z_i \sim \mathcal{N}^+(0, \sigma_{u}^2(x_i, z_i)).
$$

(28)

With the distributional assumption, the conditional mean of inefficiency can be estimated using the following relationships

$$
\sigma_u^3(x_i, z_i) = \sqrt{\frac{\pi}{2}} \left(\frac{\pi}{\pi - 4}\right) r_3(x_i, z_i),
$$

(29)

and

$$
\mu_u(x_i, z_i) = \sqrt{\frac{2}{\pi}} \sigma_u(x_i, z_i),
$$

(30)

where $r_3(x_i, z_i) = E((\varepsilon_i^*)^3|x_i, z_i)$ is the third moment of the composed error. Specifically, the residuals from the nonparametric estimation of the model in (27), $\hat{\varepsilon}_i^*$, can be utilised to obtain the nonparametrical estimates of the third moment of the composed error, $\hat{r}_3(x_i, z_i)$. The estimates of technical inefficiency then can be obtained by plugging the $\hat{r}_3(x_i, z_i)$ into (29) and (30).

6.3 Implementation of Simar et al. (2017) Model

Estimation of the Simar et al. (2017) model can be implemented using the standard Stata routines with a bit of additional programming. The key command is `npregress` which helps to perform the local least square estimation in Stata environment. As an

22The distributional assumptions on $u_i$ and $v_i$ allow obtaining a generalised version of JLMS type estimates, although more interesting in the semi/non-parametric context are the estimates of $E(u_i|x_i = x, z_i = z)$, which can be done for any values of interest for $(x, z)$. The elasticities of $E(u_i|x_i = x, z_i = z)$ can also be obtained, which can be done without any parametric assumptions on distributions, just by assuming that $u_i$ comes from a one-parameter scale family (see Section 4 in Simar et al., 2017 for more details).
illustration, we provide here in Table 9 a part of a Stata do file that implements the procedure discussed in the previous subsection to estimate the Simar et al. (2017) model.

Similarly, one can implement the estimation of the Simar et al. (2017) model in R with the local least squares estimation being carried out by the np package with a bit of additional programming similar to the one we presented here (and as was done by Parmeter and Zelenyuk (2019)). The implementation of the model in Matlab requires more effort since one needs to write his/her own codes for the local least square estimation (as was done by Simar et al. (2017)). Preparation of user-friendly packages in R and Matlab is currently in progress.
Table 9: Illustration for implementation of the Simar et al. (2017) model

```
/* Estimate the model in (27) using local linear estimator and store 
   predicted value in variable r1hat. Note that output, inputs, and 
exogenous variables are stored in global Stata variables $y, $xlist, $zlist, respectively. Output and inputs are in log forms */
npregress kernel $y $xlist $zlist, estimator(linear) predict(r1hat) noderivatives

/* Obtain the residual and the residual cubed from estimation of the 
   model in equation (27)*/
gen ehat = $y - r1hat
gen ehat3 = ehat^3

/* Estimate the third moment of the composed error using local linear 
estimator and store predicted value in variable r3hat */
npregress kernel ehat3 $xlist $zlist, estimator(linear) predict(r3hat) noderivatives

/* Calculate sigma u hat cubed using equation (29) */
gen sigmauhat3 = sqrt(_pi/2)*(_pi/(_pi-4))*r3hat

/* Calculate sigma u hat. Note that following Simar et al. (2017), we 
set negative values of sigma u hat equal zero */
gen sigmauhat = max(sigmauhat3^(1/3),0)

/* Calculate estimated values of inefficiency using equation (30)*/
gen muhat = sqrt(2/_pi)*sigmauhat
```


7 Empirical Illustration

In this section, we provide a small empirical illustration of the models discussed in the previous sections, including the basic stochastic frontier model, the stochastic panel data models, and the semi-parametric stochastic frontier model. For this purpose, we use the data set about rice producers in the Philippines, which was also utilised for the similar purpose and popularised in the literature by Coelli et al. (2005).

Specifically, the data set includes the information about 43 rice producers in the Tarlac region of the Philippines in a period of 8 years from 1990 to 1997. We extract from the data set the information on one output and three inputs including the area planted, labour used and fertiliser used. The output is measured in tonnes of freshly threshed rice, while the inputs are measured in hectares, man-days of family and hired labour, and kilograms of active ingredients, respectively (see more details about the description of the data in Coelli et al., 2005).

For this empirical illustration, we deliberately apply all the models to the data and focus our discussion on the estimated inefficiency to reflect the differences in results across the models. Moreover, for all the models that require a functional form for the production relationship, we assume a linear in log production function, i.e., the Cobb-Douglas production function. The Stata codes for implementing this analysis are provided in the Appendix.

The summary statistics of the estimated inefficiency are provided in Table 10 and their histograms are shown in Figure 1. Meanwhile, the variations of the estimated inefficiency across the years are shown in Figure 2.

At first glance, we can see that the means of estimated inefficiency vary significantly across the models, ranging from 0.20 (the Battese and Coelli (1992) model) to 0.45 (the Kumbhakar et al. (2014) model). This is understandable since each model depends on

\[ \hat{\sigma}^2_u(x_i, z_i) < 0 \]

the estimated distribution of estimated inefficiency from the Simar et al. (2017) model is showing some mass at zero (i.e., the phenomenon referred to as 'wrong skewness' in stochastic frontier analysis) because 79 out 344 observations have \( \hat{\sigma}^2_u(x_i, z_i) < 0 \) and their inefficiency is set to equal to 0.

It is important to clarify here that for all the models, the means we refer to are averages of the estimates of individual inefficiencies.
Table 10: Summary statistics of the estimated inefficiency

<table>
<thead>
<tr>
<th>Models</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aigner et al. (1977)</td>
<td>0.36</td>
<td>0.25</td>
<td>0.04</td>
<td>0.18</td>
<td>0.29</td>
<td>0.47</td>
<td>2.00</td>
</tr>
<tr>
<td>Schmidt and Sickles (1984) (fixed effects)</td>
<td>0.34</td>
<td>0.20</td>
<td>0.00</td>
<td>0.18</td>
<td>0.30</td>
<td>0.50</td>
<td>0.98</td>
</tr>
<tr>
<td>Schmidt and Sickles (1984) (random effects)</td>
<td>0.23</td>
<td>0.12</td>
<td>0.00</td>
<td>0.13</td>
<td>0.21</td>
<td>0.33</td>
<td>0.60</td>
</tr>
<tr>
<td>Pitt and Lee (1981)</td>
<td>0.21</td>
<td>0.14</td>
<td>0.05</td>
<td>0.10</td>
<td>0.16</td>
<td>0.31</td>
<td>0.70</td>
</tr>
<tr>
<td>Cornwell et al. (1990)</td>
<td>0.44</td>
<td>0.27</td>
<td>0.00</td>
<td>0.27</td>
<td>0.41</td>
<td>0.57</td>
<td>2.09</td>
</tr>
<tr>
<td>Battese and Coelli (1992)</td>
<td>0.20</td>
<td>0.15</td>
<td>0.04</td>
<td>0.08</td>
<td>0.14</td>
<td>0.28</td>
<td>0.92</td>
</tr>
<tr>
<td>Greene (2005a, b) (random effects)</td>
<td>0.33</td>
<td>0.24</td>
<td>0.03</td>
<td>0.17</td>
<td>0.27</td>
<td>0.43</td>
<td>1.89</td>
</tr>
<tr>
<td>Greene (2005a, b) (fixed effects)</td>
<td>0.35</td>
<td>0.23</td>
<td>0.02</td>
<td>0.18</td>
<td>0.30</td>
<td>0.47</td>
<td>1.87</td>
</tr>
<tr>
<td>Kumbhakar et al. (2014) (Total)</td>
<td>0.45</td>
<td>0.25</td>
<td>0.10</td>
<td>0.27</td>
<td>0.38</td>
<td>0.55</td>
<td>2.01</td>
</tr>
<tr>
<td>Kumbhakar et al. (2014) (Persistent)</td>
<td>0.15</td>
<td>0.10</td>
<td>0.02</td>
<td>0.06</td>
<td>0.13</td>
<td>0.24</td>
<td>0.49</td>
</tr>
<tr>
<td>Kumbhakar et al. (2014) (Transitory)</td>
<td>0.29</td>
<td>0.20</td>
<td>0.03</td>
<td>0.16</td>
<td>0.24</td>
<td>0.37</td>
<td>1.67</td>
</tr>
<tr>
<td>Simar et al. (2017)</td>
<td>0.25</td>
<td>0.14</td>
<td>0.00</td>
<td>0.17</td>
<td>0.30</td>
<td>0.35</td>
<td>0.45</td>
</tr>
</tbody>
</table>

different sets of assumptions. Moreover, it is important for practitioners to be aware of these differences and carefully justify the assumptions of the model of their choice before proceeding with their analysis. For example, with this data set, the difference in estimated inefficiency between the fixed effects and random effects frameworks is significant when all the unobserved individual heterogeneity is viewed as inefficiency (e.g., in the Schmidt and Sickles (1984) model), but the difference is minimal when inefficiency is distinguished from the unobserved individual heterogeneity (e.g., in the Greene (2005a, b) model).

Furthermore, recall that the temporal pattern of inefficiency is also specified differently in different models. As illustrated in Figure 2 the estimated inefficiency is constant over time in the Schmidt and Sickles (1984) and Pitt and Lee (1981) models, but follows a quadratic trend in the Cornwell et al. (1990) model and has a linear trend in Battese and Coelli (1992). Meanwhile, other models, such as Greene (2005a, b) and Kumbhakar et al. (2014), do not impose any temporal patterns on the time varying component of inefficiency.
Figure 1: Histograms of the estimated inefficiency
Figure 2: The estimated inefficiency across years
8 Concluding Remarks

This chapter discussed a variety of stochastic frontier models to estimate the technical efficiency of production units. Our chapter also documented the estimation routines used to implement these methods for practitioners, especially those who are willing to use Stata, but also with tips on where to find analogous programs for R and Matlab users.

Although many recent developments in the field were covered in this chapter, it was still a relatively brief introduction to the stochastic frontier paradigm with some other generalisations remaining untouched, such as Bayesian stochastic frontier\footnote{E.g., see \cite{van1994frontier,griffin2004frontier,liu2017frontier}}, stochastic metafrontier\footnote{E.g., see \cite{battese2004frontier,o2008frontier,huang2014frontier}}, spillovers and spatial frontiers\footnote{E.g., see \cite{glass2016frontier,orea2019frontier}}, and endogeneity\footnote{E.g., see \cite{amsler2016frontier,kutlu2010frontier,karakaplan2015frontier,karakaplan2017frontier}}. We refer interested readers to more extensive resources \cite{Sickles2019frontier,Kumbhakar2021frontier} for more detailed discussions of these and other topics.

Finally, many other important developments in the field are still in progress, thus we encourage readers to check for updates as well as contribute themselves to such developments and discoveries.

Acknowledgments

We thank the Editor and two anonymous referees for fruitful comments. Bao Hoang Nguyen and Valentin Zelenyuk acknowledge the support from The University of Queensland and the financial support from the Australian Research Council (FT170100401). We also thank Hong Ngoc Nguyen, Zhichao Wang and Evelyn Smart for their valuable feedback. These individuals and organizations are not responsible for the views expressed in this paper.

\footnotesize
\begin{itemize}
  \item \cite{van1994frontier,griffin2004frontier,liu2017frontier}
  \item \cite{battese2004frontier,o2008frontier,huang2014frontier}
  \item \cite{glass2016frontier,orea2019frontier}
  \item \cite{amsler2016frontier,kutlu2010frontier,karakaplan2015frontier,karakaplan2017frontier}
\end{itemize}
Listing 1: Stata codes for the empirical illustration

```
set more off
clear all
log using "SFABookChapter_Empirical", text replace
// Load data
import delimited "rice.txt", delimiter(space) varnames(1) ///
   encoding(ISO-8859-2)
// Generate variables
foreach X of varlist prod area labor npk {
    generate l'X' = ln(`X')
}
global y lprod
global xlist llabor larea lnpk /* Cobb-Douglas function*/
global id fmercode
global t yeardum
xtset $id $t
***********************************************************************
/* the Aigner et al. (1977) model */
***********************************************************************
sfcross $y $xlist, distribution(hnormal)
estimates store ALS
predict ineff_ALS, u /* Predict inefficiency, i.e., E(u|e) */
label variable ineff_ALS "The Aigner et al. (1977) model"
***********************************************************************
/* the Schmidt and Sickles (1984) model */
***********************************************************************
/* Partial Stata Codes**************************
/* the fixed effect framework */
sfpanel $y $xlist, model(fe)
estimates store SS_fe
predict ineff_SS_fe, u /* Predict inefficiency*/
label variable ineff_SS_fe "The Schmidt and Sickles (1984) (fixed effects)"
/* the random effect framework */
sfpanel $y $xlist, model(regls)
estimates store SS_re
predict ineff_SS_re, u /* Predict inefficiency*/
label variable ineff_SS_re "The Schmidt and Sickles (1984) (random effects)"
```
 /**************************************************************************
 */ the Pitt and Lee (1981) model */
***************************************************************************/
sfpanel \$y $xlist, model(pl81)
estimates store PL
predict ineff_PL, u /* Predict inefficiency, i.e., E(u|e) */
label variable ineff_PL "The Pitt and Lee (1981) model"
/***************************************************************************/
/* the Cornwell et al. (1990) model */
***************************************************************************/
sfpanel \$y $xlist, model(fecss)
estimates store CSS
predict ineff_CSS, u /* Predict inefficiency*/
label variable ineff_CSS "The Cornwell et al. (1990) model"
/***************************************************************************/
/* the Battese and Coelli (1992) model */
***************************************************************************/
sfpanel \$y $xlist, model(bc92)
estimates store BC
predict ineff_BC, u /* Predict inefficiency, i.e., E(u|e) */
label variable ineff_BC "The Battese and Coelli (1992) model"
/***************************************************************************/
/* the Green (2005a,b) models */
***************************************************************************/
/* the random effect framework */
sfpanel \$y $xlist, distribution(hnormal) model(tre)
estimates store G_tre
predict ineff_G_tre, u /* Predict inefficiency, i.e., E(u|e) */
label variable ineff_G_tre "The Green (2005a,b) model (random effects)"
/* the fixed effect framework */
/* Implementation using the marginal maximum likelihood estimator */
sftfe \$y $xlist, distribution(hnormal) estimator(within)
estimates store G_mmle
predict ineff_G_mmle, u /* Predict inefficiency, i.e., E(u|e) */
label variable ineff_G_mmle "The Green (2005a,b) model (fixed effects)"
/***************************************************************************/
/* the Kumbhakar et al. (2014) models */
***************************************************************************/
xtreg \$y $xlist, re
estimates store KLH
predict alp, u /* Obtain estimates of alpha */
predict esl, e /* Obtain estimates of the composed error */
/* Estimate equation (19) using the basic stochastic frontier model to obtain persistent (in)efficiency */
gen constant = 1 /* Generate constant to use in sfcross */
sfcross alp constant, distribution(hnormal) noconstant predict ineff_KHL_pers, u /* Predict inefficiency, i.e., E(u|e) */ label variable ineff_KHL_pers /// "The Kumbhakar et al. (2014) model (Persistent)"
/* Estimate equation (19) using the basic stochastic frontier model to obtain transitory (in)efficiency */
sfcross esl constant, distribution(hnormal) noconstant predict ineff_KHL_trans, u /* Predict inefficiency, i.e., E(u|e) */ label variable ineff_KHL_trans /// "The Kumbhakar et al. (2014) model (Transitory)"
/* Calculate the total inefficiency*/
gen ineff_KLH = ineff_KHL_pers + ineff_KHL_trans label variable ineff_KLH "The Kumbhakar et al. (2014) model (Total)"

* **********************************************************************
/* the Simar et al. (2017) */
* **********************************************************************
/* Estimate model in (27) using local linear estimator and store predicted value in variable r1hat. Note that output, inputs, and exogenous variables are stored in global Stata variables $y, $xlist, $zlist, respectively */
npregress kernel $y $xlist, estimator(linear) predict(r1hat) /// noderivatives
/* Obtain the residual and the residual cubed from the estimation of model in equation (27)*/
gen ehat = $y - r1hat
gen ehat3 = ehat^3
/* Estimate the third moment of the composed error using local linear estimator and store predicted value in variable r3hat */
npregress kernel ehat3 $xlist, estimator(linear) predict(r3hat) /// noderivatives
/* Calculate sigma u hat cubed using equation (29) */
gen sigmaint3 = sqrt(_pi/2)*(_pi/(_pi-4))*r3hat
/* Calculate sigma u hat. Note that following Simar et al. (2017), we set negative values of sigma u hat equal zero */
gen sigmaint = max(sigmaint3^(1/3),0)
/* Calculate estimated values of inefficiency using equation (30)*/
gen ineff_SKVZ = sqrt(2/_pi)*sigmaint label variable ineff_SKVZ "The Simar et al. (2017) model"
/* Summarising and exporting the results */
/* The estimated coefficients of the frontiers */
esttab ALS SS_fe SS_re PL CSS BC G_tre G_mmle KLH
esttab ALS SS_fe SS_re PL CSS BC G_tre G_mmle KLH ///

/* The estimated inefficiency */
global myvars ineff_ALS ineff_SS_fe ineff_SS_re ineff_PL ineff_CSS ///

estpost summarize $myvars, detail

esttab using inefficiency .csv , ///
cells("count mean sd min p25 p50 p75 max") replace

/* Histograms of estimated inefficiency */
foreach X of varlist $myvars {
    histogram `X', bin(100) normal `kden'
    graph save `X'.gph, replace
}

graph combine ineff_ALS.gph ineff_SS_fe.gph ineff_SS_re.gph ///
    ineff_PL.gph ineff_CSS.gph ineff_BC.gph ineff_G_tre.gph ///
    ineff_G_mmle.gph ineff_KLH.gph ineff_KHL_pers.gph ///
    ineff_KHL_trans.gph ineff_SKVZ.gph, col(3) scale(1)

graph export histogramineff.png, replace

/* Plot estimated inefficiency across years */
foreach X of varlist $myvars {

    by yeardum, sort: egen `X'_Q1 = pctile(`X'), p(25)
    by yeardum, sort: egen `X'_Q2 = pctile(`X'), p(50)
    by yeardum, sort: egen `X'_Q3 = pctile(`X'), p(75)
    label variable `X'_Q1 "First quartile"
    label variable `X'_Q2 "Second quartile"
    label variable `X'_Q3 "Third quartile"
    label variable `X' "Median"
    local labeltext : variable label `X'
    graph two line `X'_Q1 `X'_Q2 `X'_Q3 yeardum, ///
        title(`labeltext', size(small))
    graph save `X'_trend.gph, replace
}

graph combine ineff_ALS_trend.gph ineff_SS_fe_trend.gph ///
References


Badunenko, O., Sickles, R. C. and Zelenyuk, V. (2021), Productivity and efficiency analysis in Stata. monograph in progress, Rice University.


