The Stata Journal (2012) vv, Number ii, pp. 1–39

Stochastic frontier analysis using Stata

Federico Belotti
University of Rome Tor Vergata

Silvio Daidone
University of York

Giuseppe Ilardi
Bank of Italy

Vincenzo Atella
University of Rome Tor Vergata

Abstract. This paper describes `sfcross` and `sfpanel`, two new Stata commands for the estimation of cross-sectional and panel data stochastic frontier models. `sfcross` extends the official `frontier` capabilities by including additional models (Greene 2003; Wang 2002) and command functionality, such as the possibility to manage complex survey data characteristics. Similarly, `sfpanel` allows to estimate a much wider range of time-varying inefficiency models compared to the official `xtfrontier` command including, among the others, the Cornwell et al. (1990) and Lee and Schmidt (1993) models, the flexible model of Kumbhakar (1990), the inefficiency effects model of Battese and Coelli (1995) and the "true" fixed and random-effects models developed by Greene (2005a). A brief overview of the stochastic frontier literature, a description of the two commands and their options and illustrations using simulated and real data are provided.

Keywords: st000, stochastic frontier analysis, cross-sectional, panel data

1 Introduction

The aim of this article is to describe `sfcross` and `sfpanel`, two new Stata commands for the estimation of parametric Stochastic Frontier (SF) models using cross-sectional and panel data. Starting from the seminal papers by Meeusen and van den Broeck (1977) and Aigner et al. (1977), this class of models has become a popular tool for efficiency analysis. Since then, a continuous stream of research has produced many reformulations and extensions of the original statistical models, generating a flourishing industry of empirical studies. An extended review of these models can be found in the recent survey by Greene (2008).

The SF model is motivated by the theoretical idea that no economic agent can exceed the ideal "frontier" and the deviations from this extreme represent the individual inefficiencies. From the statistical point of view, this idea has been implemented by specifying a regression model characterized by a composite error term in which the classical idiosyncratic disturbance, aiming at capturing measurement error and any other classical noise, is included together with a one-sided disturbance which represents inefficiency.[1] Whether cross-sectional or panel data, production or cost frontier, time-invariant or varying inefficiency, parametric SF models are usually estimated by likelihood-based

1. The literature distinguishes between production and cost frontiers. The former represent the maximum amount of output that can be obtained from a given level of inputs, while the latter characterizes the minimum expenditure required to produce a bundle of outputs given the prices of the inputs used in its production.

 st000

methods, and the main interest is on making inference about both frontier parameters and inefficiency.

The estimation of SF models is already possible using official Stata routines. However, the available commands cover a restricted range of models, especially in the panel data case.

The `sfcross` command provided in this article mirrors the official `frontier` command functionality, adding new features such as: *i*) the estimation of Normal-Gamma models via Simulated Maximum Likelihood (SML) (Greene 2003); *ii*) the estimation of the Normal-Truncated Normal model proposed by (Wang 2002) in which both the location and the scale parameters of the inefficiency distribution can be expressed as a function of exogenous covariates; and *iii*) the opportunity to manage complex survey data characteristics (via the `svyset` command).

As far as panel data analysis is concerned, the official Stata `xtfrontier` command allows the estimation of a Normal-Truncated Normal model with time-invariant inefficiency (Battese and Coelli 1988) and a time-varying version, named as "time decay" model, proposed by Battese and Coelli (1992). Our `sfpanel` command allows to estimate a wider range of time-varying inefficiency models including the Cornwell et al. (1990) and Lee and Schmidt (1993) models, the flexible model of Kumbhakar (1990), the time decay and the inefficiency effects models of Battese and Coelli (Battese and Coelli 1992, 1995) and the "true" fixed (TFE) and random-effects (TRE) models developed by Greene (2005a). For the last two models, the command allows different distributional assumptions, providing the modeling of both inefficiency location and scale parameters. Furthermore, the command allows the estimation of the random-effects time-invariant inefficiency models of Pitt and Lee (1981) and Battese and Coelli (1988), as well as the fixed-effects version of the Schmidt and Sickles (1984) model, characterized by no distributional assumptions on the inefficiency term. In addition, since the main objective of the SF analysis is the estimation of inefficiency, we provide post estimation routines to compute both inefficiency and efficiency scores, as well as their confidence intervals (Jondrow et al. 1982; Battese and Coelli 1988; Horrace and Schmidt 1996). Finally, `sfcross` and `sfpanel` allow also the simultaneous modelling of heteroscedasticity in the idiosyncratic error term.

In the development of these new commands, we make extensive use of Mata to speed up the estimation process. We allow for the use of Stata factor variables, weighted estimation, constrained estimation, resampling-based variance estimation and clustering. Moreover, by using Mata structures and libraries, we provide a very readable code prone to be easily developed further by the Stata users community. All these features make the commands simple to use, extremely flexible and fast, ensuring at the same time the opportunity to estimate state-of-the-art SF models.

Finally, we would like to emphasize that `sfpanel` offers the possibility to perform a constrained fixed-effects estimation, which is not yet available with `xtreg`. Moreover, the Cornwell et al. (1990) and Lee and Schmidt (1993) models, although proposed in the SF literature, are linear panel data models with time-varying fixed-effects, thus potentially very useful also in other contexts.

The paper is organized as follows. In Section 2, we present a brief review of the SF approach evolution, focusing on the models that can be estimated using the proposed commands. Sections 3 and 4 describe the syntax of `sfcross` and `sfpanel`, focusing on the main options. Sections 5 and 6 illustrate the two commands using simulated data and two empirical applications from the SF literature. Finally, section 7 offers some conclusions.

2 A review of stochastic frontier models

We begin our discussion with a general formulation of the SF cross-sectional model and then review extensions and improvements that have been proposed in the literature, focusing on those models that can be estimated using `sfcross` and `sfpanel`. Given the large number of estimators allowed by the two commands, we deliberately do not discuss the derivation of the corresponding criterion functions. We refer the reader to the cited works for details on the estimation of each model. A synopsis guide with all estimable models and their features is reported in table 1.

2.1 Cross-sectional models

Consider the following SF model

M1

$$
\begin{aligned}
y_i &= \alpha + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, N, && (1)\\
\varepsilon_i &= v_i - u_i, && (2)\\
v_i &\sim \mathcal{N}(0, \sigma_v^2), && (3)\\
u_i &\sim \mathcal{F}, && (4)
\end{aligned}
$$

where y_i represents the logarithm of the output (or cost) of the *i-th* productive unit, $\boldsymbol{x}_i$ is a vector of inputs (input prices and quantities in the case of a cost frontier) and $\boldsymbol{\beta}$ is the vector of technology parameters. The composed error term ε_i is the sum (or the difference) of a normally distributed disturbance, v_i, representing measurement and specification error, and a one-side disturbance, u_i, representing inefficiency.[2] Moreover, u_i and v_i are assumed to be independent of each other and i.i.d. across observations. The last assumption about the distribution $\mathcal{F}$ of the inefficiency term is needed to make the model estimable. Aigner et al. (1977) assumed a Half-Normal distribution, i.e. $u_i \sim \mathcal{N}^+\left(0, \sigma_u^2\right)$, while Meeusen and van den Broeck (1977) opted for an Exponential one, $u_i \sim \mathcal{E}\left(\sigma_u\right)$. Other commonly adopted distributions are the Truncated Normal (Stevenson 1980) and the Gamma distributions (Greene 1980a,b, 2003).

The distributional assumption required for the identification of the inefficiency term implies that this model is usually estimated by Maximum Likelihood (ML), even if modified ordinary least squares or generalized method of moments estimators are pos-

2. In this section, we consider only production functions. However, the sign of the u_i term in equation (2) is positive or negative depending on whether the frontier describes a cost or a production function, respectively.

sible (often inefficient) alternatives.[3] In general, SF analysis is based on two sequential steps: in the first, estimates of the model parameters $\hat{\boldsymbol{\theta}}$ are obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}', \sigma_u^2, \sigma_v^2)'$.[4] In the second step, point estimates of inefficiency can be obtained through the mean (or the mode) of the conditional distribution $f(u_i|\hat{\varepsilon}_i)$, where $\hat{\varepsilon}_i = y_i - \hat{\alpha} - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$.

The derivation of the likelihood function is based on the independence assumption between u_i and v_i. Since the composite model error ε_i is defined as $\varepsilon_i = v_i - u_i$, its p.d.f. is the convolution of the two component densities as

$$f_\varepsilon(\varepsilon_i) = \int_0^{+\infty} f_u(u_i) f_v(\varepsilon_i + u_i) du_i. \tag{5}$$

Hence, the log-likelihood function for a sample of n productive units is

$$\ell(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{n} \log f_\varepsilon(\varepsilon_i|\boldsymbol{\theta}). \tag{6}$$

The marginalization of u_i in equation (5) leads to a convenient closed-form expressions only for the Normal-Half Normal, Normal-Exponential and Normal-Truncated Normal models. In all other cases (e.g., the Normal-Gamma model) numerical or simulation based techniques are necessary to approximate the integral in equation (5).

The second estimation step is necessary since the estimates of the model parameters allow the computation of residuals $\hat{\boldsymbol{\varepsilon}}$, but not the inefficiency estimates. Since the main objective of SF analysis is the estimation of technical (or cost) efficiency, a strategy for disentangling this unobserved component from the compounded error is required. As mentioned before, the most well-known solutions to this problem, proposed by Jondrow et al. (1982) and Battese and Coelli (1988), exploit the conditional distribution of $\boldsymbol{u}$ given $\boldsymbol{\varepsilon}$. Thus, a point estimate of the inefficiencies can be obtained using the mean $\mathbb{E}(\boldsymbol{u}|\hat{\boldsymbol{\varepsilon}})$ (or the mode $\mathbb{M}(\boldsymbol{u}|\hat{\boldsymbol{\varepsilon}})$) of this conditional distribution. Once point estimates of $\boldsymbol{u}$ are obtained, estimates of the technical (cost) efficiency can be derived as

$$\text{Eff} = \exp(-\hat{\boldsymbol{u}}).$$

where $\hat{\boldsymbol{u}}$ is either $\mathbb{E}(\boldsymbol{u}|\hat{\boldsymbol{\varepsilon}})$ or $\mathbb{M}(\boldsymbol{u}|\hat{\boldsymbol{\varepsilon}})$.[5]

2.2 Panel data models

The availability of a richer set of information in panel data allows to relax some of the assumptions previously imposed and to consider a more realistic characterization of the inefficiencies.

3. Notice that, when a distributional assumption on $\boldsymbol{u}$ is made, `sfcross` and `sfpanel` estimate model parameters by likelihood-based techniques.
4. Different model parametrizations are used in the SF literature as, for example, $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}', \sigma^2, \lambda)'$ where $\sigma^2 = \sigma_u^2 + \sigma_v^2$ and $\lambda = \sigma_u/\sigma_v$.
5. A general presentation of the post estimation procedures implemented in the `sfcross` and `sfpanel` routines is given by Kumbhakar and Lovell (2000) and Greene (2008), to which we refer the reader for further details.

Pitt and Lee (1981) were the first to extend model (1-4) to longitudinal data. They proposed the ML estimation of the following Normal-Half Normal SF model

$$\begin{aligned} y_{it} &= \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \quad i = 1, \ldots, N, \quad t = 2, \ldots, T_i, & (7)\\ \varepsilon_{it} &= v_{it} - u_i, & (8)\\ v_{it} &\sim \mathcal{N}(0, \sigma_v^2), & (9)\\ u_i &\sim \mathcal{N}^+\left(0, \sigma_u^2\right). & (10) \end{aligned}$$

The generalization of this model to the Normal-Truncated Normal case has been proposed by Battese and Coelli (1988).[6] As pointed out by Schmidt and Sickles (1984), the estimation of a SF model with time invariant inefficiency can also be performed by adapting conventional fixed-effects estimation techniques, thereby allowing inefficiency to be correlated with the frontier regressors and avoiding distributional assumptions about u_i. However, the time invariant nature of the inefficiency term has been questioned, especially in presence of empirical applications based on long panel data sets. To relax this restriction, Cornwell et al. (1990) have approached the problem proposing the following SF model with individual-specific slope parameters

$$\begin{aligned} y_{it} &= \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + v_{it} \pm u_{it}, \quad i = 1, \ldots, N, \quad t = 4, \ldots, T_i, & (11)\\ u_{it} &= \omega_i + \omega_{i1}t + \omega_{i2}t^2, & (12) \end{aligned}$$

in which the model parameters are estimated extending the conventional fixed and random-effects panel data estimators. This quadratic specification allows a unit specific temporal pattern of inefficiency but requires the estimation of a large number of parameters ($N \times 3$).

Following a slightly different estimation strategy, Lee and Schmidt (1993) proposed an alternative specification in which the u_{it} are specified as

$$u_{it} = g(t) \cdot u_i, \tag{13}$$

where $g(t)$ is represented by a set of time dummy variables. This specification is more parsimonious than (12) and it does not impose any parametric form, but it is less flexible since it restricts the temporal pattern of u_{it} to be the same for all productive units.[7] Kumbhakar (1990) was the first to propose the ML estimation of a time-varying SF model in which $g(t)$ is specified as

$$g(t) = \left[1 + \exp\left(\gamma t + \delta t^2\right)\right]^{-1}. \tag{14}$$

This model contains only two additional parameters to be estimated, γ and δ and the hypothesis of time-invariant technical efficiency can be easily tested by setting $\gamma = \delta = 0$.

6. The Normal-Exponential model is another straightforward extension allowed by `sfpanel`.

7. Ahn et al. (2005) and Ahn et al. (2001) propose to estimate through a GMM approach the Cornwell et al. (1990) and Lee and Schmidt (1993) models, respectively. They show that GMM is preferable because it is asymptotically efficient. Currently, `sfpanel` allows the estimation of Cornwell et al. (1990) and Lee and Schmidt (1993) models through modified Least Squares Dummy Variables and Iterative Least Squares approaches, respectively. We leave for future updates the implementation of the GMM estimator.

A similar model, termed as "time decay", has been proposed by Battese and Coelli (1992) in which

$$g(t) = \exp\left[-\eta\left(t - T_i\right)\right]. \tag{15}$$

The common feature of all these time-varying SF models is that the intercept α is the same across productive units, thus generating a mis-specification bias in presence of time-invariant unobservable factors, unrelated with the production process but affecting the output. As a result, the effect of these factors may be captured by the inefficiency term, producing biased results.

Greene (2005a) approached this issue through a time-varying SF Normal-Half Normal model with unit-specific intercepts, obtained by replacing (7) by the following specification

$$y_{it} = \alpha_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}. \tag{16}$$

Compared to previous models, this specification allows to disentangle time-varying inefficiency from unit specific time invariant unobserved heterogeneity. For this reason, Greene termed these models as "true" fixed (TFE) or random-effects (TRE), according to the assumptions on the unobserved unit-specific heterogeneity. While the estimation of the true random-effects specification can be easily performed using simulation-based techniques, the ML estimation of the true fixed-effects variant requires the solution of two major issues related to the estimation of nonlinear panel data models. The first is purely computational due to the large dimension of the parameters space. Nevertheless, Greene (2005a,b) showed that a Maximum Likelihood Dummy Variable (MLDV) approach is computationally feasible also in presence of a large number of nuisance parameters α_i ($N > 1000$). The second, the so-called incidental parameters problem, is an inferential issue that arises when the number of units is relatively large compared to the length of the panel. In these cases, the unit-specific intercepts are inconsistently estimated as $N \rightarrow \infty$ with fixed T, since only T_i observations are used to estimate each unit specific parameter (Neyman and Scott 1948; Lancaster 2002). As shown in Belotti and Ilardi (2012), since this inconsistency contaminates the variance parameters, which represent the key ingredients in the postestimation of inefficiencies, the MLDV approach appears to be appropriate only when the length of the panel is large enough ($T \geq 10$).[8]

Although model (16) may appear to be the most flexible and parsimonious choice among the several existing time varying specifications, it can be argued that a portion of the time-invariant unobserved heterogeneity does belong to inefficiency or that these two components should not be disentangled at all. The `sfpanel` command provides options for the estimation of these two extremes: the Schmidt and Sickles (1984), Pitt and Lee (1981) and Battese and Coelli (1988) models in which all time-invariant unobserved

8. A common approach to solve this problem is based on the elimination of the α_i through a data transformation. The consistent estimation of the fixed-effects variant of the Greene's model is still an open research issue in SF literature. Promising solutions have been proposed by Chen et al. (2011) for a homoscedastic Normal-Half Normal model and Belotti and Ilardi (2012) for a more flexible heteroscedastic specification in Normal-Half Normal and Normal-Exponential models. We are currently working to update the `sfpanel` command along these directions.

heterogeneity is considered as inefficiency, and the two "true" specifications in which all time-invariant unobserved heterogeneity is ruled out from the inefficiency component. As pointed out by Greene (2005b), neither formulation is *a priori* completely satisfactory and the choice should be driven by the features of the data at hand.[9]

Despite the usefulness of SF models in many contexts, a practical disclaimer is in order: in both cross-sectional and panel data models, the identification through distributional assumptions of the two components $\boldsymbol{u}$ and $\boldsymbol{v}$ heavily depends on how the shape of their distributions is involved in defining the shape of the $\boldsymbol{\varepsilon}$ distribution. Identification problems may arise when either the shapes are very similar (as pointed out by Ritter and Simar (1997) in the case of small samples for the Normal-Gamma cross-sectional model) or just one of the two components is responsible for most of the shape of the $\boldsymbol{\varepsilon}$ distribution. The latter is the case where the ratio between the inefficiency and measurement error variability (the so-called signal-to-noise ratio, σ_u/σ_v) is very small or very large. In these cases, the profile of the log-likelihood becomes quite "flat", producing non trivial numerical maximization problems.

2.3 Exogenous inefficiency determinants and heteroscedasticity

A very important issue in SF analysis is the inclusion in the model of exogenous variables which are supposed to affect the distribution of inefficiency. These variables, which usually are neither the inputs nor the outputs of the production process, but nonetheless affect the productive unit performance, have been incorporated in a variety of ways: *i)* they may shift the frontier function and/or the inefficiency distribution; *ii)* they may scale the frontier function and/or the inefficiency distribution; *iii)* they may shift and scale the frontier function and/or the inefficiency distribution. Moreover, Kumbhakar and Lovell (2000) stress that, differently from the linear regression model in which the mis-specification of the second moment of the errors distribution determines only efficiency losses, the presence of uncontrolled observable heterogeneity in u_i and/or v_i may affect the inference in SF models. Indeed, while neglected heteroscedasticity in v_i does not produce any bias for the frontier's parameters estimates, it leads to biased inefficiency estimates, as we show in section 5.3.

In this section, we present the approaches that introduce heterogeneity in the location parameter of the inefficiency distribution and/or heteroscedasticity of the inefficiency as well as of the idiosyncratic error term for the models implemented in the `sfcross` and `sfpanel` commands. Since these approaches can be easily extended to the panel data context, we deliberately confine the review to the cross-sectional framework.

As pointed out by Greene (2008), researchers have often incorporated exogenous effects using a two steps approach. In the first step, estimates of inefficiency are obtained without controlling for these factors while in the second, the estimated inefficiency scores are regressed (or otherwise associated) with them. Wang and Schmidt (2002) show

9. A way to disentangle unobserved heterogeneity from inefficiency is to include explanatory variables that are correlated with inefficiency but not with the remaining heterogeneity. The use of (untestable) exclusion restrictions is a quite standard econometric technique to deal with identification issues.

that this approach leads to severely biased results, thus we shall only focus on model extensions based on simultaneous estimation.

A natural starting point for introducing exogenous influences in the inefficiency model is in the location of the distribution. The most well-known approaches are those suggested by Kumbhakar et al. (1991) and Huang and Liu (1994). They proposed to parametrize the mean of the pre-truncated inefficiency distribution. Basically, model (1) - (3) can be completed with

$$u_i \sim \mathcal{N}^+\left(\mu_i, \sigma_u^2\right) \tag{17}$$

$$\mu_i = \boldsymbol{z}_i'\boldsymbol{\psi}, \tag{18}$$

where u_i is a realization from a Truncated Normal random variable, $\boldsymbol{z}_i$ is a vector of exogenous variables (including a constant term) and $\boldsymbol{\psi}$ is the vector of unknown parameters to be estimated (the so-called inefficiency effects). One interesting feature of this approach is that the vector $\boldsymbol{z}_i$ may include interactions with input variables allowing to test the hypothesis that inefficiency is neutral with respect to its impact on input usage.[10]

An alternative approach to analyze the effect of exogenous determinants on inefficiency is obtained by scaling its distribution. Then, a model that allows heteroscedasticity in u_i and/or v_i becomes a straightforward extension. For example, Caudill and Ford (1993), Caudill et al. (1995) and Hadri (1999) proposed to parametrize the variance of the pre-truncated inefficiency distribution in the following way

$$u_i \sim \mathcal{N}^+\left(0, \sigma_{ui}^2\right) \tag{19}$$

$$\sigma_{ui}^2 = \exp\left(\boldsymbol{z}_i'\boldsymbol{\psi}\right). \tag{20}$$

Hadri (1999) extends this last specification by allowing the variance of the idiosyncratic term to be heteroscedastic, so that (3) can be rewritten as

$$v_i \sim \mathcal{N}(0, \sigma_{vi}^2) \tag{21}$$

$$\sigma_{vi}^2 = \exp\left(\boldsymbol{h}_i'\boldsymbol{\phi}\right), \tag{22}$$

where the variables in $\boldsymbol{h}_i$ does not necessarily appear in $\boldsymbol{z}_i$.

As in Wang (2002), both `sfcross` and `sfpanel` allow to combine (17) and (20) for the Normal-Truncated Normal model. In postestimation, it is possible to compute non-monotonic effects of the exogenous factors $\boldsymbol{z}_i$ on u_i. A different specification has been suggested by Wang and Schmidt (2002), in which both the location and variance parameters are "scaled" by the same positive (monotonic) function $h(\boldsymbol{z}_i, \boldsymbol{\psi})$. Their model, $u_i = h(\boldsymbol{z}_i, \boldsymbol{\psi})u_i^*$ with $u_i^* \sim \mathcal{N}\left(\mu, \sigma^2\right)^+$, is equivalent to the assumption that $u_i \sim \mathcal{N}\left(\mu h(\boldsymbol{z}_i, \boldsymbol{\psi}), \sigma^2 h(\boldsymbol{z}_i, \boldsymbol{\psi})^2\right)^+$ in which the $\boldsymbol{z}_i$ vector does not include a constant term.[11]

10. Battese and Coelli (1995) proposed a similar specification for panel data.
11. We are currently working to extend the `sfcross` command allowing for Normal-Truncated Normal models with scaling property (Wang and Schmidt 2002).

Table 1: A summary of `sfcross` and `sfpanel` estimation and postestimation capabilities.

Reference	Model option	$\mathcal{F}$ [1]	Est. method[2]	Loc. eff. in $\boldsymbol{u}$	Heter. in $\boldsymbol{u}$	Heter. in $\boldsymbol{v}$	JLMS[3]	BC[4]	CI[5]
				Cross-sectional models					
Aigner et al. (1977)	als77	HN	ML		x	x	x	x	x
Meeusen and van den Broeck (1977)	mvb77	E	ML	x	x	x	x	x	x
Stevenson (1980)	stev80	TN	ML	x	x	x	x	x	x
Greene (2003)	gre00	G	SML		x	x	x	x	
				Panel data models					
Schmidt and Sickles (1984)	fe	-	W				x		
Schmidt and Sickles (1984)	regls	-	GLS				x		
Pitt and Lee (1981)	mlti	HN	ML				x	x	x
Battese and Coelli (1988)	mlti	TN	ML				x	x	x
Cornwell et al. (1990)	fecss	-	MW				x		
Lee and Schmidt (1993)	fels	-	ILS				x		
Kumbhakar (1990)	kumb90	HN	ML				x	x	x
Battese and Coelli (1992)	bc92	TN	ML				x	x	x
Battese and Coelli (1995)	bc95	TN	ML	x	x	x	x	x	x
Greene (2005a)	tfe	E	MLDV	x	x	x	x	x	x
Greene (2005a)	tfe	HN	MLDV		x	x	x	x	x
Greene (2005a)	tfe	TN	MLDV	x	x	x	x	x	x
Greene (2005a)	tre	E	SML	x	x	x	x	x	x
Greene (2005a)	tre	HN	SML		x	x	x	x	x
Greene (2005a)	tre	TN	SML	x	x	x	x	x	x

[1] Distribution $\mathcal{F}$ of $\boldsymbol{u}$: HN="Half Normal", E="Exponential", TN="Truncated Normal", G="Gamma".

[2] Estimation method: ML="Maximum Likelihood", SML="Simulated Maximum Likelihood", GLS="Generalized Least Squares", W="Within Group", MW="Modified Within Group", ILS="Iterative Least Squares", MLDV="Maximum Likelihood Dummy Variable".

[3] Inefficiency (and efficiency) estimation via the Jondrow et al. (1982) approach.

[4] Efficiency estimation via the Battese and Coelli (1988) approach.

[5] Confidence interval for inefficiencies via the Horrace and Schmidt (1996) approach.

3 The `sfcross` command

The new Stata command **`sfcross`** provides parametric ML estimators of SF models, where the default is represented by production. The general syntax of this commands is as follows

`sfcross` ***depvar*** [*indepvars*] [*if*] [*in*] [*weight*] [**`, options`**]

This command and its panel analog **`sfpanel`** are written using the **`moptimize()`** suite of functions, the optimization engine used by **`ml`**, and share the same features of all Stata estimation commands, including access to the estimation results and options for the maximization process (see **`help maximize`**). Version 11 is the earliest version of Stata that can be used to run the command. **`fweight`**, **`iweight`**, **`aweight`**, and **`pweight`** are allowed (see **`help weight`**). **`sfcross`** supports the **`svy`** prefix (See **`help survey`**). The default is the Normal-Exponential model. Most options are similar to those of other Stata estimation commands. A full description of all available options is provided in the **`sfcross`** help file.

3.1 Main options for `sfcross`

`distribution`(*distname*) specifies the distribution for the inefficiency term as Half Normal (**`hnormal`**), Truncated Normal (**`tnormal`**), Exponential (**`exponential`**) or Gamma (**`gamma`**). The default is the Exponential distribution.

`emean`(*varlist_m* [, noconstant]) may be used only with **`distribution(tnormal)`**. With this option, **`sfcross`** specifies the mean of the Truncated Normal distribution in terms of a linear function of the covariates defined in *varlist_m*. Specifying **`noconstant`** suppresses the constant in this function.

`usigma`(*varlist_u* [, noconstant]) specifies that the technical inefficiency component is heteroscedastic, with the variance expressed as a function of the covariates defined in *varlist_u*. Specifying **`noconstant`** suppresses the constant in this function.

`vsigma`(*varlist_v* [, noconstant]) specifies that the idiosyncratic error component is heteroscedastic, with the variance expressed as a function of the covariates defined in *varlist_v*. Specifying **`noconstant`** suppresses the constant in this function.

`svfrontier`() specifies a 1 x k vector of initial values for the coefficients of the frontier. The vector must have the same length of the parameters vector to be estimated.

`svemean`() specifies a 1 x k_m vector of initial values for the coefficients of the conditional mean model. This option can be specified only with **`distribution(tnormal)`**.

`svusigma`() specifies a 1 x k_u vector of initial values for the coefficients of the technical inefficiency variance function.

`svvsigma`() specifies a 1 x k_v vector of initial values for the coefficients of the idiosyncratic error variance function.

`cost` specifies that `sfcross` fits a cost frontier model.

`simtype`(*simtype*) specifies the method to generate random draws when `dist(gamma)` is specified. `runiform` generates uniformly distributed random variates; `halton` and `genhalton` create respectively Halton sequences and generalized Halton sequences where the base is expressed by the prime number in `base`(#). `runiform` is the default. See `help mata halton()` for more details on Halton sequences generation.

`nsimulations`(#) specifies the number of draws used in the simulation when `distribution(gamma)` is specified. The default is 250.

`base`(#) specifies the number, preferably a prime, used as a base for the generation of Halton sequences and generalized Halton sequences when `distribution(gamma)` is specified. The default is 7. Note that Halton sequences based on large primes (# > 10) can be highly correlated, and their coverage may be worse than that of the pseudorandom uniform sequences.

`postscore` saves an observation-by-observation matrix of scores in the estimation results list. This option is not allowed when the size of the scores' matrix is greater than Stata matrix limit; see `help limits`.

`posthessian` saves the Hessian matrix corresponding to the full set of coefficients in the estimation results list.

3.2 Postestimation command after `sfcross`

After the estimation with `sfcross`, the `predict` command can be used to compute linear predictions, (in)efficiency and score variables. Moreover, the `sfcross` postestimation command allows to compute (in)efficiency confidence interval through the option `ci` as well as non-monotonic marginal effects á la Wang (2002) using, when appropriate, the option `marginal`. The syntax of the command is the following

`predict` [*type*] *newvar* [*if*] [*in*] [, `statistics`]

`predict` [*type*] { *stub*/newvar_xb newvar_v newvar_u* } [*if*] [*in*] , `scores`

where `statistics` includes `xb`, `stdp`, `u`, `m`, `jlms`, `bc`, `ci` and `marginal`.

`xb`, the default, calculates the linear prediction.

`stdp` calculates the standard error of the linear prediction.

`u` produces estimates of inefficiency via $\mathbb{E}(s \cdot u|\varepsilon)$ using the Jondrow et al. (1982) estimator, where s=1 (s=-1) when a production (cost) frontier is estimated.

`m` produces estimates of inefficiency via $\mathbb{M}(s \cdot u|\varepsilon)$, the mode of the conditional distribution of $u|\varepsilon$. This option is not allowed when the estimation is performed with the `distribution(gamma)` option.

`jlms` produces estimates of efficiency via $\exp(-\mathbb{E}(s \cdot u|\varepsilon))$.

`bc` produces estimates of efficiency via $\mathbb{E}\left[\exp(-s \cdot u|\varepsilon)\right]$, the Battese and Coelli (1988) estimator.

`ci` computes confidence interval using the approach proposed by Horrace and Schmidt (1996). It can be used only when `u` or `bc` is specified. The default confidence level is 95, meaning a 95% confidence interval. If the option `level(#)` is used in the previous estimation command, the confidence interval will be computed using the # level. This option creates two additional variables: *newvar_LBcilevel* and *newvar_UBcilevel*, the lower and the upper bound, respectively. This option is not allowed when the estimation is performed with the `distribution(gamma)` option.

`marginal` calculates the marginal effects of the exogenous determinants on $\mathbb{E}(u)$ and Var(u). The marginal effects are observation-specific, and are saved in the new variables *varname_m_M* and *varname_u_V*, the marginal effects on the mean and the variance of the inefficiency, respectively. *varname_m* and *varname_u* are the names of each exogenous determinants specified in options `emean`(*varlist_m* [, noconstant]) and `usigma`(*varlist_u* [, noconstant]). `marginal` can be used only when the estimation is performed with the `distribution(tnormal)` option. When they are both specified, *varlist_m* and *varlist_u* must contain the same variables in the same order. This option can be specified in two ways: i) together with either `u`, `m`, `jlms` or `bc`; ii) alone without specifying *newvar*.

`score` calculates score variables. When the argument of the option `distribution()` is `hnormal`, `tnormal` or `exponential`, scores variables are generated as the derivative of the objective function with respect to the *parameters*. When the argument of the option `distribution()` is `gamma`, they are generated as the derivative of the objective function with respect to the *coefficients*. This difference is due to the different `moptimize()` *evaluator type* used to implement the estimators (See `help mata moptimize()`).

4 The `sfpanel` command

`sfpanel` allows the estimation of SF panel data models through ML and Least Squares (LS) techniques. The general `sfpanel` syntax is the following:

`sfpanel` ***depvar*** [*indepvars*] [*if*] [*in*] [*weight*] [`, options`]

As for its cross-sectional counterpart, version 11 is the earliest version of Stata that can be used to run `sfpanel`. Similarly, all type of weights are allowed but the declared `weight` variable must be constant within each unit of the panel. Moreover, the command does not support the `svy` prefix. The default model is the time-decay model of Battese and Coelli (1992). A description of the main command-specific estimation and postestimation options is provided below. A full description of all available options is provided in the `sfpanel` help file.

4.1 Main options for `sfpanel`

True fixed and random-effects models (Greene 2005a)

`distribution`(*distname*) specifies the distribution for the inefficiency term as Half-Normal (`hnormal`), Truncated Normal (`tnormal`) or Exponential (`exponential`). The default is `exponential`.

`emean`(*varlist_m* [, noconstant]) may be used only with `distribution(tnormal)`. With this option, `sfpanel` specifies the mean of the Truncated Normal distribution in terms of a linear function of the covariates defined in *varlist_m*. Specifying `noconstant` suppresses the constant in this function.

`usigma`(*varlist_u* [, noconstant]) specifies that the technical inefficiency component is heteroscedastic, with the variance expressed as a function of the covariates defined in *varlist_u*. Specifying `noconstant` suppresses the constant in this function.

`vsigma`(*varlist_v* [, noconstant]) specifies that the idiosyncratic error component is heteroscedastic, with the variance expressed as a function of the covariates defined in *varlist_v*. Specifying `noconstant` suppresses the constant in this function.

`feshow` allows the user to display estimates of individual fixed-effects, along with structural parameters. Only for `model`(tfe).

`simtype`(*simtype*) specifies the method to generate random draws for the unit-specific random-effects. `runiform` generates uniformly distributed random variates; `halton` and `genhalton` create respectively Halton sequences and generalized Halton sequences where the base is expressed by the prime number in `base`(#). `runiform` is the default. See `help mata halton()` for more details on Halton sequences generation. Only for `model`(tre).

`nsimulations`(#) specifies the number of draws used in the simulation. The default is 250. Only for `model`(tre).

`base`(#) specifies the number, preferably a prime, used as a base for the generation of Halton sequences and generalized Halton sequences. The default is 7. Note that Halton sequences based on large primes ($\# > 10$) can be highly correlated, and their coverage may be worse than that of the pseudorandom uniform sequences. Only for `model`(tre).

ML random-effects time-varying inefficiency effects model (Battese and Coelli 1995)

`emean`(*varlist_m* [, noconstant]) fits the Battese and Coelli (1995) conditional mean model in which the mean of the Truncated Normal distribution is expressed as a linear function of the covariates specified in *varlist_m*. Specifying `noconstant` suppresses the constant in this function.

`usigma`(*varlist_u* [, noconstant]) specifies that the technical inefficiency component is heteroscedastic, with the variance expressed as a function of the covariates defined in *varlist_u*. Specifying `noconstant` suppresses the constant in this function.

`vsigma`(*varlist_v* [, noconstant]) specifies that the idiosyncratic error component is heteroscedastic, with the variance expressed as a function of the covariates defined in *varlist_v*. Specifying `noconstant` suppresses the constant in this function.

ML random-effects flexible time-varying efficiency model (Kumbhakar 1990)

`bt`(*varlist_bt* [, noconstant]) fits a model that allows a flexible specification of technical inefficiency handling different types of time behavior, using the formulation $u_{it} = u_i \left[1 + \exp(\mathit{varlist\_bt})\right]^{-1}$. Typically, explanatory variables in *varlist_bt* are represented by a polynomial in time. Specifying `noconstant` suppresses the constant in the function. The default includes a linear and a quadratic term in time without constant, as in Kumbhakar (1990).

4.2 Postestimation command after `sfpanel`

After the estimation with `sfpanel`, the `predict` command can be used to compute linear predictions, (in)efficiency and score variables. Moreover, the `sfpanel` postestimation command allows to compute (in)efficiency confidence interval through the option `ci` as well as non-monotonic marginal effects á la Wang (2002) using, when appropriate, the option `marginal`. The syntax of the command is the following

`predict` [*type*] *newvar* [*if*] [*in*] [, `statistics`]

`predict` [*type*] { *stub*/newvar_xb newvar_v newvar_u* } [*if*] [*in*] , `scores`

where `statistics` includes `xb`, `stdp`, `u`, `u0`, `m`, `bc` and `jlms`, `ci`, `marginal` and `trunc`(*tlevel*).

`xb`, the default, calculates the linear prediction.

`stdp` calculates the standard error of the linear prediction.

`u` produces estimates of inefficiency via $\mathbb{E}(s \cdot u|\varepsilon)$ using the Jondrow et al. (1982) estimator, where s=1 (s=-1) when a production (cost) frontier is estimated.

`u0` produces estimates of inefficiency via $\mathbb{E}(s \cdot u|\varepsilon)$ using the Jondrow et al. (1982) estimator when the random-effect is zero. This statistic can be specified only when the estimation is performed with the `model(tre)` option.

`m` produces estimates of inefficiency via $\mathbb{M}(s \cdot u|\varepsilon)$, the mode of the conditional distribution of $u|\varepsilon$. This statistic is not allowed when the estimation is performed with the option `model(fecss)`, `model(fels)`, `model(fe)` or `model(regls)`.

`jlms` produces estimates of efficiency via $\exp(-\mathbb{E}(s \cdot u|\varepsilon))$.

`bc` produces estimates of efficiency via $\mathbb{E}\left[\exp(-s \cdot u|\varepsilon)\right]$, the Battese and Coelli (1988) estimator. This statistic is not allowed when the estimation is performed with the option `model(fecss)`, `model(fels)`, `model(fe)` or `model(regls)`.

`ci` computes confidence interval using the approach proposed by Horrace and Schmidt (1996). This option can be used only with `u`, `jlms` and `bc` statistics, but not when the estimation is performed with the option `model(fels)`, `model(bc92)`, `model(kumb90)`, `model(fecss)`, `model(fe)` or `model(regls)`. The default confidence level is 95, meaning a 95% confidence interval. If the option `level(#)` is used in the previous estimation command, the confidence interval will be computed using the # level. This option creates two additional variables: *newvar_LBcilevel* and *newvar_UBcilevel*, the lower and the upper bound, respectively.

`marginal` calculates the marginal effects of the exogenous determinants on $\mathbb{E}(u)$ and Var(u). The marginal effects are observation-specific and are saved in the new variables *varname_m_M* and *varname_u_V*, the marginal effects on the unconditional mean and variance of inefficiency, respectively. *varname_m* and *varname_u* are the names of each exogenous determinants specified in options `emean(varlist_m [, noconstant])` and `usigma(varlist_u [, noconstant])`. `marginal` can be used only when estimation is performed with the `model(bc95)` option or when the inefficiency in `model(tfe)` or `model(tre)` is `distribution(tnormal)`. When they are both specified, *varlist_m* and *varlist_u* must contain the same variables in the same order. This option can be specified in two ways: i) together with either `u`, `m`, `jlms` or `bc`; ii) alone without specifying *newvar*.

`trunc`(*tlevel*) excludes from the inefficiency estimation the units whose effects are, at least at one time period, in the upper and bottom *tlevel*% range. `trunc()` can be used only if the estimation is performed with `model(fe)`, `model(regls)`, `model(fecss)` and `model(fels)`.

`score` calculates score variables. This option is not allowed when the estimation is performed with the option `model(fecss)`, `model(fels)`, `model(fe)` or `model(regls)`. When the argument of the option `model()` is `tfe` or `bc95`, scores variables are generated as the derivative of the objective function with respect to the *parameters*. When the argument of the option `model()` is `tre`, `bc88`, `bc92`, `kumb90` or `pl81`, they are generated as the derivative of the objective function with respect to the *coefficients*. This difference is due to the different `moptimize()` *evaluator type* used to implement the estimators (See `help mata moptimize()`).

5 Examples with simulated data

In this section, we use simulated data to illustrate `sfcross` and `sfpanel` estimation capabilities, focusing on some of the models that cannot be estimated using official Stata routines.[12]

12. We report the Mata code used for the data-generating process and models' estimation syntax for each example in the `sj_examples_simdata.do` ancillary file.

5.1 The normal-gamma SF production model

There is a large debate in the SF literature about the (non-)identifiability of the Normal-Gamma cross-sectional model. Ritter and Simar (1997) pointed out that this model is difficult to distinguish from the Normal-Exponential one, and that the estimation of the shape parameter of the Gamma distribution may require large sample sizes (up to several thousand observations). On the other hand, Greene (2003) argued that their result "was a matter of degree, not a definitive result" and that the (non-)identifiability of the true value of the shape parameter remains an empirical question. In this section, we illustrate the `sfcross` command by estimating a Normal-Gamma SF production model. We consider the following Data Generating Process (DGP)

$$
\begin{aligned}
y_i &= 1 + 0.3x_{1i} + 0.7x_{2i} + v_i - u_i, \qquad i = 1, \ldots, N, && (23)\\
v_i &\sim \mathcal{N}(0,1), && (24)\\
u_i &\sim \Gamma(2,2), && (25)
\end{aligned}
$$

where the inefficiency is Gamma distributed with shape and scale parameters equal to 2, the idiosyncratic error is $\mathcal{N}(0,1)$ and the two regressors x_{1i} and x_{2i} are normally distributed with zero means and variances equal to 1 and 4, respectively. Notice that the sample size is set to 1000 observations, a large size as noted by Ritter and Simar (1997), but in general not so large given the current availability of micro data. Let us begin by fitting the Normal-Exponential model using the following syntax

```
. sfcross y x1 x2, distribution(exp) nolog

Stoc. frontier normal/exponential model              Number of obs =      1000
                                                     Wald chi2(2)  =    419.88
                                                     Prob > chi2   =    0.0000

Log likelihood = -2423.0869
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Frontier     |
          x1 |   .3709605    .068792     5.39   0.000     .2361306    .5057904
          x2 |   .6810641   .0339945    20.03   0.000     .6144361     .747692
       _cons |  -.1474677   .1131198    -1.30   0.192    -.3691784    .0742431
-------------+----------------------------------------------------------------
Usigma       |
       _cons |   2.173649   .0957468    22.70   0.000     1.985989    2.361309
-------------+----------------------------------------------------------------
Vsigma       |
       _cons |   .3827463   .1498911     2.55   0.011     .0889652    .6765274
-------------+----------------------------------------------------------------
     sigma_u |   2.964844   .1419372    20.89   0.000     2.699305    3.256505
     sigma_v |   1.210911   .0907524    13.34   0.000     1.045487     1.40251
      lambda |   2.448441   .2058941    11.89   0.000     2.044895    2.851986
------------------------------------------------------------------------------

. estimates store exp

. predict uhat_exp, u
```

It is worth noting that the Normal-Exponential model is the `sfcross` *default*, so that we might omit the option `distribution(exponential)`.[13] As can be noted, although there is only one equation to be estimated in the model, the command fits three of Mata's `[M-5] moptimize()` equations (see `help mata moptimize()`). Indeed, given that `sfcross` allows both the inefficiency and the idiosyncratic error to be heteroscedastic (see table 1), the output also reports variance parameters estimated in a transformed metric according to equation (20) and (22), respectively. Since in this example the inefficiency is assumed to be homoscedastic, `sfcross` estimates the coefficient of the constant term in equation (20) rather than estimating directly σ_u. In order to make the output easily interpretable, `sfcross` also displays the variance parameters in their natural metric.

As expected the Normal-Exponential model produces biased results, especially for the frontier's constant term and the inefficiency scale parameter σ_u. We also run the `predict` command using the `u` option. In this way, inefficiencies estimates are obtained through the Jondrow et al. (1982) approach. Since the inefficiencies are drawn from a Gamma distribution, a better fit can be obtained using the following command

```
. sfcross y x1 x2, distribution(gamma) nsim(50) simtype(genha) base(7) nolog

Stoc. frontier normal/gamma model                      Number of obs =      1000
                                                       Wald chi2(2)  =    438.00
                                                       Prob > chi2   =    0.0000

Log simulated-likelihood = -2419.0008
Number of Randomized Halton Sequences = 50
Base for Randomized Halton Sequences  = 7
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Frontier						
x1	.3809637	.0670488	5.68	0.000	.2495506	.5123769
x2	.6877522	.0336089	20.46	0.000	.6218799	.7536244
_cons	.9362409	.412162	2.27	0.023	.1284182	1.744064
Usigma						
_cons	1.535178	.2264704	6.78	0.000	1.091304	1.979051
Vsigma						
_cons	-.2734817	.3330257	-0.82	0.412	-.9262	.3792366
sigma_u	2.154565	.2439726	8.83	0.000	1.725733	2.689958
sigma_v	.8721962	.1452319	6.01	0.000	.6293297	1.208788
lambda	2.470276	.1969658	12.54	0.000	2.08423	2.856321
g_shape	1.879223	.3845289	4.89	0.000	1.125561	2.632886

```
. estimates store gamma
. predict uhat_gamma, u
```

13. The option `nolog` allows to omit the display of the criterion function iteration log. `sfcross` and `sfpanel` allow to use all `maximize` options available for `ml` estimation commands (see `help maximize`) plus the additional options `postscore` and `posthessian`, which report the score and the hessian as an `e()` vector and matrix, respectively.

In the Normal-Gamma cross-sectional model, the parameters are estimated using the Maximum Simulated Likelihood (MSL) technique. A better approximation of the log-likelihood function requires the right choice about the number of draws and the way they are created. In this example, we use generalized Halton sequences (`simtype(genhalton)`) with base equal to 7 (`base(7)`) and only 50 draws (`nsim(50)`). Indeed, a Halton sequence generally have a more uniform coverage than a sequence generated from pseudouniform random numbers. Moreover, as noted by Greene (2003), the computational efficiency compared to pseudouniform random draws appears to be at least 10 to 1, so that in our example the same results can be approximately obtained using 500 pseudouniform draws (See `help mata halton()`).[14]

As expected, in this example the parameters of the Normal-Gamma model are properly estimated. Furthermore, this model is preferable to the Normal-Exponential one, as corroborated by the following likelihood ratio test[15]

```
. lrtest exp gamma
Likelihood-ratio test                                 LR chi2(1)  =      8.17
(Assumption: exp nested in gamma)                     Prob > chi2 =    0.0043
```

Similar conclusions may be drawn by comparing the estimated mean inefficiencies with the true simulated one, even if the Spearman rank correlation with the latter is high and very similar for both `uhat_gamma` and `uhat_exp`.[16]

```
. summarize u uhat_gamma uhat_exp
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           u |      1000    4.097398     2.91035   .0259262   19.90251
  uhat_gamma |      1000    4.048885    2.839368   .4752663   20.27557
    uhat_exp |      1000    2.964844     2.64064    .363516   18.95619
. spearman u uhat_gamma uhat_exp
(obs=1000)
             |        u uhat_g~a uhat_exp
-------------+---------------------------
           u |   1.0000
  uhat_gamma |   0.9141   1.0000
    uhat_exp |   0.9145   0.9998   1.0000
```

14. For all models estimated using MSL, `sfcross` and `sfpanel` *default* options are `simtype(uniform)` with `nsim(250)`. In our opinion, small values (e.g., 50 for Halton sequences and 250 for pseudouniform random draws) are sufficient for exploratory work. On the other hand larger values, in the order of several hundreds, are advisable to get more precise results. Our advise is to use Halton sequences rather than pseudorandom random draws. However, as pointed out by Drukker and Gates (2006), "Halton sequences based on large primes (`d` > 10) can be highly correlated, and their coverage can be worse than that of the pseudorandom uniform sequences".

15. Notice that `exp` and `gamma` are the names of the Exponential and Gamma models' estimation results saved using the `estimates store` command.

16. In line with Ritter and Simar (1997), our simulation results indicate that in the Normal-Gamma model a relatively large samples is needed to achieve a reasonable degree of precision in the estimates of inefficiency distribution parameters.

5.2 Panel data time-varying inefficiency models

Cornwell et al. (1990) and Lee and Schmidt (1993) provide a fixed-effect treatment of models like those proposed by Kumbhakar (1990) and Battese and Coelli (1992). Currently, `sfpanel` allows the estimation of Cornwell et al. (1990) and Lee and Schmidt (1993) models by means of Modified Least Squares Dummy Variables (MLSDV) and Iterative Least Squares (ILS), respectively. An interesting aspect of these models is that, although they have been proposed in the SF literature, actually they are linear panel data models with time-varying fixed-effects, thus potentially very useful also in other contexts. However, their consistency requires white noise errors and they are less efficient than the GMM estimator proposed by Ahn et al. (2001) and Ahn et al. (2005).

In this section, we report the main syntax to estimate such models. We start specifying the following stochastic production frontier *translog* model

$$\begin{aligned} y_{it} &= u_{it} + 0.2x_{1it} + 0.6x_{2it} + 0.6x_{3it} + 0.2x_{1it}^2 + 0.1x_{2it}^2 + 0.2x_{3it}^2 + \\ &\quad +0.15x_{1it}x_{2it} - 0.3x_{1it}x_{3it} - 0.3x_{2it}x_{3it} + v_{it}, \qquad (26) \\ v_{it} &\sim \mathcal{N}(0, 0.25), \quad i = 1, \ldots, n, \;\; t = 1, \ldots, T. \qquad (27) \end{aligned}$$

As already mentioned, the main feature of these models is the absence of any distributional assumption about inefficiency. In this example, the DGP follows the Lee and Schmidt (1993) model, where $u_{it} = \delta_i \boldsymbol{\xi}$. For each unit, the parameter δ_i is drawn from a uniform distribution in $[0, \sqrt{12\tau + 1} - 1]$ with $\tau = 0.8$. The elements of the vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_T)$ are equally spaced between -2 and 2. This set-up implies a standard deviation of the inefficiency term $\sigma_u \approx 1.83$.

Once the sample is declared to be a panel (see `help xtset`), the Lee and Schmidt (1993) and the Cornwell et al. (1990) models can be estimated using the following syntaxes

```
. sfpanel y x1 x2 x3 x1_sq x2_sq x3_sq x1_x2 x1_x3 x2_x3, model(fels)
  (output omitted)
. estimates store fels
. predict uhat_fels, u

. sfpanel y x1 x2 x3 x1_sq x2_sq x3_sq x1_x2 x1_x3 x2_x3, model(fecss)
  (output omitted)
. estimates store fecss
. predict uhat_fecss, u
```

Notice that we use the `predict` command with the `u` option to post-estimate inefficiency. As an additional source of comparison, we use the same simulated data to assess the behavior of the Schmidt and Sickles (1984) time-invariant inefficiency model. The fixed-effects version of this model can be estimated using `sfpanel` as well as the official `xtreg` command. However, when the estimation is performed using `sfpanel`, the `predict` command with the aforementioned option `u` can be used to obtain inefficiency estimates[17]

17. Both `xtreg` and `sfpanel` also allow the estimation of the random-effects version of this model through the FGLS approach.

```
. sfpanel y x1 x2 x3 x1_sq x2_sq x3_sq x1_x2 x1_x3 x2_x3, model(fe)
  (output omitted)
. estimates store fess_sf
. predict uhat_fess, u

. xtreg y x1 x2 x3 x1_sq x2_sq x3_sq x1_x2 x1_x3 x2_x3, fe
  (output omitted)
. estimates store fess_xt
```

Table 2 reports the estimation results from the three models. Unsurprisingly, both the frontier and variance parameters are well estimated in the `ls93` and `css90` models. This result shows that, when the DGP follows the model by Lee and Schmidt, the estimator by Cornwell, Schmidt, and Sickles provides reliable results. On the other hand, being the data generated from a time-varying model, variance estimates from the `ss84` model show a substantial bias.

Table 2: Schmidt and Sickles (`ss84`), Cornwell, Schmidt, and Sickles (`css90`) and Lee and Schmidt (`ls93`) estimation results

	ss84	**css90**	**ls93**
x1	0.254 ***	0.185 ***	0.171 ***
	(0.0695)	(0.0167)	(0.0230)
x2	0.626 ***	0.619 ***	0.611 ***
	(0.0354)	(0.0085)	(0.0117)
x3	0.602 ***	0.591 ***	0.596 ***
	(0.0220)	(0.0052)	(0.0075)
x1_sq	0.193 ***	0.204 ***	0.209 ***
	(0.0234)	(0.0055)	(0.0076)
x2_sq	0.099 ***	0.103 ***	0.101 ***
	(0.0080)	(0.0019)	(0.0026)
x3_sq	0.198 ***	0.201 ***	0.201 ***
	(0.0036)	(0.0008)	(0.0012)
x1_x2	0.149 ***	0.142 ***	0.145 ***
	(0.0198)	(0.0047)	(0.0064)
x1_x3	-0.293 ***	-0.295 ***	-0.295 ***
	(0.0130)	(0.0030)	(0.0043)
x2_x3	-0.306 ***	-0.300 ***	-0.301 ***
	(0.0076)	(0.0018)	(0.0025)
_cons	-0.050		
	(0.0866)		
σ_u	0.223	1.859	1.832
σ_v	2.096	0.352	0.497

We do not expect large differences in terms of inefficiency scores, given the similarities in terms of variance estimates between `css90` and `ls93`. It is worth noting that for these models (including also `ss84`), inefficiency scores are retrieved in postestimation assuming that the best decision making unit is fully efficient.[18] As it can be seen from the following `summarize` command, both `css90` and `ls93` average inefficiencies are close to the true values, while Spearman rank correlations are almost equal to 1. As expected, the `ss84` estimated inefficiencies are highly biased and the corresponding units' ranking

18. This assumption involves calculating $\hat{u}_i = \hat{\alpha} - \hat{\alpha}_i$ with $\hat{\alpha} = \max_{i=1,\ldots,n}(\hat{\alpha}_i)$, normalizing the frontier in terms of the best unit in the sample.

completely unreliable.

```
. summarize u uhat_fels uhat_fecss uhat_fess

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           u |       2500    4.510559    1.828692          0   9.021117
  uhat_fels  |       2500    5.068159    1.832078          0   10.11807
 uhat_fecss  |       2500    5.510969    1.859123          0   10.89882
  uhat_fess  |       2500     .645184    .2232496          0    1.27254

. spearman u uhat_fels uhat_fecss uhat_fess
(obs=2500)

             |        u uhat_~ls uhat~css uhat~ess
-------------+------------------------------------
           u |   1.0000
  uhat_fels  |   0.9824   1.0000
 uhat_fecss  |   0.9603   0.9652   1.0000
  uhat_fess  |   0.0000   0.0061   0.1331   1.0000
```

Finally, we show additional features of `sfpanel`, namely: *i)* the possibility to compute elasticities via the official `lincom` command; *ii)* the possibility to perform a constrained fixed-effects estimation, which is not yet available with `xtreg`.

With respect to the former point, it is well known that parameters in a *translog* production frontier do not represent output elasticities. In particular, a linear combination of frontier parameters is needed for computing such elasticities. Moreover, in order to calculate output elasticities at means, we first need to compute and store the mean for each input variable using the following syntax

```
. quietly summarize x1
. scalar x1m = r(mean)
. quietly summarize x2
. scalar x2m = r(mean)
. quietly summarize x3
. scalar x3m = r(mean)
```

Then, the `lincom` command can be used to combine estimated frontier parameters using the following standard syntax

```
 ( 1)  x1 + 1.108946*x1_sq + 1.074533*x1_x2 + 1.05167*x1_x3 = 0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3203578     .05348     5.99   0.000     .2154752    .4252405
------------------------------------------------------------------------------

. lincom x2 + x2_sq * x2m + x1_x2*x1m + x2_x3*x3m
 ( 1)  x2 + 1.074533*x2_sq + 1.108946*x1_x2 + 1.05167*x2_x3 = 0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
             |
```

(*Continued on next page*)

```
         (1) |   .5751999   .0254143    22.63   0.000     .5253585    .6250413
------------------------------------------------------------------------------

. lincom x3 + x3_sq * x3m + x1_x3*x1m + x2_x3*x2m
 ( 1)  x3 + 1.05167*x3_sq + 1.108946*x1_x3 + 1.074533*x2_x3 = 0
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .156379   .0158945     9.84   0.000     .1252075    .1875505
------------------------------------------------------------------------------
```

Finally, the Constant Return to Scale (CRS) hypothesis can be trivially tested by using the following syntax

```
. lincom (x1 + x1_sq * x1m + x1_x2*x2m + x1_x3*x3m) ///
>          + (x2 + x2_sq * x2m + x1_x2*x1m + x2_x3*x3m) ///
>          + (x3 + x3_sq * x3m + x1_x3*x1m + x2_x3*x2m) - 1
 ( 1)  x1 + x2 + x3 + 1.108946*x1_sq + 1.074533*x2_sq + 1.05167*x3_sq +
       2.18348*x1_x2 + 2.160617*x1_x3 + 2.126204*x2_x3 = 1
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0519367   .0609852     0.85   0.395    -.0676648    .1715383
------------------------------------------------------------------------------
```

In this example, the CRS hypothesis cannot be rejected. In order to run a constrained fixed-effects estimation, the required set of constraints to impose CRS may be defined through the official Stata command `constraint` using the following syntax

```
. /// Constraints definition
. constraint define 1 x1 + x2 + x3 = 1
. constraint define 2 x1_sq + x1_x2 + x1_x3 = 0
. constraint define 3 x2_sq + x1_x2 + x2_x3 = 0
. constraint define 4 x3_sq + x1_x3 + x2_x3 = 0
```

Then, the constrained model can be estimated using `sfpanel` with the options `model(fe)` and `constraints(1 2 3 4)`

```
. sfpanel y x1 x2 x3 x1_sq x2_sq x3_sq x1_x2 x1_x3 x2_x3, model(fe) constraints
> (1 2 3 4)
Time-invariant fixed-effects model (LSDV)          Number of obs =      2500
Group variable: id                               Number of groups =       500
Time variable: time                             Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5

 ( 1)  x1 + x2 + x3 = 1
 ( 2)  x1_sq + x1_x2 + x1_x3 = 0
 ( 3)  x2_sq + x1_x2 + x2_x3 = 0
 ( 4)  x3_sq + x1_x3 + x2_x3 = 0
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

(*Continued on next page*)

x1	.3530365	.0851901	4.14	0.000	.1860671	.520006
x2	.5092917	.0434568	11.72	0.000	.4241179	.5944655
x3	.1376718	.0270375	5.09	0.000	.0846792	.1906644
x1_sq	-.0343576	.0287476	-1.20	0.232	-.0907019	.0219868
x2_sq	.1282553	.0098209	13.06	0.000	.1090067	.1475039
x3_sq	.21594	.004442	48.61	0.000	.2072339	.2246461
x1_x2	.0610211	.0242651	2.51	0.012	.0134624	.1085799
x1_x3	-.0266635	.0159577	-1.67	0.095	-.0579401	.0046131
x2_x3	-.1892764	.0092834	-20.39	0.000	-.2074716	-.1710813
_cons	.2326412	.1062126	2.19	0.029	.0244682	.4408141
sigma_u	.7140381					
sigma_v	2.5700643					

It is worth noting that the constrained frontier estimates are more biased than the unconstrained ones, but are still not too far from the true values. This is an artifact of our DGP since the scale elasticity has been simulated without imposing CRS.

5.3 "True" fixed and random-effects models

As already discussed in section 2.2, the "true" fixed and random-effects models allow to disentangle time-invariant heterogeneity from time-varying inefficiency. In this section, we present the main syntax and some of the options useful to estimate such models. We start our exercise by specifying the following Normal-Exponential stochastic production frontier model

$$y_{it} = 1 + \alpha_i + 0.3x_{1it} + 0.7x_{2it} + v_{it} - u_{it}, \tag{28}$$

$$v_{it} \sim \mathcal{N}(0,1), \tag{29}$$

$$u_{it} \sim \mathcal{E}(2), \quad i = 1, \ldots, n, \quad t = 1, \ldots, T. \tag{30}$$

where the nuisance parameters α_i $(i = 1, \ldots, n)$ are drawn from a $\mathcal{N}(0, \theta^2)$ with $\theta = 1.5$. In the fixed-effects design (TFE_{DGP}), the two regressors x_{1it} and x_{2it} are distributed for each unit according to a Normal distribution centered in the corresponding unit-effect α_i with variances equal to 1 and 4, respectively. This design ensures correlation between regressors and individual effects, a typical scenario in which the fixed-effects specification represents the consistent choice.[19]

As far as the random-effects design is concerned (TRE_{DGP}), x_{1it} and x_{2it} are not correlated with the unit-specific effects and are distributed according to a Normal distribution with zero mean and variances equal to 1 and 4, respectively.

The generated sample consists of a balanced panel of 1,000 units observed for 10 periods, for a total of 10,000 observations. Once the sample is declared to be a panel, we estimate the following models: *i)* a Normal-Exponential TFE model on TFE_{DGP} data (`tfe1`) [20]

19. Notice that, higher values of θ correspond to higher correlations between the regressors and the unit-specific effects.

20. Note that `yf`, `x1_c` and `x2_c` are the variables from the TFE_{DGP} while `yr`, `x1_nc` and `x2_nc` are from

```
. sfpanel yf x1_c x2_c, model(tfe) distribution(exp) rescale
  (output omitted)
. estimate store tfe_c
. predict u_tfe_c, u
```

ii) a Normal-Exponential TRE model on TFE_{DGP} data (`tre1`)

```
. sfpanel yf x1_c x2_c, model(tre) distribution(exp) nsim(50) simtype(genhalton)
> base(7) rescale
  (output omitted)
. estimate store tre_c
. predict u_tre_c, u
```

iii) a Normal-Exponential TRE model on TRE_{DGP} data (`tre2`)

```
. sfpanel yr x1_nc x2_nc, model(tre) distribution(exp) nsim(50) simtype(genhalton)
> base(7) rescale
  (output omitted)
. estimate store tre_nc
. predict u_tre_nc, u
. predict u0_tre_nc, u0
```

As shown in the first column of table 3, when the model is correctly specified, the frontier parameters are properly estimated. However, in this example, the MLDV estimator of σ_v is slightly biased by the incidental parameter problem even if the length of the panel is quite large.[21] This problem does not seem to affect variance estimates in the `tre1` model. In this case, the parameters are estimated using the MSL technique assuming that *i*) the unobserved heterogeneity is distributed as $\mathcal{N}(0, \theta^2)$ (where θ represents the standard deviation of the unobserved heterogeneity), and *ii*) $\mathbb{E}(\alpha_i|x_{1it}, x_{2it}) = 0$. Thus, since the estimates are obtained using the TFE_{DGP} data, the frontier and θ parameter estimates are biased.

Table 3: TFE and TRE estimation results

	tfe1	**tre1**	**tre2**
x1_c	0.304 ***	0.776 ***	
	(0.0164)	(0.0198)	
x2_c	0.700 ***	0.811 ***	
	(0.0081)	(0.0094)	
x1_nc			0.295 ***
			(0.0176)
x2_nc			0.706 ***
			(0.0089)
cons		1.062 ***	1.090 ***
		(0.0342)	(0.0540)
σ_u	2.075	2.035	2.023
σ_v	0.770	1.095	0.973
θ		0.602	1.542

On the contrary, by estimating a correctly specified TRE model on TRE_{DGP} data (column `tre2` in table 3), all parameters, including the frontier ones, are accurately estimated.

the TRE_{DGP}.

21. See section 2.2 for a discussion of the MLDV estimator problems in the TFE model.

After each estimation, we use the `predict` command in order to obtain inefficiency estimates. As already mentioned, option `u` instructs the postestimation routine to compute inefficiencies through the Jondrow et al. (1982) estimator (see `help sfpanel postestimation`). Notice that, in the case of the TRE model, the `predict` command also allows the option `u0` to estimate inefficiencies assuming the random-effects are zero. At this point, we can `summarize` the estimated inefficiencies to compare them with the actual values

```
. summarize u u_tfe_c u_tre_c u_tre_nc u0_tre_nc
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           u |     10000    2.004997     2.00852   .0003777   20.83139
     u_tfe_c |     10000    2.075017    1.948148   .2008319   20.42197
     u_tre_c |     10000    2.034946    1.818154   .2430926   18.76244
    u_tre_nc |     10000    2.025002    1.831147   .2656734   19.98998
   u0_tre_nc |     10000    2.200728    2.086419   .1338621   19.47739
. spearman u u_tfe_c u_tre_c u_tre_nc u0_tre_nc
(obs=10000)
             |        u  u_tfe_c  u_tre_c u_tre_nc u0_tre~c
-------------+---------------------------------------------
           u |   1.0000
     u_tfe_c |   0.7654   1.0000
     u_tre_c |   0.7541   0.9291   1.0000
    u_tre_nc |   0.7700   0.9925   0.9464   1.0000
   u0_tre_nc |   0.6297   0.7313   0.8168   0.7965   1.0000
```

All the JLMS estimates are very close to the true simulated ones (`u`). Actually, the estimated average inefficiency after a correctly specified TRE model shows a lower bias than the estimated average inefficiency after a correctly specified TFE model. This is a consequence of the incidental parameters problem. It is worth noting also the good performances of the TRE model when it is fitted on the TFE_{DGP} data (`u_tre_c`).

Introducing heteroscedasticity

Finally, we deal with the problem of heteroscedasticity, a very important issue for applied research. In what follows, we adopt the same presentation strategy. For both TFE and TRE models, we compare the estimates obtained from a model that neglects heteroscedasticity with those obtained from a heteroscedastic one. In order to introduce heteroscedasticity, equations (29)-(30) are replaced by the following

$$\begin{aligned} v_{it} &\sim \mathcal{N}(0, \sigma_{vit}), & (31)\\ u_{it} &\sim \mathcal{E}\left(\sigma_{uit}\right), & (32)\\ \sigma_{vit} &= \exp\left[0.5(1 + .5 \times zv_{it})\right], & (33)\\ \sigma_{uit} &= \exp\left[0.5(2 + 1 \times zu_{it})\right], & (34) \end{aligned}$$

where both inefficiency and idiosyncratic error scale parameters are now a function of a constant term and of an exogenous covariate (zu_{it} and zv_{it}) drawn from a standard normal random variable. Notice that, due to the introduction of heteroscedasticity, we will deal with "average" σ_u and σ_v, which in our simulated sample are approximately 3.1

and 1.7, respectively. In this case, each observation has a different signal-to-noise ratio, implying an average of about 1.9. We estimate four different models: *i)* a homoscedastic TFE model on heteroscedastic TFE_{DGP} data (`tfe1`)

```
. sfpanel yf x1_c x2_c, model(tfe) distribution(exp) rescale
  (output omitted)
. estimates store tfe_hom
. predict u_tfe_hom, u
```

ii) a heteroscedastic TFE model on heteroscedastic TFE_{DGP} data (`tfe2`)

```
. sfpanel yf x1_c x2_c, model(tfe) distribution(exp) usigma(zu) vsigma(zv)
  (output omitted)
. estimates store tfe_het
. predict u_tfe_het, u
```

iii) a homoscedastic TRE model on heteroscedastic TRE_{DGP} data (`tre1`)

```
. sfpanel yr x1_nc x2_nc, model(tre) distribution(exp)  ///
>      nsim(50) simtype(genhalton) base(7) rescale
  (output omitted)
. estimates store tre_hom
. predict u_tre_hom, u
```

vi) a heteroscedastic TRE model on heteroscedastic TRE_{DGP} data (`tre2`)

```
. sfpanel yr x1_nc x2_nc, model(tre) distribution(exp) usigma(zu) vsigma(zv) ///
>     nsim(50) simtype(genhalton) base(7) rescale
  (output omitted)
. estimates store tre_het
. predict u_tre_het, u
. predict u0_tre_het, u0
```

Estimation results are reported in table 4. As expected, `tfe1` variance parameter estimates are biased by both the incidental parameters problem and the neglected heteroscedasticity in $\boldsymbol{u}$ and $\boldsymbol{v}$. These estimates can be significantly improved by taking into account both sources of heteroscedasticity using the options `usigma(`*varlist*`)` and `vsigma(`*varlist*`)` (`tfe2`). Exactly the same argument applies in the TRE case (`tre1` VS `tre2`), but without incidental parameters problem.

As we have mentioned in section 2.3, neglecting heteroscedasticity in $\boldsymbol{u}$ and/or $\boldsymbol{v}$ leads to biased inefficiency estimates. This conclusion is confirmed by the following `summarize` command

```
. summarize u u_tfe_hom u_tfe_het u_tre_hom u_tre_het u0_tre_het

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           u |     10000    3.091925    3.915396    .000169   52.20689
   u_tfe_hom |     10000    3.717061    3.941147   .3442658   51.54804
   u_tfe_het |     10000    3.271297    3.828366   .2642199   52.06564
   u_tre_hom |     10000    3.641955    3.788298   .3739219   51.76109
   u_tre_het |     10000    3.173224    3.709123   .3241621   51.83721
-------------+--------------------------------------------------------
  u0_tre_het |     10000      3.2855    3.844297   .1828969    54.2632
```

Table 4: TFE and TRE estimation results (homoscedasticity VS heteroscedasticity)

	tfe1	**tfe2**	**tre1**	**tre2**
x1_c	0.324 ***	0.295 ***		
	(0.0271)	(0.0245)		
x2_c	0.723 ***	0.732 ***		
	(0.0134)	(0.0121)		
x1_nc			0.316 ***	0.310 ***
			(0.0290)	(0.0264)
x2_nc			0.681 ***	0.689 ***
			(0.0147)	(0.0135)
cons			1.576 ***	1.113 ***
			(0.0637)	(0.0652)
σ_u	3.717	3.264	3.642	3.168
σ_v	1.185	1.402	1.526	1.693
θ			1.579	1.565

The average inefficiency is upward biased (by about 15%) for both TFE and TRE models in which heteroscedasticity has been neglected. A slightly better result is obtained also in terms of Spearman rank correlation.

```
. spearman u u_tfe_hom u_tfe_het u_tre_hom u_tre_het u0_tre_het
(obs=10000)

             |        u u_tfe_~m u_tfe_~t u_tre_~m u_tre_~t u0_tre~t
-------------+------------------------------------------------------
           u |   1.0000
   u_tfe_hom |   0.7287   1.0000
   u_tfe_het |   0.7536   0.9589   1.0000
   u_tre_hom |   0.7380   0.9830   0.9531   1.0000
   u_tre_het |   0.7623   0.9461   0.9835   0.9642   1.0000
  u0_tre_het |   0.7039   0.8173   0.8455   0.8944   0.9121   1.0000
```

6 Empirical applications

In this section we illustrate `sfcross` and `sfpanel` capabilities through two empirical applications from the SF literature. The first analyzes Switzerland railways cost inefficiency using data from the Swiss Federal Office of Statistics on public transport companies, while the second focuses on Spanish diary farms technical inefficiency using data from a voluntary Record Keeping Program.[22]

6.1 Swiss railways

This application is based on a unbalanced panel of 50 railway companies from 1985 to 1997, resulting in 605 observations. We think that this application is interesting for at least two reasons: a) cost frontiers are much less diffuse in the literature compared to production frontiers, given the lack of reliable cost and price data; b) the length of the

22. Both data sets are freely available from the webpage of prof. William Greene (http://people.stern.nyu.edu/wgreene/).

panel makes this database quite unusual in the SF literature. A detailed description of the Switzerland railways transport system and complete information on the variables used are available in Farsi et al. (2005).

In order to estimate a Cobb-Douglas cost frontier we impose linear homogeneity by normalizing total costs and prices through the price of energy. Therefore, the model can be written as

$$\begin{aligned}\ln\left(\frac{TC_{it}}{Pe_{it}}\right) &= \beta_0 + \beta_Y \ln Y_{it} + \beta_Q \ln Q_{it} + \beta_N \ln N_{it} + \\ &+ \beta_{Pk} \ln\left(\frac{Pk_{it}}{Pe_{it}}\right) + \beta_{Pl} \ln\left(\frac{Pl_{it}}{Pe_{it}}\right) + \sum_{t=1986}^{1997} \beta_t dyear_t + u_{it} + v_{it}, \quad (35)\end{aligned}$$

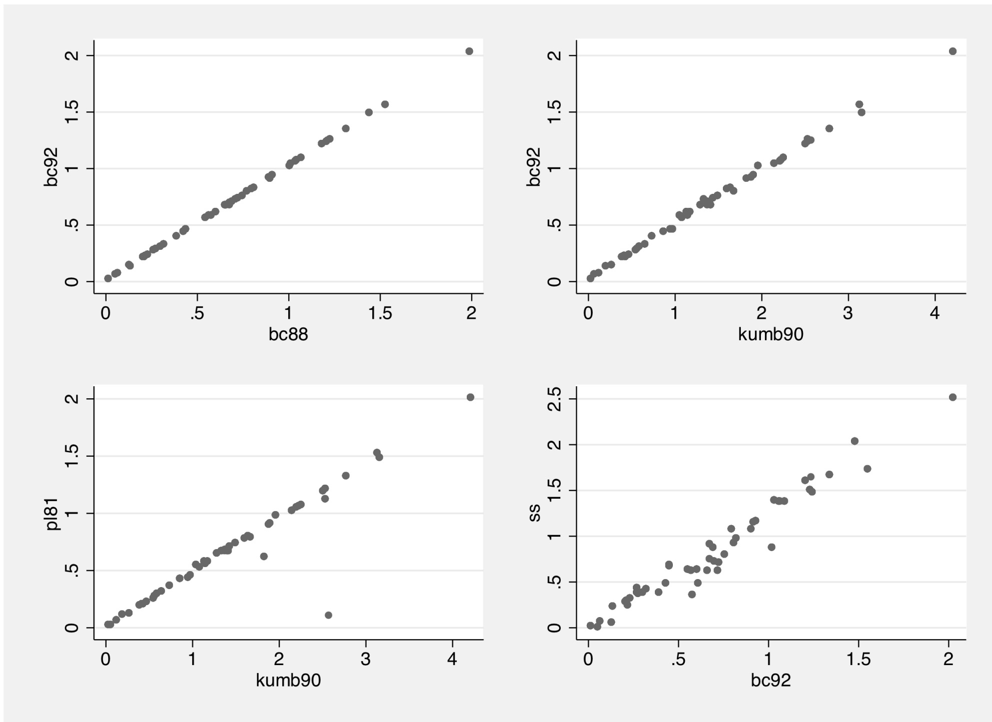
where i and t are the subscripts denoting the railway company and year, respectively. As common, u_{it} is interpreted as a measure of cost inefficiency. Two output measures are included in the cost function: passenger output and freight output. Length of network is included as output characteristic. Further, we have price data for three inputs: capital, labor and energy. All monetary values, including total costs, are in 1997 Swiss Francs (CHF). We have included also a set of time dummies, $dyear_t$, to control for unobserved time dependent variation in costs.

We consider three time-varying inefficiency specifications, that is the Kumbhakar (1990) model (`kumb90`), the Battese and Coelli (1992) model (`bc92`) and the Greene (2005a) random-effects model (`tre`), and three time-invariant models. With respect to the latter group, we estimate the fixed-effects version of the Schmidt and Sickles (1984) model (`ss84`), the Pitt and Lee (1981) (`pl81`) and the Battese and Coelli (1988) (`bc88`) specifications. All models are estimated assuming that the inefficiency is half-normally distributed, with the exception of `bc88` and `bc92` in which $u \sim \mathcal{N}^+(\mu, \sigma_u^2)$ and the `ss84` model in which no distributional assumption is made. The choice of including also the Greene's specification is driven by the multi-output technology that characterizes a railway company, for which unmeasured quality, captured by the random-effects, may play an important role in the production process. Finally, as a benchmark, we estimate a pooled cross-sectional model (`pcs`).

Table 5 shows the results. Coefficients estimates of input prices and outputs are all significant across the seven models, and with the expected signs (positive marginal costs and positive own-price elasticities). Looking at table 6, we further observe that the three time-invariant specifications provide inefficiency estimates that are highly correlated. Perhaps the most interesting result comes from the fact that inefficiency scores obtained from `kumb90` and `bc92` models are also highly correlated with those coming from time-invariant models (table 6 and figure 1). This is not surprising since the two time-invariance hypotheses, $H_0 : t = t^2 = 0$ in the `kumb90` model and $H_0 : \eta = 0$ in `bc92` specification, cannot be rejected at 5% level. Hence, we may conclude that there is evidence of time-invariant technical inefficiency in the Switzerland railways transport system, at least for the study period.

Consistently with this result, we also find that the `tre` model provides inefficiency estimates which have no link with those obtained from any of the other models. More-

Figure 1: Swiss railways, inefficiencies scatterplots



over, due to a very low estimate of the inefficiency variance, the estimated signal-to-noise ratio $\hat{\lambda}$ is the lowest one. In our opinion, these results are driven from the peculiar time-varying inefficiency specification of this model. Indeed, when the inefficiency term is constant over time, the `tre` specification does not allow to disentangle time-invariant unobserved heterogeneity from inefficiency. The variance parameters support this interpretation, since the estimated standard deviation of the random-effects (θ) dominates the inefficiencies one.

Table 5: Swiss railways, estimation results (50 firms for a total of 605 observations)

	pcs	**ss**	**pl81**	**bc88**	**kumb90**	**bc92**	**tre**
	b/se	b/se	b/se	b/se	b/se	b/se	b/se
lnY	0.492***	0.114***	0.200***	0.199***	0.193***	0.199***	0.324***
	(0.015)	(0.032)	(0.034)	(0.033)	(0.033)	(0.033)	(0.019)
lnQ	0.030***	0.014*	0.021***	0.021***	0.020***	0.020***	0.034***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
lnN	0.393***	0.448***	0.485***	0.503***	0.477***	0.499***	0.609***
	(0.027)	(0.051)	(0.045)	(0.047)	(0.044)	(0.047)	(0.049)
lnpk	0.171***	0.318***	0.310***	0.311***	0.311***	0.313***	0.294***
	(0.032)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.020)
lnpl	0.592***	0.546***	0.548***	0.546***	0.538***	0.543***	0.538***
	(0.074)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)	(0.039)
dyear1986	0.009	0.010	0.009	0.009	0.015	0.008	0.011
	(0.056)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
dyear1987	0.003	0.020	0.012	0.012	0.023	0.009	0.004
	(0.056)	(0.015)	(0.015)	(0.015)	(0.017)	(0.015)	(0.016)
dyear1988	0.010	0.039*	0.028	0.027	0.043*	0.023	0.017
	(0.057)	(0.015)	(0.015)	(0.015)	(0.019)	(0.016)	(0.016)
dyear1989	0.036	0.065***	0.052***	0.052***	0.070***	0.046**	0.040*
	(0.057)	(0.016)	(0.016)	(0.016)	(0.021)	(0.016)	(0.016)
dyear1990	0.024	0.084***	0.068***	0.068***	0.086***	0.060***	0.054**
	(0.058)	(0.016)	(0.016)	(0.016)	(0.022)	(0.017)	(0.017)
dyear1991	0.030	0.098***	0.078***	0.078***	0.096***	0.069***	0.059***
	(0.058)	(0.017)	(0.018)	(0.017)	(0.024)	(0.019)	(0.018)
dyear1992	0.046	0.111***	0.094***	0.094***	0.109***	0.083***	0.078***
	(0.058)	(0.017)	(0.017)	(0.017)	(0.023)	(0.019)	(0.018)
dyear1993	0.015	0.100***	0.081***	0.081***	0.092***	0.069***	0.062***
	(0.057)	(0.017)	(0.017)	(0.017)	(0.023)	(0.020)	(0.017)
dyear1994	-0.001	0.082***	0.063***	0.063***	0.069**	0.049*	0.042*
	(0.056)	(0.017)	(0.017)	(0.017)	(0.022)	(0.020)	(0.017)
dyear1995	0.019	0.059***	0.048**	0.047**	0.045*	0.031	0.032
	(0.057)	(0.016)	(0.016)	(0.016)	(0.022)	(0.021)	(0.017)
dyear1996	0.027	0.037*	0.028	0.027	0.018	0.010	0.019
	(0.057)	(0.017)	(0.016)	(0.016)	(0.022)	(0.022)	(0.018)
dyear1997	0.019	0.038*	0.030	0.029	0.009	0.009	0.016
	(0.060)	(0.018)	(0.017)	(0.017)	(0.023)	(0.024)	(0.019)
Constant	-8.310***	-2.682***	-4.895***	-4.929***	-4.626***	-4.871***	-6.577***
	(0.976)	(0.652)	(0.643)	(0.634)	(0.637)	(0.637)	(0.505)
t	-	-	-	-	0.023	-	-
	-	-	-	-	(0.015)	-	-
t^2	-	-	-	-	-0.002	-	-
	-	-	-	-	(0.001)	-	-
η	-	-	-	-	-	-0.002	-
	-	-	-	-	-	(0.002)	-
λ	2.882	7.900	11.366	7.716	23.930	7.887	1.310
σ	0.464	0.566	0.807	0.551	1.682	0.562	0.097
σ_u	0.438	0.562	0.804	0.546	1.681	0.557	0.077
σ_v	0.152	0.071	0.071	0.071	0.070	0.071	0.059
θ	-	-	-	-	-	-	0.271
Estimated cost inefficiencies, $\hat{u}_{it}$							
Mean	0.350	0.813	0.663	0.679	1.399	0.692	0.068
SD	0.233	0.550	0.429	0.425	0.906	0.434	0.039
Min	0.060	0.000	0.015	0.020	0.032	0.020	0.018
Max	1.134	2.507	2.006	1.991	4.220	2.031	0.362
Log-likelihood	-116.572	-	595.159	596.523	597.649	597.285	577.898

Notes: Standard errors for ancillary parameters not reported.

Table 6: Swiss railways, correlation of inefficiency estimates

Variables	pcs	ss84	pl81	bc88	kumb90	bc92	tre
pcs	1.000						
ss84	0.439	1.000					
pl81	0.595	0.969	1.000				
bc88	0.608	0.971	0.991	1.000			
kumb90	0.573	0.984	0.991	0.998	1.000		
bc92	0.603	0.974	0.992	1.000	0.999	1.000	
tre	-0.029	-0.222	-0.249	-0.248	-0.243	-0.247	1.000

6.2 Spanish dairy farms

This application is based on a balanced panel of 247 dairy farms located in Northern Spain over a six years period (1993 – 1998). This dataset is interesting as it represents what is generally available to researchers: short panel, information only on input and output volumes, heterogeneity of output and less than ideal proxies for inputs. The output variable is given by the liters of milk produced per year. This measure explains only partially the final output of this industry, as milk can be also considered as an intermediate input to produce dairy products. Furthermore, variables like slaughtered animals should be also considered as part of the final output.

The functional form employed in the empirical analysis is the following *translog* production function with time dummy variables to control for neutral technical change

$$\begin{aligned} \ln y_{it} &= \beta_0 + \sum_{j=1}^{4} \beta_j \ln x_{jit} + \frac{1}{2} \sum_{j=1}^{4} \sum_{k=1}^{4} \beta_{jk} \ln x_{jit} \ln x_{kit} \\ &+ \sum_{t=1993}^{1998} \beta_t dyear_t - u_{it} + v_{it} \end{aligned} \tag{36}$$

where j and t are the subscripts denoting farm and year, respectively. Four inputs have been employed in the production frontier: number of milking cows (`x1`), number of man-equivalent units (`x2`), hectares of land devoted to pasture and crops (`x3`) and kilograms of feedstuffs fed to the dairy cows (`x4`). More details on these variables are available in Cuesta (2000) and Alvarez and Arias (2004).

We have estimated three models with time-varying inefficiency: the Normal-Half Normal Kumbhakar (1990) model (`kumb90`), a random effect model by means of the Feasible Generalized Least Squares (FGLS) method, the Cornwell et al. (1990) model (`css90`) estimated through the modified-LSDV technique and, finally, the Lee and Schmidt (1993) model (`ls93`) estimated using ILS. It is worth noting that the latter two models are estimated using approaches that do not allow intercept (β_0) and time dummies ($dyear_t$) to be simultaneously included into the frontier equation. Finally, we also considered two models with time-invariant inefficiency, i.e. the u_{it} term boils down to be $u_i$in equation (36): the first proposed by Schmidt and Sickles (1984) and estimated without any distributional assumption through the LSDV approach (`ss84`) and the second proposed by Pitt and Lee (1981) estimated through ML assuming a Half Normal inefficiency (`pl81`).

Table 7 reports the results of our exercise. There is a certain degree of similarity between the different models, as both parameters significance and magnitudes are comparable. Since for `ss84`, `css90` and `ls93` models the most efficient firm in the sample is considered as fully efficient, the smallest value of inefficiency is 0. On average and as expected, the `css90` model shows a higher level of inefficiency, whose distribution has also more variability while the other models seem to behave very similarly in this application. Finally, as we can see in table 8, linear correlations between inefficiencies are very high. This does not come as a surprise given the similarity of the estimated frontier parameters and it looks like an indication that in medium-short panels and in certain economic sectors/contexts, a time-invariant inefficiency specification is a valid solution.

Table 7: Spanish dairy farms, estimation results (247 firms for a total of 1482 observations)

	ss84	**css90**	**ls93**	**kumb90**	**pl81**
x1	0.642 ***	0.527 ***	0.641 ***	0.661 ***	0.660 ***
	(0.036)	(0.046)	(0.036)	(0.028)	(0.028)
x2	0.037 *	0.043 *	0.037 *	0.038 **	0.041 **
	(0.017)	(0.019)	(0.017)	(0.015)	(0.015)
x3	0.011	0.079	0.010	0.050 **	0.049 **
	(0.025)	(0.044)	(0.025)	(0.018)	(0.018)
x4	0.308 ***	0.226 ***	0.307 ***	0.351 ***	0.356 ***
	(0.020)	(0.024)	(0.020)	(0.018)	(0.017)
x11	0.135	-0.187	0.133	0.308	0.314
	(0.157)	(0.135)	(0.155)	(0.171)	(0.178)
x22	-0.002	0.060	-0.001	-0.111	-0.112
	(0.069)	(0.078)	(0.068)	(0.064)	(0.067)
x33	-0.242	-0.168	-0.243	-0.129	-0.131
	(0.188)	(0.223)	(0.187)	(0.119)	(0.115)
x44	0.105 *	-0.125 *	0.105 *	0.112 *	0.118 *
	(0.050)	(0.059)	(0.050)	(0.048)	(0.049)
x12	-0.010	0.059	-0.009	-0.060	-0.064
	(0.073)	(0.070)	(0.072)	(0.077)	(0.081)
x13	0.084	-0.114	0.085	0.088	0.091
	(0.102)	(0.111)	(0.101)	(0.090)	(0.090)
x14	-0.075	0.142	-0.074	-0.140	-0.146
	(0.083)	(0.093)	(0.082)	(0.084)	(0.088)
x23	0.001	0.067	0.002	0.020	0.011
	(0.050)	(0.076)	(0.050)	(0.049)	(0.050)
x24	-0.011	-0.062	-0.011	0.025	0.025
	(0.041)	(0.042)	(0.041)	(0.039)	(0.040)
x34	-0.012	0.110	-0.013	-0.015	-0.017
	(0.046)	(0.060)	(0.046)	(0.041)	(0.041)
dyear1994	0.035 ***			0.042 ***	0.027 ***
	(0.007)			(0.010)	(0.007)
dyear1995	0.062 ***			0.072 ***	0.048 ***
	(0.009)			(0.014)	(0.008)
dyear1996	0.072 ***			0.078 ***	0.052 ***
	(0.010)			(0.016)	(0.009)
dyear1997	0.075 ***			0.074 ***	0.051 ***
	(0.010)			(0.017)	(0.009)
dyear1998	0.092 ***			0.077 ***	0.064 ***
	(0.012)			(0.018)	(0.010)
Constant	11.512 ***			11.695 ***	11.711 ***
	(0.016)			(0.019)	(0.016)
t	-	-	-	-0.347	-
	-	-	-	(0.212)	-
t^2	-	-	-	0.045	-
	-	-	-	(0.028)	-
λ	1.948	4.807	2.010	4.485	2.775
σ	0.168	0.234	0.171	0.356	0.230
σ_u	0.149	0.229	0.153	0.348	0.216
σ_v	0.077	0.048	0.076	0.077	0.078
Estimated technical inefficiencies, $\hat{u}_{it}$					
Mean	0.315	0.685	0.353	0.288	0.179
SD	0.149	0.229	0.153	0.188	0.117
Min	0.000	0.000	0.000	0.014	0.009
Max	0.873	1.412	0.966	1.008	0.623
Log-likelihood	-	-	-	1355.248	1351.826

Notes: Cluster-robust standard errors in parenthesis. Standard errors for ancillary parameters are not reported.

Table 8: Spanish dairy farms, correlation of inefficiency estimates

Variables	ss84	css90	ls93	kumb90	pl81
ss84	1.000				
css90	0.861	1.000			
ls93	0.980	0.890	1.000		
kumb90	0.942	0.721	0.921	1.000	
pl81	0.931	0.704	0.910	0.999	1.000

7 Concluding remarks

In this article we introduce the new Stata commands `sfcross` and `sfpanel`, which implement an extensive array of SF models for cross-sectional and panel data. With respect to the available official Stata commands, `frontier` and `xtfrontier`, we add multiple features for estimating frontier parameters and for postestimating unit inefficiency/efficiency. In the development of the commands we widely exploit Mata potentiality. By using Mata structures, we provide a very readable code prone to be easily developed further by the Stata users community.

We illustrate the commands estimation capabilities through simulated data, focusing on some of the models that cannot be estimated using official Stata commands. Finally, we illustrate the proposed routines using real data sets under different possible empirical scenarios: short vs. long panels, cost vs. production frontiers, homogenous vs. heterogeneous outputs.

Acknowledgement

We are grateful to David Drukker and all participants at the 2009 Italian Stata Users Group meeting for useful comments. We thank William Greene for valuable advice and discussions and for maintaining an excellent webpage and making available several databases, two of which we have extracted and used in the empirical applications.

8 References

Ahn, S. C., Y. Hoon Lee, and P. Schmidt. 2001. GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics* 101(2): 219–255.

Ahn, S. C., L. Orea, and P. Schmidt. 2005. Estimation of a panel data model with parametric temporal variation in individual effects. *Journal of Econometrics* 126(2): 241–267.

Aigner, D., C. Lovell, and P. Schmidt. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6(1): 21–37.

Alvarez, A., and C. Arias. 2004. Technical efficiency and farm size: a conditional analysis. *Agricultural Economics* 30: 241–250.

Battese, G., and T. Coelli. 1988. Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. *Journal of Econometrics* 38: 387–399.

———. 1992. Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India. *Journal of Productivity Analysis* 3(1/2): 153–169.

———. 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics* 20: 325–332.

Belotti, F., and G. Ilardi. 2012. Consistent estimation of the “true” fixed-effects stochastic frontier model. *CEIS Research Papers* (231).

Caudill, S., and J. Ford. 1993. Biases in frontier estimation due to heteroscedasticity. *Economic Letters* 41: 17–20.

Caudill, S., J. Ford, and D. Gropper. 1995. Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. *Journal of Business Economics and Statistics* 13: 105–111.

Chen, Y., H. Wang, and P. Schmidt. 2011. Consistent estimation of the fixed effects stochastic frontier model. Mimeo.

Cornwell, C., P. Schmidt, and R. Sickles. 1990. Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics* 46: 185–200.

Cuesta, R. 2000. A production model with firm-specific temporal variation in technical inefficiency: With application to Spanish dairy farms. *Journal of Productivity Analysis* 13: 139–158.

Drukker, D. M., and R. Gates. 2006. Generating Halton sequences using Mata. *Stata Journal* 6(2): 214–228. http://www.stata-journal.com/article.html?article=st0103.

Farsi, M., M. Filippini, and W. Greene. 2005. Efficiency measurement in network industries: Application to the Swiss railway companies. *Journal of Regulatory Economics* 28:1: 69–90.

Greene, W. 1980a. Maximum likelihood estimation of econometric frontier functions. *Journal of Econometrics* 13: 27–56.

———. 1980b. On the estimation of a flexible frontier production model. *Journal of Econometrics* 13: 101–115.

———. 2003. Simulated Likelihood Estimation of the Normal-Gamma Stochastic Frontier Function. *Journal of Productivity Analysis* 19: 179–190.

———. 2005a. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics* 126: 269–303.

———. 2005b. Fixed and random effects in stochastic frontier models. *Journal of Productivity Analysis* 23: 7–32.

———. 2008. *The Measurement of Efficiency*, chap. The Econometric Approach to Efficiency Analysis. Oxford University Press.

Hadri, K. 1999. Estimation of a Doubly Heteroscedastic Stochastic Frontier Cost Function. *Journal of Business Economics and Statistics* 17(3): 359–363.

Horrace, W., and P. Schmidt. 1996. Confidence staements for efficiency estimates from Stochastic Frontier Models. *Journal of Productivity Analysis* 7: 257–282.

Huang, C. J., and J. T. Liu. 1994. Estimation of a Non-Neutral Stochastic Frontier Production Function. *The Journal of Productivity Analysis* 5: 171–180.

Jondrow, J., C. Lovell, I. Materov, and P. Schmidt. 1982. On the estimation of technical efficiency in the stochastic production function model. *Journal of Econometrics* 19: 233–238.

Kumbhakar, S. 1990. Production frontiers, panel data and time-varying technical inefficiency. *Journal of Econometrics* 46: 201–212.

Kumbhakar, S., and C. Lovell. 2000. *Stochastic frontier analysis*. Cambridge University Press.

Kumbhakar, S. C., S. Ghosh, and J. T. McGuckin. 1991. A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in U.S. Dairy Farms. *Journal of Business & Economic Statistics* 9: 279–286.

Lancaster, T. 2002. The incidental parameters problem since 1948. *Journal of Econometrics* 95: 391–414.

Lee, Y., and P. Schmidt. 1993. *The measurement of productive efficiency: techniques and applications*, chap. A production frontier model with flexible temporal variation in technical inefficiency. Oxford University Press.

Meeusen, W., and J. van den Broeck. 1977. Efficiency estimation from Cobb-Douglas production function with composed errors. *International Economic Review* 18(2): 435–444.

Neyman, J., and E. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16: 1–32.

Pitt, M., and L. Lee. 1981. The measurement and sources of technical inefficiency in the Indonesian weaving industry. *Journal of Development Economics* 9: 43–64.

Ritter, C., and L. Simar. 1997. Pitfalls of Normal-Gamma stochastic frontier models. *Journal of Productivity Analysis* 8: 167–182.

Schmidt, P., and R. Sickles. 1984. Production frontiers and panel data. *Journal of Business Economics and Statistics* 2(4): 367–374.

Stevenson, R. 1980. Likelihood functions for generalized stochastic frontier functions. *Journal of Econometrics* 13: 57–66.

Wang, H. 2002. Heteroscedasticity and non-monotonicity efficiency effects of a stochastic frontier model. *Journal of Productivity Analysis* 18: 241–253.

Wang, H., and P. Schmidt. 2002. One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis* 18(2): 289–296.

About the author

Federico Belotti is a research fellow at the University of Rome Tor Vergata.

Silvio Daidone is a research fellow at the Centre for Health Economics (CHE) of the University of York.

Giuseppe Ilardi is researcher at the Economic and Financial Statistics Department of the Bank of Italy.

Vincenzo Atella is an associate professor at the University of Rome Tor Vergata.