The Stata Journal (2013)
13, Number 4, pp. 719–758

Stochastic frontier analysis using Stata

Federico Belotti
Centre for Economic and International Studies
University of Rome Tor Vergata
Rome, Italy
federico.belotti@uniroma2.it

Silvio Daidone
Centre for Health Economics
University of York
York, UK
silvio.daidone@york.ac.uk

Giuseppe Ilardi
Economic and Financial Statistics Department
Bank of Italy
Rome, Italy
giuseppe.ilardi@bancaditalia.it

Vincenzo Atella
Centre for Economic and International Studies
University of Rome Tor Vergata
Rome, Italy
atella@uniroma2.it

Abstract. This article describes `sfcross` and `sfpanel`, two new Stata commands for the estimation of cross-sectional and panel-data stochastic frontier models. `sfcross` extends the capabilities of the `frontier` command by including additional models (Greene, 2003, *Journal of Productivity Analysis* 19: 179–190; Wang, 2002, *Journal of Productivity Analysis* 18: 241–253) and command functionalities, such as the possibility of managing complex survey data characteristics. Similarly, `sfpanel` allows one to fit a much wider range of time-varying inefficiency models compared with the `xtfrontier` command, including the model of Cornwell, Schmidt, and Sickles (1990, *Journal of Econometrics* 46: 185–200); the model of Lee and Schmidt (1993, in *The Measurement of Productive Efficiency: Techniques and Applications*), a production frontier model with flexible temporal variation in technical efficiency; the flexible model of Kumbhakar (1990, *Journal of Econometrics* 46: 201–211); the inefficiency effects model of Battese and Coelli (1995 *Empirical Economics* 20: 325–332); and the "true" fixed- and random-effects models of Greene (2005a, *Journal of Econometrics* 126: 269–303). A brief overview of the stochastic frontier literature, a description of the two commands and their options, and examples using simulated and real data are provided.

Keywords: st0315, sfcross, sfpanel, stochastic frontier analysis, production frontier, cost frontier, cross-sectional, panel data

1 Introduction

This article describes `sfcross` and `sfpanel`, two new Stata commands for the estimation of parametric stochastic frontier (SF) models using cross-sectional and panel data.

Since the publication of the seminal articles by Meeusen and van den Broeck (1977) and Aigner, Lovell, and Schmidt (1977), this class of models has become a popular tool for efficiency analysis; a 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 (2012).

The SF model is motivated by the theoretical idea that no economic agent can exceed the ideal "frontier", and deviations from this extreme represent 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 with a one-sided disturbance that 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 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 mirrors the `frontier` command's functionality but adds 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 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 the "time decay" model (Battese and Coelli 1992). Our `sfpanel` command allows one to fit a wider range of time-varying inefficiency models, including the model of Cornwell, Schmidt, and Sickles (1990), that of Lee and Schmidt (1993), the flexible model of Kumbhakar (1990), the time decay and the inefficiency effects models of Battese and Coelli (1992, 1995), and the "true" fixed- (TFE) and random-effects (TRE) models of 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 model of Schmidt and Sickles (1984), which is characterized by no distributional assumptions on the inefficiency term. In addition, because the main objective

1. The literature distinguishes between production and cost frontiers. The former represents 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.

of SF analysis is the estimation of inefficiency, we provide postestimation 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` also allow the simultaneous modeling of heteroskedasticity 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 that can be easily developed further by other Stata users. All of these features make the commands simple to use, extremely flexible, and fast; they also ensure the opportunity to fit state-of-the-art SF models.

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

The article 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 summary of all estimable models and their features is reported in table 1.

2.1 Cross-sectional models

Consider the following SF model,

$$y_i = \alpha + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, N \tag{1}$$

$$\varepsilon_i = v_i - u_i \tag{2}$$

$$v_i \sim \mathcal{N}(0, \sigma_v^2) \tag{3}$$

$$u_i \sim \mathcal{F} \tag{4}$$

where y_i represents the logarithm of the output (or cost) of the ith productive unit, $\mathbf{x}_i$ 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-sided disturbance, u_i, representing inefficiency.[2] Moreover, u_i and v_i are assumed to be independent of each other and independent and identically distributed across observations. The last assumption about the distribution $\mathcal{F}$ of the inefficiency term is needed to make the model estimable. Aigner, Lovell, and Schmidt (1977) assumed a half-normal distribution, that is, $u_i \sim \mathcal{N}^{+}\left(0, \sigma_u^2\right)$, 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 fit by maximum likelihood (ML), even if modified ordinary least-squares or generalized method of moments estimators are possible (often inefficient) alternatives.[3] In general, SF analysis is based on two sequential steps: in the first, estimates of the model parameters $\widehat{\boldsymbol{\theta}}$ 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 $f(u_i|\widehat{\varepsilon}_i)$, where $\widehat{\varepsilon}_i = y_i - \widehat{\alpha} - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$.

The derivation of the likelihood function is based on the independence assumption between u_i and v_i. Because the composite model error ε_i is defined as $\varepsilon_i = v_i - u_i$, its probability density function 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

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

The marginalization of u_i in (5) leads to a convenient closed-form expression only for the normal-half normal, normal-exponential, and normal-truncated normal models. In all other cases (for example, the normal-gamma model), numerical- or simulation-based techniques are necessary to approximate the integral in (5).

The second estimation step is necessary because the estimates of the model parameters allow for the computation of residuals $\widehat{\boldsymbol{\varepsilon}}$ but not for the inefficiency estimates.

2. In this section, we consider only production functions. However, the sign of the u_i term in (2) is positive or negative depending on whether the frontier describes a cost or a production function, respectively.
3. Notice that when a distributional assumption on $\mathbf{u}$ is made, `sfcross` and `sfpanel` estimate model parameters by likelihood-based techniques.
4. Different model parameterizations 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 $\lambda = \sigma_u/\sigma_v$.

Because 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 $\mathbf{u}$ given $\boldsymbol{\varepsilon}$. Thus a point estimate of the inefficiencies can be obtained using the mean $\mathbb{E}(\mathbf{u}|\widehat{\boldsymbol{\varepsilon}})$ (or the mode $\mathbb{M}[\mathbf{u}|\widehat{\boldsymbol{\varepsilon}}]$) of this conditional distribution. Once point estimates of $\mathbf{u}$ are obtained, estimates of the technical (cost) efficiency can be derived as

$$\text{Eff} = \exp\left(-\widehat{\mathbf{u}}\right)$$

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

2.2 Panel-data models

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

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

$$\begin{aligned} y_{it} &= \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T_i \\ \varepsilon_{it} &= v_{it} - u_i \\ v_{it} &\sim \mathcal{N}(0, \sigma_v^2) \\ u_i &\sim \mathcal{N}^{+}\left(0, \sigma_u^2\right) \end{aligned} \tag{6}$$

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 an 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 the presence of empirical applications based on long-panel datasets. To relax this restriction, Cornwell, Schmidt, and Sickles (1990) have approached the problem by proposing the following SF model with individual-specific slope parameters,

$$\begin{aligned} y_{it} &= \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it} \pm u_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T_i \\ u_{it} &= \omega_i + \omega_{i1} t + \omega_{i2} t^2 \end{aligned} \tag{7}$$

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

5. A general presentation of the postestimation procedures implemented in the `sfcross` and `sfpanel` routines is given by Kumbhakar and Lovell (2000) and Greene (2012), to whom we refer the reader for further details.
6. The normal-exponential model is another straightforward extension allowed by `sfpanel`.

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

$$u_{it} = g(t) \times u_i$$

where $g(t)$ is represented by a set of time dummy variables. This specification is more parsimonious than (7), and it does not impose any parametric form, but it is less flexible because 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

$$g(t) = \left\{1 + \exp\left(\gamma t + \delta t^2\right)\right\}^{-1}$$

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 $\gamma = \delta = 0$. A similar model called "time decay" has been proposed by Battese and Coelli (1992) in which

$$g(t) = \exp\left\{-\eta\left(t - T_i\right)\right\}$$

The common feature of all of these time-varying SF models is that the intercept α is the same across productive units, thus generating a misspecification bias in the presence of time-invariant unobservable factors, which are unrelated with the production process but affect the output. Therefore, 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 (6) by the following specification:

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \tag{8}$$

Compared with previous models, this specification allows one to disentangle time-varying inefficiency from unit-specific time-invariant unobserved heterogeneity. For this reason, Greene (2005a) termed these models TFE or TRE according to the assumptions on the unobserved unit-specific heterogeneity. While the estimation of the TRE specification can be easily performed using simulation-based techniques, the ML estimation of the TFE variant requires the solution of two major issues related to the estimation of nonlinear panel-data models. The first is purely computational because of the large dimension of the parameter space. Nevertheless, Greene (2005a,b) showed that a maximum-likelihood dummy variable (MLDV) approach is computationally feasible also in the 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

7. Han, Orea, and Schmidt (2005) and Ahn, Hoon Lee, and Schmidt (2001) propose estimating the models of Cornwell, Schmidt, and Sickles (1990) and Lee and Schmidt (1993), respectively, through a generalized method of moments (GMM) approach. They show that GMM is preferable because it is asymptotically efficient. Currently, `sfpanel` allows for the estimation of the models of Cornwell, Schmidt, and Sickles (1990) and Lee and Schmidt (1993) through modified least-squares dummy variables and iterative least-squares approaches, respectively. We leave for future updates the implementation of the GMM estimator.

when the number of units is relatively large compared with the length of the panel. In these cases, the unit-specific intercepts are inconsistently estimated as $N \to \infty$ with fixed T because 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), because 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 (8) may appear to be the most flexible and parsimonious choice among the several existing time-varying specifications, one can argue that a portion of the time-invariant unobserved heterogeneity does belong to inefficiency, or these two components should not be disentangled at all. The `sfpanel` command provides options for the estimation of these two extremes: for the models of Schmidt and Sickles (1984), Pitt and Lee (1981), and Battese and Coelli (1988), in which all time-invariant unobserved heterogeneity is considered as inefficiency, and for 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 completely satisfactory a priori, 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 $\mathbf{u}$ and $\mathbf{v}$ 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 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 nontrivial numerical maximization problems.

2.3 Exogenous inefficiency determinants and heteroskedasticity

A very important issue in SF analysis is the inclusion in the model of exogenous variables that 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

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 Greene's (2005a) model is still an open research issue in SF literature. Promising solutions have been proposed by Chen, Schmidt, and Wang (2011) for a homoskedastic normal-half normal model and Belotti and Ilardi (2012) for a more flexible heteroskedastic specification in normal-half normal and normal-exponential models. We are currently working to update the `sfpanel` command along these directions.

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.

affect the productive unit performance, have been incorporated in a variety of ways: i) they may shift the frontier function and the inefficiency distribution; ii) they may scale the frontier function and the inefficiency distribution; and iii) they may shift and scale the frontier function and the inefficiency distribution. Moreover, Kumbhakar and Lovell (2000) stress that in contrast with the linear regression model, in which the misspecification of the second moment of the errors distribution determines only efficiency losses, the presence of uncontrolled observable heterogeneity in u_i and v_i may affect the inference in SF models. Indeed, while neglected heteroskedasticity in v_i does not produce any bias for the frontier's parameter 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 heteroskedasticity of the inefficiency as well as in the parameter of the idiosyncratic error term for the models implemented in the `sfcross` and `sfpanel` commands. Because 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 (2012), researchers have often incorporated exogenous effects using a two-step 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 that this approach leads to severely biased results; thus we shall focus only on model extensions based on simultaneous estimation.

A natural starting point for introducing exogenous variables in the inefficiency model is in the location of the distribution. The most well-known approaches are those suggested by Kumbhakar, Ghosh, and McGuckin (1991) and Huang and Liu (1994). They proposed to parameterize the mean of the pretruncated inefficiency distribution. Basically, model (1)–(3) can be completed with

$$\begin{aligned} u_i &\sim \mathcal{N}^{+}\left(\mu_i, \sigma_u^2\right) \\ \mu_i &= \mathbf{z}_i'\boldsymbol{\psi} \end{aligned} \tag{9}$$

where u_i is a realization from a truncated normal random variable, $\mathbf{z}_i$ 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 $\mathbf{z}_i$ may include interactions with input variables, thus allowing one to test the hypothesis that inefficiency is neutral with respect to its impact on input usage.[10]

10. Battese and Coelli (1995) proposed a similar specification for panel data.

An alternative approach to analyzing the effect of exogenous determinants on inefficiency is obtained by scaling its distribution. Then a model that allows heteroskedasticity in u_i and v_i becomes a straightforward extension. For example, Caudill and Ford (1993), Caudill, Ford, and Gropper (1995), and Hadri (1999) proposed to parameterize the variance of the pretruncated inefficiency distribution in the following way:

$$u_i \sim \mathcal{N}^+\left(0, \sigma_{ui}^2\right) \tag{10}$$

$$\sigma_{ui}^2 = \exp\left(\mathbf{z}_i'\boldsymbol{\psi}\right) \tag{11}$$

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

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

$$\sigma_{vi}^2 = \exp\left(\mathbf{h}_i'\boldsymbol{\phi}\right) \tag{12}$$

where the variables in $\mathbf{h}_i$ do not necessarily appear in $\mathbf{z}_i$.

As in Wang (2002), both `sfcross` and `sfpanel` allow one to combine (9) and (10) for the normal-truncated normal model. In postestimation, it is possible to compute nonmonotonic effects of the exogenous factors $\mathbf{z}_i$ 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 $h(\mathbf{z}_i, \boldsymbol{\psi})$. Their model, $u_i = h(\mathbf{z}_i, \boldsymbol{\psi})u_i^*$ with $u_i^* \sim \mathcal{N}\left(\mu, \sigma^2\right)^+$, is equivalent to the assumption that $u_i \sim \mathcal{N}\{\mu h(\mathbf{z}_i, \boldsymbol{\psi}), \sigma^2 h(\mathbf{z}_i, \boldsymbol{\psi})^2\}^+$, in which the $\mathbf{z}_i$ vector does not include a constant term.[11]

11. We are currently working to extend the `sfcross` command to normal-truncated normal models with scaling property (Wang and Schmidt 2002).

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

Reference	**Arguments of `d()` and `m()`**[1]	**F** [2]	**Est. method**[3]	**Loc. eff. in u**	**Heter. in u**	**Heter. in v**	**JLMS**[4]	**BC**[5]	**CI**[6]
			Cross-sectional models						
Aigner, Lovell, and Schmidt (1977)	`hnormal`	HN	ML		x	x	x	x	x
Meeusen and van den Broeck (1977)	`exponential`	E	ML	x	x	x	x	x	x
Stevenson (1980)	`tnormal`	TN	ML	x	x	x	x	x	x
Greene (2003)	`gamma`	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)	`pl81`	HN	ML				x	x	x
Battese and Coelli (1988)	`bc88`	TN	ML				x	x	x
Cornwell, Schmidt, and Sickles (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] The `model()` option is available in `sfpanel`. The `distribution()` option is available in `sfcross`.

[2] Distribution $\mathcal{F}$ of $\mathbf{u}$: HN = half normal, E = exponential, TN = truncated normal, G = gamma.

[3] 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.

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

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

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

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 command is as follows:

```
sfcross depvar [indepvars] [if] [in] [weight] [, distribution(distname)
    emean(varlist_m [, noconstant]) usigma(varlist_u [, noconstant])
    vsigma(varlist_v [, noconstant]) svfrontier(listname) svemean(listname)
    svusigma(listname) svvsigma(listname) cost simtype(simtype)
    nsimulations(#) base(#) postscore posthessian]
```

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`). Stata 11.2 is the earliest version that can be used to run the commands. `aweight`, `fweight`, `iweight`, 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 `distribution(exponential)`.

`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 heteroskedastic, 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 heteroskedastic, with the variance expressed as a function of the covariates defined in *varlist_v*. Specifying `noconstant` suppresses the constant in this function.

`svfrontier(`*listname*`)` specifies a $1 \times k$ vector of initial values for the coefficients of the frontier. The vector must have the same length of the parameter vector to be estimated.

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

`svusigma(`*listname*`)` specifies a $1 \times k_u$ vector of initial values for the coefficients of the technical inefficiency variance function.

`svvsigma(`*listname*`)` specifies a $1 \times k_v$ vector of initial values for the coefficients of the idiosyncratic error variance function.

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

`simtype(`*simtype*`)` specifies the method for random draws when `distribution(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(#)`. The default is `simtype(runiform)`. See `help mata halton()` for more details on generating Halton sequences.

`nsimulations(#)` specifies the number of draws used when `distribution(gamma)` is specified. The default is `nsimulations(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 `base(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. Scores are defined as the derivative of the objective function with respect to the parameters. This option is not allowed when the size of the scores' matrix is greater than the Stata matrix limit; see [R] **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 one to compute the (in)efficiency confidence interval through the option `ci` as well as nonmonotonic marginal effects in the manner of Wang (2002) using, when appropriate, the option `marginal`. The syntax of the command is the following,

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

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

where *statistic* 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 (technical or cost) inefficiency via $\mathbb{E}(u|\varepsilon)$ using the estimator of Jondrow et al. (1982).

`m` produces estimates of (technical or cost) inefficiency via $\mathbb{M}(u|\varepsilon)$, 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 (technical or cost) efficiency via $\exp\{-\mathbb{E}(u|\varepsilon)\}$.

`bc` produces estimates of (technical or cost) efficiency via $\mathbb{E}\{\exp(-u|\varepsilon)\}$, the estimator of Battese and Coelli (1988). This option is not allowed when the estimation is performed with the `distribution(gamma)` option.

`ci` computes the confidence interval using the approach proposed by Horrace and Schmidt (1996). It can be used only when `u` or `bc` is specified. The default is `level(95)`, or 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 determinant specified in options `emean(`*varlist_m* [`, noconstant`]`)` and `usigma(`*varlist_u* [`, noconstant`]`)`. `marginal` can be used only if 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 `u`, `m`, `jlms`, or `bc`; and ii) alone without specifying *newvar*.

`scores` calculates score variables. When the argument of the option `distribution()` is `hnormal`, `tnormal`, or `exponential`, score 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 for the estimation of SF panel-data models through ML and least-squares techniques. The general `sfpanel` syntax is the following:

`sfpanel` *depvar* [*indepvars*] [*if*] [*in*] [*weight*], `model(`*modeltype*`)` [*options*]

As for its cross-sectional counterpart, version 11.2 is the earliest version of Stata that can be used to run `sfpanel`. Similarly, all types 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,b)

`distribution(`*distname*`)` specifies the distribution for the inefficiency term as half-normal (`hnormal`), truncated normal (`tnormal`), or exponential (`exponential`). The default is `distribution(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 heteroskedastic, 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 heteroskedastic, 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(`#`)`. The default is `simtype(runiform)`. See `help mata halton()` for more details on generating Halton sequences. Only for `model(tre)`.

`nsimulations(`#`)` specifies the number of draws used in the simulation. The default is `nsimulations(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 `base(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 heteroskedastic, 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 heteroskedastic, 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 $u_{it} = u_i \left[1 + \exp(\textit{varlist\_bt})\right]^{-1}$. 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 one to compute the (in)efficiency confidence interval through the option `ci` as well as nonmonotonic marginal effects in the manner of Wang (2002) using, when appropriate, the option `marginal`. The syntax of the command is the following,

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

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

where *statistic* includes `xb`, `stdp`, `u`, `u0`, `m`, `jlms`, `bc`, `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 (technical or cost) inefficiency via $\mathbb{E}(u|\varepsilon)$ using the estimator of Jondrow et al. (1982).

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

`m` produces estimates of (technical or cost) inefficiency via $\mathbb{M}(u|\varepsilon)$, 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 (technical or cost) efficiency via $\exp\{-\mathbb{E}(u|\varepsilon)\}$.

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

`ci` computes the confidence interval using the approach of Horrace and Schmidt (1996). This option can only be used with the `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 is `level(95)`, or 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 $\text{Var}(u)$. 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 the variance of inefficiency, respectively. *varname_m* and *varname_u* are the names of each exogenous determinant specified in options `emean(`*varlist_m* [`, noconstant`]`)` and `usigma(`*varlist_u* [`, noconstant`]`)`. `marginal` can only be used if estimation is performed with the `model(bc95)` option or if 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 `u`, `m`, `jlms`, or `bc`; and 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 only be used if the estimation is performed with `model(fe)`, `model(regls)`, `model(fecss)`, and `model(fels)`.

`scores` 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`, score 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]

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 fitting a normal-gamma SF production model. We consider the following data-generating process (DGP),

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

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 0 means and variances equal to 1 and 4, respectively. The sample size is set to 1,000 observations, a large size as noted by Ritter and Simar (1997) but in general not so large given the current availability of microdata. Let us begin by fitting the normal-exponential model using the following syntax:

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

```
. 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
```

Note that the normal-exponential model is the `sfcross` default, so we might omit the option `distribution(exponential)`.[13] As can be seen, although there is only one equation to be estimated in the model, the command fits three of Mata's `moptimize()` equations (see [M-5] **moptimize( )**). Indeed, given that `sfcross` allows both the inefficiency and the idiosyncratic error to be heteroskedastic (see table 1), the output also reports variance parameters estimated in a transformed metric according to (11) and (12), respectively. In this example, the inefficiency is assumed to be homoskedastic, so `sfcross` estimates the coefficient of the constant term in (11) rather than directly estimating σ_u. 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, inefficiency estimates are obtained through the approach of Jondrow et al. (1982). Because the inefficiencies are drawn from a gamma distribution, a better fit can be obtained using the following command:

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

```
. 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)  =    438.02
                                                   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	.3809769	.0670487	5.68	0.000	.2495639	.5123899
x2	.6877634	.0336088	20.46	0.000	.6218914	.7536354
_cons	.9361791	.4121864	2.27	0.023	.1283087	1.74405
Usigma						
_cons	1.53519	.226486	6.78	0.000	1.091286	1.979094
Vsigma						
_cons	-.2734356	.333033	-0.82	0.412	-.9261682	.379297
sigma_u	2.154578	.2439909	8.83	0.000	1.725717	2.690016
sigma_v	.8722163	.1452384	6.01	0.000	.6293397	1.208825
lambda	2.470234	.1969744	12.54	0.000	2.084171	2.856297
g_shape	1.879186	.3845502	4.89	0.000	1.125482	2.632891

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

In the normal-gamma cross-sectional model, the parameters are estimated using simulated maximum likelihood (SML). 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 has a more uniform coverage than a sequence generated from pseudouniform random numbers. Moreover, as noted by Greene (2003), the computational efficiency, when compared with that of pseudouniform random draws, appears to be at least 10 to 1. Thus, in our example, the same results can be approximately obtained using 500 pseudouniform draws (see `help mata halton()`).[14]

14. For all models fit using SML, the default option of `sfcross` and `sfpanel` is `simtype(uniform)` with `nsim(250)`. In our opinion, small values (for example, 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 for more precise results. We suggest using 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".

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
```

5.2 Panel-data time-varying inefficiency models

Cornwell, Schmidt, and Sickles (1990) and Lee and Schmidt (1993) provide a fixed-effects treatment of models like those proposed by Kumbhakar (1990) and Battese and Coelli (1992). Currently, `sfpanel` allows for the estimation of the models of Cornwell, Schmidt, and Sickles (1990) and Lee and Schmidt (1993) by means of modified least-squares dummy variables and iterative least squares (ILS), respectively. An interesting aspect of these models is that although they have been proposed in the SF literature, they are actually linear panel-data models with time-varying fixed effects and thus potentially very useful in other contexts. However, their consistency requires white noise errors, and they are less efficient than the GMM estimator proposed by Ahn, Hoon Lee, and Schmidt (2001) and Han, Orea, and Schmidt (2005).

15. Notice that `exp` and `gamma` are the names of the exponential and gamma models' estimation results saved with 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 sample is needed to achieve a reasonable degree of precision in the estimates of inefficiency distribution parameters.

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

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

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 $\mathbf{u}_i = \delta_i \boldsymbol{\xi}$. 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 setup 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 models of Lee and Schmidt (1993) and Cornwell, Schmidt, and Sickles (1990) 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 postestimate 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 fit using `sfpanel` and the official `xtreg` command. However, when the estimation is performed using `sfpanel`, the `predict` command with the option `u` can be used to obtain inefficiency estimates.[17]

```
. 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 (1993), the

17. Both `xtreg` and `sfpanel` also allow for the estimation of the random-effects version of this model through the feasible generalized least-squares approach.

estimator by Cornwell, Schmidt, and Sickles (1990) 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)	(0.0237)	(0.0230)
`x2`	0.626***	0.619***	0.611***
	(0.0354)	(0.0121)	(0.0117)
`x3`	0.602***	0.591***	0.596***
	(0.0220)	(0.0073)	(0.0075)
`x1_sq`	0.193***	0.204***	0.209***
	(0.0234)	(0.0078)	(0.0076)
`x2_sq`	0.099***	0.103***	0.101***
	(0.0080)	(0.0027)	(0.0026)
`x3_sq`	0.198***	0.201***	0.201***
	(0.0036)	(0.0012)	(0.0012)
`x1_x2`	0.149***	0.142***	0.145***
	(0.0198)	(0.0066)	(0.0064)
`x1_x3`	−0.293***	−0.295***	−0.295***
	(0.0130)	(0.0043)	(0.0043)
`x2_x3`	−0.306***	−0.300***	−0.301***
	(0.0076)	(0.0026)	(0.0025)
`_cons`	−0.050		
	(0.0866)		
σ_u	0.223	1.859	1.832
σ_v	2.096	0.499	0.497

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

18. This assumption involves calculating $\widehat{u}_i = \widehat{\alpha} - \widehat{\alpha}_i$ with $\widehat{\alpha} = \max_{i=1,\ldots,n}(\widehat{\alpha}_i)$ in the case of time-invariant inefficiency models and $\widehat{u}_{it} = \widehat{\alpha}_t - \widehat{\alpha}_{it}$ with $\widehat{\alpha}_t = \max_{i=1,\ldots,n}(\widehat{\alpha}_{it})$ in the case of time-varying inefficiency models.

```
. summarize u uhat_fels uhat_fecss uhat_fess
```

Variable	Obs	Mean	Std. Dev.	Min	Max
u	2500	1.345894	1.265155	0	4.486315
uhat_fels	2500	1.669869	1.419268	0	5.550043
uhat_fecss	2500	2.214224	1.314245	0	6.500714
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	0.9794	1.0000		
uhat_fecss	0.8974	0.9129	1.0000	
uhat_fess	0.0005	0.0092	0.1991	1.0000

Finally, we show additional features of `sfpanel`: i) the possibility of computing elasticities via the official `lincom` command; and ii) the possibility of performing 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, 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:

```
. lincom x1 + x1_sq * x1m + x1_x2*x2m + x1_x3*x3m
 ( 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]
(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. To run a constrained fixed-effects estimation, we can define the required set of constraints to impose CRS 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 `model(fe)` and `constraints(1 2 3 4)` options.

```
. 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]
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					

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 because 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 one to disentangle time-invariant heterogeneity from time-varying inefficiency. In this section, we present the main syntax and some of the options to fit such models. We start by specifying the following normal-exponential stochastic production frontier model,

$$y_{it} = 1 + \alpha_i + 0.3x_{1it} + 0.7x_{2it} + v_{it} - u_{it}$$
$$v_{it} \sim \mathcal{N}(0,1) \tag{13}$$
$$u_{it} \sim \mathcal{E}(2), \quad i = 1,\ldots,n, \quad t = 1,\ldots,T \tag{14}$$

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]

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

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 0 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 as a panel, we fit the following models:

i) a normal-exponential TFE model on TFE_{DGP} data (`tfe1`)[20]

```
. sfpanel yf x1_c x2_c, model(tfe) distribution(exp) rescale
  (output omitted)
. estimates 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)
. estimates 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(100)
> simtype(genhalton) base(7) rescale
  (output omitted)
. estimates 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 SML technique assuming that the unobserved heterogeneity is distributed as $\mathcal{N}(0, \theta^2)$ (where θ represents the standard deviation of the unobserved heterogeneity) and that $\mathbb{E}(\alpha_i|x_{1it}, x_{2it}) = 0$. Thus, because the estimates are obtained using the TFE_{DGP} data, the frontier and θ parameter estimates are biased.

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

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

After each estimation, we use the `predict` command to obtain inefficiency estimates. As already mentioned, option `u` instructs the postestimation routine to compute inefficiencies through the estimator of Jondrow et al. (1982) (see `help sfpanel postestimation`). In the case of the TRE model, the `predict` command also allows for 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	.1338385	19.47738

```
. 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 estimates of Jondrow et al. (1982) 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 due to the incidental parameters problem. Also note the good performances of the TRE model when it is fit on the TFE_{DGP} data (`u_tre_c`).

Introducing heteroskedasticity

Finally, we deal with the problem of heteroskedasticity, a very important issue for applied research. For both TFE and TRE models, we compare the estimates obtained from a model that neglects heteroskedasticity with those obtained from a heteroskedastic one. To introduce heteroskedasticity, we replace equations (13)–(14) with the following,

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

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. Note that because of the introduction of heteroskedasticity, 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, which implies an average of about 1.9. We estimate four different models:

i) a homoskedastic TFE model on heteroskedastic 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 heteroskedastic TFE model on heteroskedastic 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 homoskedastic TRE model on heteroskedastic 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
```

iv) a heteroskedastic TRE model on heteroskedastic 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 estimates are biased by both the incidental parameters problem and the neglected heteroskedasticity in **u** and **v**. These estimates can be significantly improved by considering both sources of heteroskedasticity using the options `usigma(`*varlist_u*`)` and `vsigma(`*varlist_v*`)` (`tfe2`). Exactly the same argument applies in the TRE case (`tre1` versus `tre2`) but without the incidental parameters problem.

Table 4. TFE and TRE estimation results (homoskedasticity versus heteroskedasticity)

	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

As we mentioned in section 2.3, neglecting heteroskedasticity in **u** and **v** leads to biased inefficiency estimates. This conclusion is confirmed by the `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
```

The average inefficiency is upward biased (by about 15%) for both TFE and TRE models in which heteroskedasticity has been neglected. A slightly better result is also obtained in terms of Spearman's 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 the cost inefficiency of Swiss railways using data from the Swiss Federal Office of Statistics on public transport companies; the second focuses on the technical inefficiency of Spanish dairy farms, using data from a voluntary record-keeping program.[22]

6.1 Swiss railways

This application is based on an unbalanced panel of 50 railway companies over the period 1985–1997, which resulted in 605 observations. We think that this application is interesting for at least two reasons: i) cost frontiers are much less diffuse in the literature compared with production frontiers, given the lack of reliable cost and price data; and ii) the length of the panel makes this database quite unusual in the SF literature. A detailed description of the Swiss railway transport system and complete information on the variables used are available in Farsi, Filippini, and Greene (2005).

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

$$\ln\left(\frac{TC_{it}}{Pe_{it}}\right) = \beta_0 + \beta_Y \ln Y_{it} + \beta_Q \ln Q_{it} + \beta_N \ln N_{it} + \\ + \beta_{Pk} \ln\left(\frac{Pk_{it}}{Pe_{it}}\right) + \beta_{Pl} \ln\left(\frac{Pl_{it}}{Pe_{it}}\right) + \sum_{t=1986}^{1997} \beta_t \text{dyear}_t + u_{it} + v_{it} \quad (15)$$

where i and t are the subscripts denoting the railway company and year, respectively. As is 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

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

is included as an 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 also included a set of time dummies, dyear_t, to control for unobserved time-dependent variation in costs.

We consider three time-varying inefficiency specifications—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`) model, and the Battese and Coelli (1988) (`bc88`) model. All models are fit assuming that the inefficiency is half-normally distributed—that is, all except `bc88` and `bc92`, in which $u \sim \mathcal{N}^{+}(\mu, \sigma_u^2)$, and `ss84`, in which no distributional assumption is made. The choice of also including Greene's specification is driven by the multioutput 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 fit a pooled cross-sectional model (`pcs`).

Table 5 shows the results. Coefficient estimates of input prices and outputs are all significant across the seven models and have 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 is 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, because the two time-invariance hypotheses, $H_0 : t = t^2 = 0$ in the `kumb90` model and $H_0 : \eta = 0$ in the `bc92` specification, cannot be rejected at the 5% level. Hence, we may conclude that there is evidence of time-invariant technical inefficiency in the Swiss railway transport system, at least for the study period.

Consistently with this result, we also find that the `tre` model provides inefficiency estimates that have no link with those obtained from any of the other models. Moreover, because of a very low estimate of the inefficiency variance, the estimated signal-to-noise ratio, $\widehat{\lambda}$, 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 one to disentangle time-invariant unobserved heterogeneity from inefficiency. This interpretation is supported by the fact that the estimated standard deviation of the random effects (θ) dominates the inefficiency one (σ_u).

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***	0.201***
	(0.015)	(0.032)	(0.034)	(0.033)	(0.033)	(0.033)	(0.026)
lnQ	0.030***	0.014*	0.021***	0.021***	0.020***	0.020***	0.028***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)
lnN	0.393***	0.448***	0.485***	0.503***	0.477***	0.499***	0.583***
	(0.027)	(0.051)	(0.045)	(0.047)	(0.044)	(0.047)	(0.034)
lnpk	0.171***	0.318***	0.310***	0.311***	0.311***	0.313***	0.311***
	(0.032)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
lnpl	0.592***	0.546***	0.548***	0.546***	0.538***	0.543***	0.560***
	(0.074)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)	(0.037)
dyear1986	0.009	0.010	0.009	0.009	0.015	0.008	0.015
	(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	0.018
	(0.056)	(0.015)	(0.015)	(0.015)	(0.017)	(0.015)	(0.015)
dyear1988	0.010	0.039*	0.028	0.027	0.043*	0.023	0.034*
	(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**	0.058***
	(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***	0.074**
	(0.058)	(0.016)	(0.016)	(0.016)	(0.022)	(0.017)	(0.016)
dyear1991	0.030	0.098***	0.078***	0.078***	0.096***	0.069***	0.086***
	(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***	0.101***
	(0.058)	(0.017)	(0.017)	(0.017)	(0.023)	(0.019)	(0.017)
dyear1993	0.015	0.100***	0.081***	0.081***	0.092***	0.069***	0.089***
	(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*	0.070***
	(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	0.050**
	(0.057)	(0.016)	(0.016)	(0.016)	(0.022)	(0.021)	(0.016)
dyear1996	0.027	0.037*	0.028	0.027	0.018	0.010	0.017
	(0.057)	(0.017)	(0.016)	(0.016)	(0.022)	(0.022)	(0.017)
dyear1997	0.019	0.038*	0.030	0.029	0.009	0.009	0.029
	(0.060)	(0.018)	(0.017)	(0.017)	(0.023)	(0.024)	(0.017)
Constant	−8.310***	−2.682***	−4.895***	−4.929***	−4.626***	−4.871***	−4.894***
	(0.976)	(0.652)	(0.643)	(0.634)	(0.637)	(0.637)	(0.531)

Continued on next page

	pcs b/se	ss b/se	pl81 b/se	bc88 b/se	kumb90 b/se	bc92 b/se	tre b/se
t	-	-	-	-	0.023	-	-
	-	-	-	-	(0.015)	-	-
t^2	-	-	-	-	−0.002	-	-
	-	-	-	-	(0.001)	-	-
η	-	-	-	-	-	−0.002	-
	-	-	-	-	-	(0.002)	-
λ	2.882	7.803	11.366	7.716	23.930	7.887	1.634
σ	0.464	0.560	0.807	0.551	1.682	0.562	0.098
σ_u	0.438	0.555	0.804	0.546	1.681	0.557	0.083
σ_v	0.152	0.071	0.071	0.071	0.070	0.071	0.051
θ	-	-	-	-	-	-	0.347
Estimated cost inefficiencies, $\widehat{u}_{it}$							
Mean	0.350	0.807	0.663	0.679	0.687	0.682	0.091
SD	0.233	0.550	0.429	0.425	0.445	0.428	0.076
Min	0.060	0.000	0.015	0.020	0.015	0.019	0.018
Max	1.134	2.507	2.006	1.991	2.124	2.031	0.629
Log likelihood	−116.572	-	595.159	596.523	597.649	597.285	595.516

Notes: Standard errors for ancillary parameters are 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	0.975	1.000				
bc88	0.608	0.971	0.991	1.000			
kumb90	0.573	0.984	0.990	0.998	1.000		
bc92	0.603	0.974	0.991	1.000	0.998	1.000	
tre	−0.140	−0.378	−0.405	−0.407	−0.400	−0.406	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-year period (1993–1998). This dataset is interesting because it represents what is generally available to researchers: short panel, information only on input

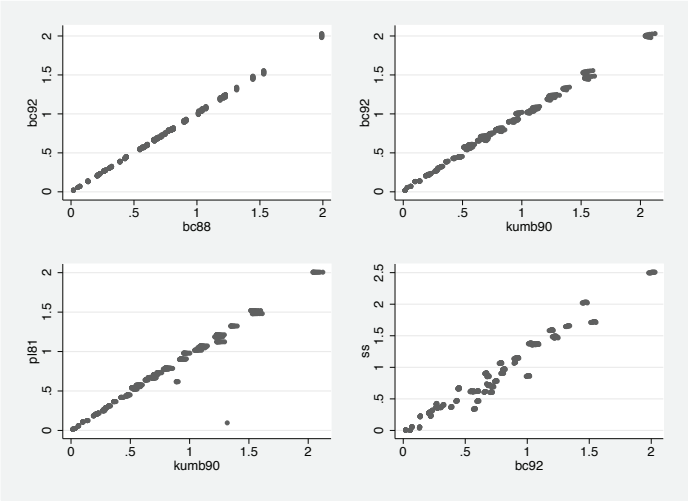

Figure 1. Swiss railways inefficiency scatterplots

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: milk can also be considered an intermediate input to produce dairy products. Furthermore, variables such as slaughtered animals should also be considered 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,

$$\ln y_{it} = \beta_0 + \sum_{j=1}^{4} \beta_j \ln x_{jit} + \frac{1}{2} \sum_{j=1}^{4} \sum_{k=1}^{4} \beta_{jk} \ln x_{jit} \ln x_{kit} + \sum_{t=1993}^{1998} \beta_t \text{dyear}_t - u_{it} + v_{it} \tag{16}$$

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 fit three models with time-varying inefficiency: the normal-half normal Kumbhakar (1990) model (`kumb90`), a random-effects model, fit through the feasible generalized least-squares method; the Cornwell, Schmidt, and Sickles (1990) model (`css90`), fit through the modified least-squares dummy variable technique; and finally, the Lee and Schmidt (1993) model (`ls93`), fit through ILS. Note that the last two models are fit using approaches that do not allow intercept (β_0) and time dummies (dyear_t)

to be simultaneously included in the frontier equation. Finally, we also considered two models with time-invariant inefficiency, that is, the u_{it} term reduced to u_i in (16): the first was proposed by Schmidt and Sickles (1984) and estimated without any distributional assumption through the least-squares dummy variable approach (`ss84`); the second was proposed by Pitt and Lee (1981) and 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 because both parameter significance and magnitudes are comparable. For the `ss84`, `css90`, and `ls93` models, the most efficient firm in the sample for each period is considered as fully efficient; thus the smallest value of inefficiency is 0. On average and as expected, the `css90` model shows a higher level of inefficiency, and its distribution also has 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 and contexts, a time-invariant inefficiency specification is a valid solution.

Table 7. Spanish dairy farms estimation results (247 firms for a total of 1,482 obs.)

	ss84	css90	ls93	kumb90	pl81
	b/se	b/se	b/se	b/se	b/se
x1	0.642***	0.527***	0.641***	0.661***	0.660***
	(0.036)	(0.065)	(0.036)	(0.028)	(0.028)
x2	0.037*	0.043	0.037*	0.038**	0.041**
	(0.017)	(0.027)	(0.017)	(0.015)	(0.015)
x3	0.011	0.079	0.010	0.050**	0.049**
	(0.025)	(0.063)	(0.025)	(0.018)	(0.018)
x4	0.308***	0.226***	0.307***	0.351***	0.356***
	(0.020)	(0.035)	(0.020)	(0.018)	(0.017)
x11	0.135	−0.187	0.133	0.308	0.314
	(0.157)	(0.192)	(0.155)	(0.171)	(0.178)
x22	−0.002	0.060	−0.001	−0.111	−0.112
	(0.069)	(0.111)	(0.068)	(0.064)	(0.067)
x33	−0.242	−0.168	−0.243	−0.129	−0.131
	(0.188)	(0.317)	(0.187)	(0.119)	(0.115)
x44	0.105*	−0.125	0.105*	0.112*	0.118*
	(0.050)	(0.084)	(0.050)	(0.048)	(0.049)
x12	−0.010	0.059	−0.009	−0.060	−0.064
	(0.073)	(0.100)	(0.072)	(0.077)	(0.081)
x13	0.084	−0.114	0.085	0.088	0.091
	(0.102)	(0.158)	(0.101)	(0.090)	(0.090)
x14	−0.075	0.142	−0.074	−0.140	−0.146
	(0.083)	(0.132)	(0.082)	(0.084)	(0.088)
x23	0.001	0.067	0.002	0.020	0.011
	(0.050)	(0.107)	(0.050)	(0.049)	(0.050)
x24	−0.011	−0.062	−0.011	0.025	0.025
	(0.041)	(0.060)	(0.041)	(0.039)	(0.040)
x34	−0.012	0.110	−0.013	−0.015	−0.017
	(0.046)	(0.085)	(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)

Continued on next page

	ss84 b/se	css90 b/se	ls93 b/se	kumb90 b/se	pl81 b/se
t	-	-	-	−0.347	-
	-	-	-	(0.212)	-
t^2	-	-	-	0.045	-
	-	-	-	(0.028)	-
λ	1.948	3.711	2.010	4.485	2.775
σ	0.168	0.237	0.171	0.356	0.230
σ_u	0.149	0.229	0.153	0.348	0.216
σ_v	0.077	0.062	0.076	0.077	0.078
Estimated technical inefficiencies, $\widehat{u}_{it}$					
Mean	0.315	0.531	0.316	0.182	0.179
SD	0.149	0.227	0.150	0.119	0.117
Min	0.000	0.000	0.000	0.008	0.009
Max	0.873	1.412	0.879	0.667	0.623
Log likelihood	-	-	-	1355.248	1351.826

Notes: Cluster–robust standard errors are in parentheses. 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	0.868	1.000			
ls93	1.000	0.871	1.000		
kumb90	0.938	0.726	0.936	1.000	
pl81	0.931	0.709	0.929	0.995	1.000

7 Concluding remarks

In this article, we introduced 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 and efficiency. In the development of the commands, we widely exploit Mata potentiality. By using Mata structures, we provide a very readable code that can be easily developed further by other Stata users.

We illustrated the commands' estimation capabilities through simulated data, focusing on some of the models that cannot be estimated using official Stata commands.

Finally, we illustrated the proposed routines using real datasets under different possible empirical scenarios: short versus long panels, cost versus production frontiers, homogeneous versus heterogeneous outputs.

8 Acknowledgments

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.

9 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: 219–255.

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

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

Battese, G. E., and T. J. 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: 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 Tor Vergata: Research Paper Series.

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

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

Chen, Y.-Y., P. Schmidt, and H.-J. Wang. 2011. Consistent estimation of the fixed effects stochastic frontier model.
http://www.economics.uwo.ca/newsletter/misc/2011/schmidt_nov11.pdf.

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

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

Drukker, D. M., and R. Gates. 2006. Generating Halton sequences using Mata. *Stata Journal* 6: 214–228.

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

Greene, W. 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.

Greene, W. H. 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.

———. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.

Hadri, K. 1999. Estimation of a doubly heteroscedastic stochastic frontier cost function. *Journal of Business and Economic Statistics* 17: 359–363.

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

Horrace, W. C., and P. Schmidt. 1996. Confidence statements 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. *Journal of Productivity Analysis* 5: 171–180.

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

Kumbhakar, S. C. 1990. Production frontiers, panel data, and time-varying technical inefficiency. *Journal of Econometrics* 46: 201–211.

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 and Economic Statistics* 9: 279–286.

Kumbhakar, S. C., and C. A. K. Lovell. 2000. *Stochastic Frontier Analysis*. Cambridge: Cambridge University Press.

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

Lee, Y. H., and P. Schmidt. 1993. A production frontier model with flexible temporal variation in technical efficiency. In *The Measurement of Productive Efficiency: Techniques and Applications*, ed. H. O. Fried, C. A. Knox Lovell, and S. S. Schmidt, 237–255. New York: Oxford University Press.

Meeusen, W., and J. van den Broeck. 1977. Efficiency estimation from Cobb–Douglas production functions with composed error. *International Economic Review* 18: 435–444.

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

Pitt, M. M., and L.-F. 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. C. Sickles. 1984. Production frontiers and panel data. *Journal of Business and Economic Statistics* 2: 367–374.

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

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

Wang, H.-J., 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: 129–144.

About the authors

Federico Belotti is a researcher at the Centre for Economics and International Studies (CEIS) of the University of Rome Tor Vergata.

Silvio Daidone is a research associate at the Centre for Health Economics (CHE) of the University of York and an economist at the Agricultural Development Economics Division of the Food and Agriculture Organization of the United Nations.

Giuseppe Ilardi is a 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.