A Skew Normal Stochastic Frontier Model for Panel Data

Roberto Colombi

Dipartimento di Ingegneria dell' Informazione e Metodi Matematici
Università di Bergamo
Italy

Riunione Scientifica SIS
Padova June 2010

Stochastic frontiers

A stochastic frontier defines a random upper bound for the log of a response variable called *output* as a function of variables called *inputs* plus an idiosyncratic random error

$$y_{it}^{SUP} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + e_{it}$$

Stochastic frontiers and random inefficiency

The observable log of the *output* is given by the stochastic frontier value minus a non negative random variable called subject random inefficiency component

- time dependent inefficiencies

$$y_{it} = y_{it}^{SUP} - u_{it}$$

- time invariant inefficiency

$$y_{it} = y_{it}^{SUP} - u_{i}$$

Heterogeneity and inefficiency random components of a Stochastic frontier for panel data

- subject unobservable heterogeneity is modeled by a random component
- the presence of both a time invariant and a time dependent random inefficiency component is allowed

  subject heterogeneity is not wrongly modeled as inefficiency (Greene, 2005) and it is possible to disentangle a time persistent component from the total inefficiency

The full model for panel data

$$y_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + b_i - u_{i0} - u_{it} + e_{it} \quad (1)$$

y_{it} is the logarithm of the output of the *i-th* unit, $i = 1, 2, ...n$, at time t, $t = 1, 2, ..., T$, T.

The deterministic component $\beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta}$

- $\boldsymbol{x}'_{it}$ is a row vector of p regressors
- the vector $\boldsymbol{\beta}$ and β_0 are unknown parameters

The full model

$$y_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + b_i - u_{i0} - u_{it} + e_{it} \tag{2}$$

The random component $b_i - u_{i0} - u_{it} + e_{it}$

- the random variable e_{it} is the *idiosyncratic random component*
- b_i is the *random subject effect*
- u_{i0} is the *time invariant stochastic inefficiency*
- u_{it} is the *time dependent stochastic inefficiency*

The independence assumptions

- (A1a) for $i = 1, 2, ..., n$, the $(T + 1)$ random variables $u_{i0},\ b_i,\ u_{it},\ e_{it}$, $t = 1, 2, ..., T$, are independent;
- (A1b) the random vectors

$$\boldsymbol{\epsilon_i} = (u_{i0}, b_i, u_{i1}, ..., u_{iT}, e_{i1}, ..., e_{iT}),\ \ i = 1, 2, ..., n$$

are independent;

the distributional assumptions

- (A2a) for every i, u_{i0} is a normal random variable with null expected value and variance σ^2_{1u} left truncated at zero
- (A2b) b_i is a normal random variable with null expected value and variance σ^2_b ;
- (A3a) for every i and t, u_{it} is a normal random variable with null expected value and variance σ^2_{2u} left truncated at zero
- (A3b) e_{it} is a normal random variable with null expected value and variance σ^2_e;
- (A4) the $\boldsymbol{x}_{it}$ are vectors of known constants.

New models that are sub-models of the full model TTT

the model FTT without subject specific component:

$$y_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} - u_{i0} - u_{it} + e_{it} \tag{3}$$

the model TFT without time dependent inefficiency:

$$y_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + b_i - u_{i0} + e_{it} \tag{4}$$

Well known sub-models of the full model TTT

the Greene (2005) true random effect stochastic frontier TTF without time independent inefficiency:

$$y_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + b_i - u_{it} + e_{it} \tag{5}$$

the model FFT (Pitt and Lee model I, 1981) without time dependent inefficiency and subject specific component:

$$y_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} - u_{i0} + e_{it} \tag{6}$$

the pooled stochastic frontier FTF (Pitt and Lee model II, 1981) without time independent inefficiency and subject specific component:

$$y_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} - u_{it} + e_{it}. \tag{7}$$

Sub models that are not stochastic frontiers

the random effect regression model TFF

$$y_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + b_i + e_{it} \tag{8}$$

the classical regression model FFF

$$y_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + e_{it} \tag{9}$$

Comparing nested models

Testing one of the previous models against the full model TTT or comparing nested models (for example TFT against FFT) is a non standard problem because under the null hypothesis one or two parameters are on the boundary of the parametric space. In fact, under reasonable assumptions, in this case the asymptotic distribution of the log-likelihood ratio test statistic is a mixture of chi squared distributions known as *chi-bar-squared distribution* (Silvapulle and Sen, 2005)

The Closed Skew Normal Distribution (Gonzales-Farias *et al.* 2004)

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

(p,q)-CSN distribution

A random vector $\boldsymbol{x}$ has a (p, q) closed skew normal distribution if its probability density function is:

$$f(\boldsymbol{x}, \mu, \boldsymbol{\Gamma}, \boldsymbol{D}, \nu, \boldsymbol{\Delta}, p, q) = \frac{\phi_p(\boldsymbol{x}, \mu, \boldsymbol{\Gamma})\bar{\Phi}_q(\boldsymbol{D}(\boldsymbol{y} - \mu) - \nu, \boldsymbol{\Delta})}{\bar{\Phi}_q(-\nu, \boldsymbol{\Delta} + \boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{D}')}.$$

the special case $\nu = 0$

$$f(\boldsymbol{x}, \mu, \boldsymbol{\Gamma}, \boldsymbol{D}, \boldsymbol{0}, \boldsymbol{\Delta}, p, q) = 2^q \phi_p(\boldsymbol{x}, \mu, \boldsymbol{\Gamma}) \bar{\Phi}_q(\boldsymbol{D}(\boldsymbol{y} - \mu), \boldsymbol{\Delta}).$$

Skew Normality and Closed Skew Normality

In the stochastic frontiers context, Skew Normality and Closed Skew Normality arise from the presence of error components that are defined by a sum of a normal random variable with a left truncated normal random variable.
Closed Skew Normality was implicitly used by Pitt and Lee (1981, Appendix 2) when they discussed their model III (a stochastic frontier for panel data with correlated time dependent inefficiencies).

Skew Normality and Closed Skew Normality

- the time invariant component $v_i = b_i - u_i$ is Skew Normal
- the time dependent components $v_{it} = e_{it} - u_{it}$, $t = 1, 2, ..., T$ are skew normal
- the T+1 random variables $v_i, v_{i1}, v_{i2}, ..., v_{iT}$ are independent
- the T+1 random variables $v_i, v_{i1}, v_{i2}, ..., v_{iT}$ have a $(T+1, T+1)$-Closed Skew Normal joint Distribution
- the T random variables $v_i + v_{i1}, v_i + v_{i2}, ..., v_i + v_{iT}$ have a $(T, T+1)$-Closed Skew Normal joint Distribution

Closed Skew Normality

Under the assumptions A1-A4, the log-likelihood of nT observations is:

$$\sum_{i=1}^{n}(\ln \phi_T(\boldsymbol{y_i}-\boldsymbol{X_i}\beta, \boldsymbol{1_T}\beta_0, \boldsymbol{\Sigma}+\boldsymbol{AVA'})+ln\bar{\Phi}_{T+1}(\boldsymbol{R}(\boldsymbol{y_i}-\boldsymbol{X_i}\beta-\boldsymbol{1_T}\beta_0), \boldsymbol{\Lambda})$$

which is the log-likelihood of the n independent $(T, T+1)$ closed skew normal random variables $\boldsymbol{y_i} - \boldsymbol{X_i}\beta$.

$\boldsymbol{A} = -[\ \boldsymbol{1}_T \quad \boldsymbol{I}_T\ ]$
$\boldsymbol{X}_i$ is the $T \times p$ matrix of regressors

$$\boldsymbol{V} = \begin{bmatrix} \sigma^2_{1u} & \boldsymbol{0}'_T \\ \boldsymbol{0_T} & \sigma^2_{2u}\boldsymbol{I_T} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \sigma^2_e\boldsymbol{I_T} + \sigma^2_b\boldsymbol{1_T}\boldsymbol{1}'_{\boldsymbol{T}} \tag{10}$$

$$\boldsymbol{\Lambda} = \boldsymbol{V} - \boldsymbol{VA'}(\boldsymbol{\Sigma} + \boldsymbol{AVA'})^{-1}\boldsymbol{AV}, \quad \boldsymbol{R} = \boldsymbol{VA'}(\boldsymbol{\Sigma} + \boldsymbol{AVA'})^{-1}. \tag{11}$$

The computational problem

For the models TTT, TTF, FTT the computational complexity of the evaluation of the log-likelihood comes from the $T+1$ multiple integrals

$$\bar{\Phi}_{T+1}(\boldsymbol{R}(\boldsymbol{y_i} - \boldsymbol{X_i}\beta - \mathbf{1_T}\beta_0), \boldsymbol{\Lambda})$$

- For the Pitt and Lee models TFT, FFT with and without the subject random effect, the above integral is one dimensional
- for the pooled model FTF it is a product of T one dimensional integrals

For a survey on efficient numerical and Monte Carlo methods to compute the previous multinormal integral see Genz and Bretz (2009).

Rice data

The Rice panel data set Battese *et al.* (2005, pp. 325-326). In this data set the *annual rice production* of $n = 43$ rice producers in the Tarlac region of the Philippines is reported for $T = 8$ years (from 1990 to 1997).

- *Rice production* (in tonnes) output
- *area planted* (hectares) input ,
- *labor* used (man days) input
- *fertiliser* used (kg)input .

Rice Example

model[1]	max. log-lik.	σ^2_b	σ^2_{2u}	σ^2_{1u}	σ^2_e
TTT	-65.340	0.0001	0.1744	0.0661	0.0161
TFT	-86.430	0.000	0	0.0721	0.0832
FTT	-65.371	0	0.1778	0.650	0.0159
TTF	-68.117	0.0228	0.1794	0	0.0154
FFT	-86.431	0	0	0.0723	0.0832
FTF	-104.907	0	0.000	0	0.1077
TFF	-88.605	0.1584	0	0	0.2901
FFF	-104.907	0	0	0	0.3283

[1] The R (R Development Core Team, 2009) functions: SNF_twostage and SNF_maxlik, developed by the author, were used to fit the models of the Table.

Wagepan data

The WAGEPAN panel data set examined by Wooldridge (2006). In this data set the wages of $n = 545$ workers are reported for $T = 8$ years.

- *wage* output
- *hours of work* input
- *years of education*, *experience hispanic*, *black*, *married*, *union*

Wagepan example

model	max. log-lik.	σ_b^2	σ_{2u}^2	σ_{1u}^2	σ_e^2
TTT	-1807.251	0.109	0.245	0.000	0.022
TFT	-2182.970	0.110	0	0.000	0.123
FTT	-1977.245	0	0.209	0.408	0.040
TTF	-1807.514	0.109	0.239	0	0.023
FFT	-2226.459	0	0	0.508	0.136
FTF	-2856.730	0	0.341	0	0.100
TFF	-2182.970	0.110	0	0	0.123
FFF	-2984.890	0	0	0	0.231