Journal of Econometrics 190 (2016) 341–348

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Some models for stochastic frontiers with endogeneity

William E. Griffiths *, Gholamreza Hajargasht

Department of Economics, University of Melbourne, Australia

ARTICLE INFO

Article history:
Available online 25 June 2015

JEL classification:
C11, D24, C23, C12

Keywords:
Technical efficiency
Instrumental variables
Gibbs sampling

ABSTRACT

We consider mostly Bayesian estimation of stochastic frontier models where one-sided inefficiencies and/or the idiosyncratic error term are correlated with the regressors. We begin with a model where a Chamberlain–Mundlak device is used to relate a transformation of time-invariant effects to the regressors. This basic model is then extended in two directions: first an extra one-sided error term is added to allow for time-varying efficiencies. Second, a model with an equation for instrumental variables and a more general error covariance structure is introduced to accommodate correlations between both error terms and the regressors. An application of the first and second models to Philippines rice data is provided.

1. Introduction

Studies of stochastic frontier models that allow for correlation between inefficiency effects and regressors are few and have been mainly done under a fixed effects framework in which a panel data model with a two-sided error term is estimated first, and the inefficiency effects are later estimated by subtracting the effects from their maximum (see e.g. Sickles, 2005 and references cited therein). Given that stochastic frontier models are more commonly estimated based on a one-sided random effects assumption, it is useful to investigate estimation within a framework where the one-sided random effects are correlated with the regressors. Also of interest are methods for accommodating correlation between the idiosyncratic error term and the regressors. The purpose of this paper is to propose a relatively general approach to modelling of stochastic frontiers with endogeneity, where one-sided efficiency effects, and idiosyncratic error terms, can be correlated with the regressors. We show that by transforming the inefficiency term into a normally distributed random term and modelling endogeneity through the mean or covariance of the normal errors, a range of stochastic frontier models with endogeneity can be handled.

We first consider a panel stochastic frontier model in which correlations between the effects and the regressors are based on a generalisation of the correlated random effects model proposed by Mundlak (1978), extended by Chamberlain (1984), and described further by Wooldridge (2010). Inefficiency effects are assumed to be correlated with the regressors through the mean of a transformation of the inefficiency errors. The main focus is on a log transformation implying the inefficiency errors have a lognormal distribution whose first argument depends on the regressors. Pursuing Bayesian estimation of the model, we derive conditional posterior densities for the parameters and the inefficiency errors for use in a Gibbs sampler. We then extend the model in two directions. Following Colombi et al. (2011), we add a time-varying inefficiency error leading to a model with both time invariant (permanent) and time-varying (transient) inefficiency errors; endogeneity is assumed to occur through correlation between the regressors and the time-invariant error. Necessary changes to the previously specified conditional posteriors are described. The second extension is to a more general model where endogeneity can exist because both the inefficiency errors and the idiosyncratic errors are correlated with the regressors. So that estimation can proceed, a "reduced form" type equation with instrumental variables is added to the earlier model. Details of how to estimate the model using both maximum simulated likelihood and Bayesian methods are provided.

The paper is organised as follows. The basic Mundlak-type model where the mean of the transformed error is a function of the regressors is considered in Section 2. In Section 3 we extend this model to include both permanent and transient inefficiency errors. Specification and estimation of the model that makes provision for instrumental variables and accommodates endogeneity more generally are considered in Section 4. An application using Philippine rice data and the models from Sections 2 and 3 is provided in Section 5.

* Correspondence to: Economics Department, University of Melbourne, Vic 3010, Australia. Tel.: +61 3 8344 3622; fax: +61 3 8344 6899.

E-mail addresses: wegrif@unimelb.edu.au (W.E. Griffiths), har@unimelb.edu.au (G. Hajargasht).

http://dx.doi.org/10.1016/j.jeconom.2015.06.012
0304-4076/

2. Modelling correlation with a Chamberlain–Mundlak device

In the first instance we consider the following random effects stochastic production frontier model with a time invariant inefficiency term

$$y_{it} = \mathbf{x}_{1,it}\boldsymbol{\beta} - u_i + v_{it}. \tag{2.1}$$

In Eq. (2.1), $i = 1, \ldots, N$ indexes the firms and $t = 1, \ldots, T$ indexes time, $\mathbf{x}_{1,it}$ is a row vector of functions of inputs (e.g., logs of inputs and squared logs of inputs), y_{it} represents the logarithm of output, $\mathbf{x}_{1,it}\boldsymbol{\beta}$ is the log of the frontier production function (e.g., translog), u_i is a non-negative random error which accounts for time-invariant inefficiency of firm i, and v_{it} is an idiosyncratic error assumed to be *i.i.d.* $N\left(0, \sigma^2\right)$. The model can also represent a stochastic cost frontier, with y_{it} being the logarithm of cost, by changing "$-u_i$" to "$+u_i$".

In view of recent developments in the stochastic frontier literature – see, for example, Parmeter and Kumbhakar (2014) – having a model with time-invariant inefficiencies can be considered too restrictive. However, we include this model in the first instance as a stepping stone to more realistic time-varying inefficiency models considered in Sections 3 and 4.

To model correlation between the inefficiency error u_i and some or all of the inputs we assume that there is a transformation of u_i, call it $H(u_i)$, that is normally distributed with a mean that depends on the firm averages of some of the inputs or functions of them. These functions of the inputs are collected in the vector $\mathbf{x}_{2,it}$ and their firm averages are given by $\bar{\mathbf{x}}_{2,i} = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{2,it}$. The resulting endogeneity model for describing how the inefficiency error is correlated with the inputs is given by

$$H(u_i) = \bar{\mathbf{x}}_{2,i}\boldsymbol{\gamma} + e_i, \tag{2.2}$$

with $e_i \sim$ *i.i.d.* $N(0, \lambda^2)$. The most convenient transformation in the sense that it leads to recognisable conditional posterior distributions for implementing Gibbs sampling is the logarithmic one, $H(u_i) = \ln(u_i)$, implying that u_i has a lognormal distribution. Other transformations [e.g., $(u_i^{\rho} - 1)/\rho$ for some values of ρ] are possible.[1]

Eq. (2.2) is an extension of the model considered by Mundlak (1978) for a conventional random effects panel data model with correlated effects. Modelling of endogeneity in this way, and its extension by Chamberlain (1984), have been referred to as the Chamberlain–Mundlak device, a device which has proved to be very useful in the context of nonlinear panel data models with endogeneity. It has been applied to model endogeneity in probit, fractional response, Tobit, sample selection, count data, double hurdle, unbalanced panel models, and models with cluster sampling. See Wooldridge (2010) for a review and for references to these applications. Also, when $H(u_i) = \ln(u_i)$, Eq. (2.2) can be written as $u_i = \exp\left\{\bar{\mathbf{x}}_{2,i}\boldsymbol{\gamma}\right\}u_i^*$ where $u_i^* = \exp(e_i)$, implying the model can also be viewed as a stochastic frontier model with scaling properties. Alvarez et al. (2006) have studied and argued in favour of the scaling property in the context of models with environmental variables.

2.1. Prior specification

For Bayesian estimation of the model in (2.1)–(2.2), we begin by specifying prior distributions, and then present the conditional posterior densities that can be used for Gibbs sampling. For $\boldsymbol{\beta}$, we use the noninformative prior $p(\boldsymbol{\beta}) \propto 1$; for the variance of v_{it}, we use $\sigma^{-2} \sim G(A_\sigma, B_\sigma)$ where $G(A_\sigma, B_\sigma)$ denotes a gamma density with shape parameter A_σ and scale parameter B_σ; a truncated normal distribution denoted by $\boldsymbol{\gamma} \sim TN\left(\underline{\boldsymbol{\gamma}}, \mathbf{V}_\gamma; \mathbf{L}, \mathbf{U}\right)$ is used for $\boldsymbol{\gamma}$. The truncated normal parameters $\underline{\boldsymbol{\gamma}}$ and $\mathbf{V}_\gamma$ are what would be the prior mean vector and covariance matrix for $\boldsymbol{\gamma}$ if there were no truncation; $\mathbf{L}$ and $\mathbf{U}$ are vectors containing the lower and upper truncation points for each of the elements in $\boldsymbol{\gamma}$. For λ two alternative priors were considered: a gamma prior on λ^{-2} and a truncated uniform prior on λ, written as $\lambda^{-2} \sim G(A_\lambda, B_\lambda)$ and $\lambda \sim U(a_\lambda, b_\lambda)$, respectively.

The choice of priors for $\boldsymbol{\beta}$ and σ^{-2} is standard. For $\boldsymbol{\gamma}$ and λ, we experimented with several alternative priors, considering in each case their implications for (1) MCMC convergence, and (2) the marginal prior distributions of the inefficiency errors and their efficiencies, defined as $r_i = \exp(-u_i)$. Truncating a normal prior for $\boldsymbol{\gamma}$ to values that lead to reasonable efficiency values led to more precise estimates and improved MCMC convergence. A gamma prior for λ^{-2} is in line with most traditional priors specified for variance parameters, while use of a uniform prior for standard deviations in hierarchical models (which bear some similarity to our model) has been advocated by Gelman (2006). We defer discussion on the setting of values for the prior parameters to the application in Section 5.

2.2. Conditional posterior densities

To use Gibbs sampling for estimation we begin by considering the conditional posterior densities when $H(u_i) = \ln(u_i)$ and the prior $\lambda^{-2} \sim G(A_\lambda, B_\lambda)$ is used. Define $\mathbf{u}' = (u_1, u_2, \ldots, u_N)$; let $\mathbf{X}$ be a matrix with NT rows and typical row $\mathbf{x}_{1,it}$ and $\mathbf{X}_2$ be a matrix with N rows and typical row $\bar{\mathbf{x}}_{2,i}$. The joint posterior kernel for $\boldsymbol{\Theta} = \left(\boldsymbol{\beta}, \sigma^{-2}, \boldsymbol{\gamma}, \lambda^{-2}, \mathbf{u}\right)$ is

$$\begin{aligned}
p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X}, \mathbf{X}_2) &\propto p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^{-2}, \mathbf{u}\right) p\left(\mathbf{u}|\mathbf{X}_2, \boldsymbol{\gamma}, \lambda^{-2}\right) p(\boldsymbol{\beta}) \\
&\quad \times p\left(\sigma^{-2}\right) p(\boldsymbol{\gamma}) p\left(\lambda^{-2}\right) \\
&\propto \left(\sigma^{-2}\right)^{NT/2+A_\sigma-1} \exp\left\{-\frac{\sigma^{-2}}{2}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \mathbf{x}_{1,it}\boldsymbol{\beta} + u_i\right)^2 \right.\right. \\
&\quad \left.\left. + \ 2B_\sigma\right]\right\} \left(\lambda^{-2}\right)^{N/2+A_\lambda-1}\left[\prod_{i=1}^{N} u_i^{-1}\right] \\
&\quad \times \exp\left\{-\left[\frac{\lambda^{-2}}{2}\sum_{i=1}^{N}\left(\ln u_i - \bar{\mathbf{x}}_{2,i}\boldsymbol{\gamma}\right)^2 + 2B_\lambda\right]\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\gamma} - \underline{\boldsymbol{\gamma}}\right)'\mathbf{V}_\gamma^{-1}\left(\boldsymbol{\gamma} - \underline{\boldsymbol{\gamma}}\right)\right\} \\
&\quad \times \left[\prod_{s=1}^{S} I\left(L_s \le \gamma_s \le U_s\right)\right]
\end{aligned} \tag{2.3}$$

where $I(L_s \le \gamma_s \le U_s)$ is an indicator function, L_s, U_s and γ_s are elements of $\mathbf{L}$, $\mathbf{U}$ and $\boldsymbol{\gamma}$, respectively, and S is the dimension of $\boldsymbol{\gamma}$. If we use the uniform prior $\lambda \sim U(a_\lambda, b_\lambda)$, then the joint posterior density can be obtained from (2.3) by setting $A_\lambda = 1$, $B_\lambda = 0$, and including the indicator function $I(a_\lambda \le \lambda \le b_\lambda)$. From Eq. (2.3), and using $\mathbf{D} = \{\mathbf{y}, \mathbf{X}, \mathbf{X}_2\}$ to denote the available data, the following conditional posterior densities can be derived:

$$(\boldsymbol{\beta}|\boldsymbol{\Theta}_{-\boldsymbol{\beta}}, \mathbf{D}) \sim N\left\{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{y} + \mathbf{u} \otimes \mathbf{i}_T), \quad \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right\}, \tag{2.4}$$

$$\left(\sigma^{-2}|\ \boldsymbol{\Theta}_{-\sigma^{-2}}, \mathbf{D}\right) \sim G\left(A_\sigma + NT/2, \quad B_\sigma + \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \mathbf{x}_{1,it}\boldsymbol{\beta} + u_i\right)^2\right), \tag{2.5}$$

---

[1] One can in fact assume any distribution for u_i (e.g., exponential), with its cdf denoted by $F(u_i)$, and use the transformation $H(u_i) = \lambda\Phi^{-1}(F(u_i)) + \bar{\mathbf{x}}_{2,i}\boldsymbol{\gamma}$, but the posterior density must then include extra parameters from $F(u_i)$.