Journal of Econometrics 190 (2016) 280–288

Contents lists available at ScienceDirect

Journal of Econometrics

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

Endogeneity in stochastic frontier models

Christine Amsler[a], Artem Prokhorov[b], Peter Schmidt[a,*]

[a] Michigan State University, USA
[b] University of Sydney, Australia

ARTICLE INFO

Article history:
Available online 23 June 2015

JEL classification:
C10
C26
C36

Keywords:
Endogeneity
Stochastic frontier
Efficiency measurement

ABSTRACT

Stochastic frontier models are typically estimated by maximum likelihood (MLE) or corrected ordinary least squares. The consistency of either estimator depends on exogeneity of the explanatory variables (inputs, in the production frontier setting). We will investigate the case that one or more of the inputs is endogenous, in the simultaneous equation sense of endogeneity. That is, we worry that there is correlation between the inputs and statistical noise or inefficiency.

In a standard regression setting, simultaneity is handled by a number of procedures that are numerically or asymptotically equivalent. These include 2SLS; using the residual from the reduced form equations for the endogenous variables as a control function; and MLE of the system that contains the equation of interest plus the unrestricted reduced form equations for the endogenous variables (LIML). We will consider modifications of these standard procedures for the stochastic frontier setting.

The paper is mostly a survey and combination of existing results from the stochastic frontier literature and the classic simultaneous equations literature, but it also contains some new results.

1. Introduction

In this paper we consider the stochastic frontier (SF) model

$$y_i = \alpha + x_i'\beta + v_i - u_i, \quad i = 1, \ldots, n, \tag{1}$$

where y_i is log output, x_i is a vector of inputs or functions of inputs, v_i is random noise distributed as $N(0, \sigma_v^2)$, and $u_i \geq 0$ represents technical inefficiency. Here i indexes firms and n is the number of firms. We are interested in the case that some of the x's may be endogenous, in the sense that they are correlated with v or u or both. This can occur when there is feedback from either statistical noise or inefficiency to the choice of inputs, or when the inputs influence the level of inefficiency as well as the frontier. Endogeneity needs to be dealt with because the usual procedures for estimating SF models depend on the assumption that the inputs are exogenous.

In a standard regression setting, simultaneity is handled by a number of procedures that are numerically or asymptotically equivalent. These include instrumental variables (2SLS); using the residual from the reduced form equations for the endogenous variables as a control function; and MLE of the system that contains the equation of interest plus the reduced form equations for the endogenous variables (LIML). We will consider modifications of these standard procedures for the SF setting. An important issue is that procedures that are numerically or asymptotically equivalent in the usual linear regression model may not be equivalent for the SF model. Another important issue is that it is definitely not appropriate to insert "fitted values" for the endogenous variables and then proceed with standard SF procedures such as the usual SF MLE.

Modification of the first three of these procedures to the SF model is straightforward. However, appropriate modification of LIML is not straightforward, because it is not clear how best to model the joint distribution of the composed error in the SF model and the error in the reduced form equations for the endogenous inputs. This is a potentially important issue because correlation between the reduced form errors and either noise or inefficiency can be helpful in the decomposition of the composed error into its noise and inefficiency components.

This paper is mostly a survey and combination of existing results from the SF literature and the classic simultaneous equations literature, but it also contains some new results. The material in this paper may be assumed to be part of the existing literature unless it is specifically claimed to be new. The plan of the paper is as follows. In Section 2 we give a brief review of estimation of stochastic frontier models, and in Section 3 we give a brief review of 2SLS

[*] Correspondence to: Department of Economics, Michigan State University, East Lansing, MI 48824, USA. Tel.: +1 517 355 8381.
E-mail address: schmidtp@msu.edu (P. Schmidt).

http://dx.doi.org/10.1016/j.jeconom.2015.06.013
0304-4076/© 2015 Elsevier B.V. All rights reserved.

and LIML in the usual linear simultaneous equations model. In Section 4 we consider stochastic frontier models with endogeneity, and we discuss how the simple 2SLS and LIML estimators can be modified for use in the stochastic frontier model. We also discuss some issues that are relevant in the case of a translog model (or other nonlinear models). In Section 5 we give an empirical example. Finally, Section 6 gives our concluding remarks.

2. A brief review of estimation in SF models

This section will give a very brief review of the estimation of SF models under exogeneity. This is all standard material but it allows us to define some necessary notation and to summarize the relevant results for readers who are not knowledgeable about SF models.

The most common way to estimate the SF model is by MLE. Following standard terminology, we define $\varepsilon_i = v_i - u_i = y_i - \alpha - x_i'\beta$, which is the *composed error*. We will make the standard assumptions (Aigner et al., 1977) that we have random sampling (and therefore independence) over i, that x_i, v_i and u_i are mutually independent, that $v_i \sim N(0, \sigma_v^2)$, and that $u_i \sim N^+(0, \sigma_u^2)$. (That is, u_i has the so-called *half normal* distribution.) The implied density of ε_i is

$$f_\varepsilon(\varepsilon_i) = \int_0^\infty f_v(\varepsilon_i + u) f_u(u)\, du = \frac{2}{\sigma}\varphi\left(\frac{\varepsilon_i}{\sigma}\right)\Phi\left(-\frac{\lambda\varepsilon_i}{\sigma}\right), \qquad (2)$$

where: $\sigma^2 = \sigma_u^2 + \sigma_v^2$; $\lambda = \sigma_u/\sigma_v$; φ is the standard normal density function; and Φ is the standard normal cdf. We can then form the likelihood function: $\ln L = \sum_i \ln f_\varepsilon(y_i - \alpha - x_i'\beta)$.

The MLE's of the parameters of the model are obtained by maximizing the likelihood function with respect to the parameters $\alpha, \beta, \lambda, \sigma^2$ (or, equivalently, $\alpha, \beta, \sigma_u^2, \sigma_v^2$).

An alternative to MLE is corrected ordinary least squares (COLS), which was defined in Aigner et al. (1977) and Olson et al. (1980). We can make the same assumptions as above, or the slightly weaker assumptions that, conditional on x_i, the first three moments of v_i are the moments of $N(0, \sigma_v^2)$, the first three moments of u_i are the moments of $N^+(0, \sigma_u^2)$, and v_i and u_i are independent. Define $\mu = E(u) = \sqrt{\frac{2}{\pi}}\sigma_u$. Let $\hat{\alpha}$ and $\hat{\beta}$ be the OLS estimates when y is regressed on x. These are consistent estimators of $(\alpha - \mu)$ and β, respectively. Now define the OLS residuals $e_i = y_i - \hat{\alpha} - x_i'\hat{\beta}$. The second and third sample moments of the residuals are $\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_i e_i^2$ and $\hat{\mu}_3' = \frac{1}{n}\sum_i e_i^3$. These are consistent estimators of $\sigma_\varepsilon^2 = \sigma_v^2 + \frac{\pi-2}{\pi}\sigma_u^2$ and $\mu_3' = E[\varepsilon - E(\varepsilon)]^3 = \frac{\pi-4}{\pi}\sqrt{\frac{2}{\pi}}\sigma_u^3$. Solving for σ_u^2 and σ_v^2, in terms of sample quantities we have

$$\hat{\sigma}_u^2 = \left(\frac{\pi}{\pi-4}\sqrt{\frac{\pi}{2}}\hat{\mu}_3'\right)^{2/3}, \qquad \hat{\sigma}_v^2 = \hat{\sigma}_\varepsilon^2 - \frac{\pi-2}{\pi}\hat{\sigma}_u^2. \qquad (3)$$

This presumes that $\hat{\mu}_3' < 0$. (It is the case that $\mu_3' < 0$, but because of estimation error it is possible that $\hat{\mu}_3' > 0$.) If $\hat{\mu}_3' > 0$, the so-called *wrong skew problem*, we set $\hat{\sigma}_u^2 = 0$ (Waldman, 1982). We can now correct the intercept: $\tilde{\alpha} = \hat{\alpha} + \sqrt{\frac{2}{\pi}}\hat{\sigma}_u$. Then the COLS estimates are $\tilde{\alpha}, \hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_v^2$.

There is no real case for preferring the COLS estimate to the MLE in the current setting, but as we will see it is easy to generalize to models with endogeneity.

Once the parameters have been estimated, the ultimate aim is to estimate (or, more properly, *predict*) the values of the inefficiency terms u_i. Under the assumptions that were made in the discussion of MLE above, Jondrow et al. (1982) showed that the distribution of u_i conditional on ε_i is $N^+\left(a_i, \sigma_*^2\right)$ where $a_i = -\varepsilon_i\sigma_u^2/\sigma^2$ and $\sigma_*^2 = \sigma_u^2\sigma_v^2/\sigma^2$. Then the prediction of u_i is the mean of this distribution:

$$\hat{u}_i = E(u_i|\varepsilon_i) = \sigma_*\left[\frac{\varphi(b_i)}{1 - \Phi(b_i)} - b_i\right] \quad \text{where } b_i = \varepsilon_i\lambda/\sigma. \qquad (4)$$

To implement this formula, it must be evaluated at the estimated parameters ($\hat{\alpha}, \hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_v^2$ and the implied values of $\hat{\lambda}$ and $\hat{\sigma}^2$) and at $\hat{\varepsilon}_i = y_i - \hat{\alpha} - x_i'\hat{\beta}$. (Here, with a slight abuse of notation, $\hat{\alpha}$, $\hat{\beta}$, etc. can be either the MLE or the COLS estimates.)

3. A brief review of 2SLS and LIML

This section will give a very brief review of the estimation of linear models (not SF models) when some variables may be endogenous. This is all standard material but the discussion allows us to define some necessary notation and to summarize the relevant results that will be generalized to the stochastic frontier model.

The model of interest is

$$y_i = x_i'\beta + v_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + v_i, \quad i = 1, \ldots, n. \qquad (5)$$

Here x_{1i} is *exogenous*, meaning $E(v_i|x_{1i}) = 0$ (loosely, x_{1i} is not correlated with v_i) and x_{2i} is *endogenous*, meaning $E(v_i|x_{2i}) \neq 0$ (loosely, x_{2i} is correlated with v_i). There are k_1 variables in x_{1i} and k_2 variables in x_{2i}. The intercept is part of x_{1i}. In matrix terms we write the model as $y = X\beta + v = X_1\beta_1 + X_2\beta_2 + v$ where y is $n \times 1$, X_1 is $n \times k_1$, etc.

We assume there are some *instruments* $z_i = \begin{bmatrix} x_{1i} \\ w_i \end{bmatrix}$ with w_i of dimension $k_w \geq k_2$, so there are at least as many instruments as x's. We say that the model is *exactly identified* when $k_w = k_2$ and that it is *overidentified* when $k_w > k_2$. The instruments are exogenous, in the sense that $E(v_i|z_i) = 0$. We can think in terms of a reduced form for the endogenous variables, which we write in matrix terms as

$$X_2 = Z\Pi + \eta \qquad (6)$$

where $Z = (X_1, W)$ and where η_i is uncorrelated with z_i. Then endogeneity of X_2 corresponds to $\text{cov}(\eta v) \neq 0$.

The problem that endogeneity causes (*simultaneous equations bias*) is that ordinary least squares is inconsistent. This occurs because $E(v|X_2) \neq 0$, and therefore $E(y|X_1, X_2) \neq X_1\beta_1 + X_2\beta_2$, so the regression model is not valid.

We now discuss standard methods to obtain consistent estimates in the presence of endogeneity. These are the methods that we will later generalize to the SF model.

3.1. Two stage least squares (2SLS)

Let $\hat{\Pi} = (Z'Z)^{-1}Z'X_2$ be the ordinary least squares estimate of the reduced form (6), and let $\hat{X}_2 = Z\hat{\Pi}$ and $\hat{\eta} = X_2 - \hat{X}_2$ be the corresponding fitted values and residuals, respectively. Also define $\hat{X} = (X_1, \hat{X}_2)$. Then the 2SLS (or *instrumental variables*, IV) estimator of β in (5) is

$$\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'y = (\hat{X}'X)^{-1}\hat{X}'y. \qquad (7)$$

This estimator is consistent if the instruments are exogenous (as defined above) and if there are enough relevant instruments (the model is identified).

An alternative approach that is equivalent to 2SLS in the linear model uses a so-called *control function*. In principle, we could control for the effect of η on v by including η in the regression (7). This would be a regression of the form $y = X\beta + \eta\xi + error$, where $\xi = \Sigma_{\eta\eta}^{-1}\Sigma_{\eta v}$ and $error = v - \eta\xi$. Least squares applied