8-16-2017

Endogeneity in Panel Stochastic Frontier Models: An Application to the Japanese Cotton Spinning Industry

Mustafa U. Karakaplan
Governors State University

Levent Kutlu
The University of Texas Rio Grande Valley, levent.kutlu@utrgv.edu

Endogeneity in Panel Stochastic Frontier Models:

An Application to the Japanese Cotton Spinning Industry

Mustafa U. Karakaplan[a] and Levent Kutlu[b]

Abstract

We present a panel stochastic frontier model that handles the endogeneity problem. This model can treat the endogeneity of both frontier and inefficiency variables. We apply our method to examine the technical efficiency of Japanese cotton spinning industry. Our results indicate that market concentration is endogenous, and when its endogeneity is properly handled, it has a larger negative impact on the technical efficiency of cotton spinning plants. We find that the exogenous model substantially overestimates efficiency in concentrated markets.

Keywords: Stochastic Frontier; Panel Data; Endogeneity; Efficiency; Japan; Cotton Industry

JEL Classification: C13, D24

[a] Georgetown University, Department of Economics, Washington, DC, 20057, USA.

E-mail: muk3@georgetown.edu, Phone: +1-202-687-6109

[b] Georgia Institute of Technology, School of Economics, Atlanta, GA, 30332, USA.

E-mail: levent.kutlu@econ.gatech.edu, Phone: +1-404-385-1363

About the authors: Mustafa U. Karakaplan has a PhD in Economics from Texas A&M University. Levent Kutlu has a PhD in Economics from Rice University.

1. Introduction

Maximum likelihood estimation is probably the most widely used method in the stochastic frontier literature. However, if the model has endogeneity problem, then the traditional maximum likelihood estimation for stochastic frontier models (SFM) gives inconsistent parameter estimates. This necessitates a proper instrumental variable (IV) approach in order to deal with the endogeneity issue. A standard way to handle this problem is modeling the joint distribution of the left-hand-side variable and endogenous variables; and then maximizing the corresponding log-likelihood. Due to the special nature of the error term in the SFMs, this is a relatively more difficult task compared to standard maximum likelihood models involving only two-sided error terms.

In the panel data framework, Kutlu (2010) provides a maximum likelihood model that enables estimation of producer specific cost (or technical) efficiencies when some of the frontier regressors are correlated with the two-sided error term. Tran and Tsionas (2013) propose estimating the same model with GMM.[1] Both of these studies assume that the one-sided error term (inefficiency component) is independent from two-sided error term. This assumption is not unlikely to be violated in practice. Karakaplan and Kutlu (2017a) solve the endogeneity problem for both cases in the cross-sectional data setting.[2] Panel data can potentially give more reliable information about the efficiency. Hence, we provide a panel data model that can handle both types of endogeneity.

In the empirical section, we examine the technical efficiency of Japanese cotton spinning industry.[3] In particular, we examine the relationship between technical efficiency and market

[1] Guan et al. (2009) is another GMM based estimator that is solving endogeneity of frontier variables.

[2] See also Amsler et al. (2016) for another cross sectional study.

[3] We programmed the estimator using Stata 13. The Stata ado files are available upon request.

concentration, and use our empirical model as an example to illustrate the dangers of ignoring endogeneity in stochastic frontier models. In our model, we find that market concentration has a negative effect on efficiency, which is in line with the quiet life hypothesis of Hicks (1935). The quiet life hypothesis argues that in concentrated markets, due to lack of competitive pressure, the managers are likely to show less effort, which in turn results in suboptimal profit or production levels. Due to lack of econometric tools, historically potential endogeneity of market concentration in stochastic frontier models is either ignored or tried to be handled by pseudo econometric techniques. We overcome this difficulty using our panel stochastic frontier model, and show that, as expected, market concentration is endogenous in the model, which undermines the estimates from a standard SFM.

2. Panel Stochastic Frontier Model and Endogeneity Test

Our stochastic frontier panel data model is given as follows:

$$y_{it} = x'_{\mathrm{y}it}\beta + v_{it} - \mathrm{s}u_{it} \tag{1}$$

$$x_{it} = Z_{it}\delta + \varepsilon_{it}$$

$$\begin{bmatrix} \tilde{\varepsilon}_{it} \\ v_{it} \end{bmatrix} \equiv \begin{bmatrix} \Omega^{-1/2}\varepsilon_{it} \\ v_{it} \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_p & \sigma_v\rho \\ \sigma_v\rho' & \sigma_v^2 \end{bmatrix} \right)$$

$$u_{it} = \mathrm{h}(x'_{uit}\varphi_u)u_i^*$$

$$s = \begin{cases} 1 & \text{for production functions} \\ -1 & \text{for cost functions} \end{cases}$$

where y_{it} is the logarithm of the output or cost of the i^{th} productive unit at time t; x_{yit} is a vector of exogenous and endogenous variables; x_{it} is a $p \times 1$ vector of all endogenous variables (excluding y_{it}), $Z_{it} = I_p \otimes z'_{it}$ where z_{it} is a $q \times 1$ vector of all exogenous variables, v_{it} and ε_{it} are two-sided error terms, $u_{it} \geq 0$ is a one-sided error term capturing the inefficiency, $h_{it} =$

$\mathrm{h}(x'_{uit}\varphi_u) > 0$, $x_{\mathrm{u}it}$ is a vector of exogenous and endogenous variables excluding the constant, and u_i^* is a producer-specific random component independent from v_{it} and ε_{it}. Here, Ω is the variance-covariance matrix of ε_{it}, σ_v^2 is the variance of v_{it}, and ρ is the vector representing correlation between $\tilde{\varepsilon}_{it}$ and v_{it}. Hence, u_{it} and v_{it} can be correlated with x_{it}, yet u_{it} and v_{it} are conditionally independent given x_{it} and z_{it}. Similarly, u_{it} and ε_{it} are conditionally independent given x_{it} and z_{it}. By a Cholesky decomposition of the variance-covariance matrix of $(\tilde{\varepsilon}'_{it}, v_{it})'$, we can represent $(\tilde{\varepsilon}'_{it}, v_{it})'$ as follows:

$$\begin{bmatrix} \tilde{\varepsilon}_{it} \\ v_{it} \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ \sigma_v\rho' & \sigma_v\sqrt{1-\rho'\rho} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_{it} \\ \widetilde{w}_{it} \end{bmatrix} \tag{2}$$

where $\tilde{\varepsilon}_{it}$ and $\widetilde{w}_{it} \sim \mathbf{N}(0,1)$ are independent. The frontier equation can be written as:

$$\begin{aligned} y_{it} &= x'_{\mathrm{y}it}\beta + \sigma_v\rho'\tilde{\varepsilon}_{it} + w_{it} - \mathrm{s}u_{it} \\ &= x'_{\mathrm{y}it}\beta + (x_{it} - Z_{it}\delta)'\eta + e_{it} \end{aligned} \tag{3}$$

where $e_{it} = w_{it} - \mathrm{s}u_{it}$, $w_{it} = \sigma_v\sqrt{1-\rho'\rho}\widetilde{w}_{it} = \sigma_w\widetilde{w}_{it}$, and $\eta = \sigma_w\Omega^{-1/2}\,\rho/\sqrt{1-\rho'\rho}$. An important aspect of this setup is that e_{it} is conditionally independent from the regressors given x_{it} and z_{it}. In Equation (3) the term $(x_{it} - Z_{it}\delta)'\eta$ serves as a bias correction term. We assume that:

$$\begin{aligned} u_i^* &\sim \mathbf{N}^+(\mu, \sigma_u^2) \\ h_{it}^2 &= \exp(x'_{\mathrm{u}it}\varphi_u). \end{aligned} \tag{4}$$

A vector of observations corresponding to the panel i will be represented by a subscipt i. For example, $h_{i.} = (h_{i1}, h_{i2}, \ldots, h_{iT_i})'$ is a $T_i \times 1$ vector where T_i is the number of time periods for panel i. The log-likelihood function of panel i is given by:

$$\ln L_i = \ln L_{i,y|x} + \ln L_{i,x} \tag{5}$$

where

$$\ln L_{i,y|x} = -\frac{1}{2}\left(T_i \ln(2\pi\sigma_w^2) + \frac{e_{i.}' e_{i.}}{\sigma_w^2} + \left(\frac{\mu^2}{\sigma_u^2} - \frac{\mu_{i*}^2}{\sigma_{i*}^2}\right)\right) + \ln\left(\frac{\sigma_{i*}\Phi\left(\frac{\mu_{i*}}{\sigma_{i*}}\right)}{\sigma_u \Phi\left(\frac{\mu}{\sigma_u}\right)}\right)$$

$$\ln L_{i,x} = -\frac{1}{2}\sum_{t=1}^{T_i} (\ln(|2\pi\Omega|) + \varepsilon_{it}' \, \Omega^{-1} \varepsilon_{it})$$

$$\mu_{i*} = \frac{\sigma_w^2 \mu - s\sigma_u^2 e_{i.}' h_{i.}}{\sigma_u^2 h_{i.}' h_{i.} + \sigma_w^2}$$

$$\sigma_{i*}^2 = \frac{\sigma_u^2 \sigma_w^2}{\sigma_u^2 h_{i.}' h_{i.} + \sigma_w^2}$$

$$e_{it} = y_{it} - x_{1it}'\beta - \varepsilon_{it}'\eta$$

$$\varepsilon_{it} = x_{it} - Z_{it}\delta$$

where Φ denotes the standard normal CDF. We predict the efficiency, $EFF_{it} = \exp(-u_{it})$, by:

$$\exp(-\mathrm{E}[u_{it}|e_i]) = \exp\left(-h_{it}\left(\mu_{i*} + \frac{\sigma_{i*}\phi\left(\frac{\mu_{i*}}{\sigma_{i*}}\right)}{\Phi\left(\frac{\mu_{i*}}{\sigma_{i*}}\right)}\right)\right) \tag{6}$$

where ϕ denotes the standard normal PDF.

Note that unlike the standard control function methods where estimations are done in two-stages, our model estimates the parameters in a single stage. Compared to two-stage methods, our model has the advantage that it is statistically more efficient and does not require a bootstrap procedure to correct standard errors.

It is possible to test endogeneity relying on similar ideas with the standard Durbin-Wu-Hausman test for endogeneity. This is done by testing joint significance of the components of η term. If η is jointly significant, this would indicate that there is endogeneity in our model. If η is not jointly significant, then the correction term is not necessary and efficiency can be estimated by

traditional SFMs.

3. Application to the Japanese Cotton Spinning Industry

3.1. Data

Our main panel dataset is borrowed from Braguinsky et al. (2015). The dataset consists of annual plant-level cotton yarn production records of 134 plants over 1896-1920 gathered by Japanese prefectural governments.[4] Note that the dataset is historical, and over a century, Japan changes tremendously and becomes one of the leading industrialized countries in the world. However, besides its historical importance, the advantage of using this dataset is that it has excellent details to illustrate consequences of ignoring endogeneity of concentration measures in a stochastic frontier model. Also, we supplement this dataset with original data that we collect from the Geospatial Information Authority of Japan. Hence, our analysis is enriched with new data. We consider a production technology with one output and three inputs. The output is the quality-adjusted amount of cotton yarns produced (Y). The inputs are: gender-adjusted labor (L), number of installed spindles as capital (K), and materials (M). We also integrate the age of plants (AGE) in the model.

3.2. Empirical Model and Results

In order to examine the relationship between the technical efficiency of plants and market competitiveness, we construct a year and prefecture specific Herfindahl-Hirschman Index (HHI) of market concentration. We expect that competition improves technical efficiency of the plants. Since HHI is potentially endogenous, as an instrumental variable, we create and use a count of

[4] For more details about the dataset, see Braguinsky et al. (2015).

mountains by prefectures that are higher than the average height of all Japanese mountains (1,698 meters or 5,571 feet).[5] Researchers such as Hoxby (2000) use similar topographical measures effectively as instrumental variables. In our case, the F-statistic of our topographical measure in the prediction equation of HHI is 632.23 which is substantially greater than 10 and passes the rule-of-thumb test for not being a weak IV. In order to check the sensitivity of our regression results to the selection of the IV, we also try using a count of all mountains by prefectures as an IV.

We estimate a translog production function with three inputs: L, K, and M. In the estimations, all inputs are demeaned. The estimation results are presented in Table 1. Model EX represents the model that ignores endogeneity, and Model EN represents the model that uses our methodology to handle endogeneity. Evaluated at the mean values of variables, we cannot reject that the production function has constant returns to scale at any conventional level. The η endogeneity test indicates that HHI is endogenous. We find that HHI has a positive and significant effect on inefficiency, which agrees with quiet life hypothesis, and this effect is larger when its endogeneity is handled. This finding supports some other studies from the literature such as Karakaplan and Kutlu (2017b), which shows that HHI is an endogenous inefficiency variable. The point estimates for mean and median efficiencies of cotton spinning plants under Model EN are somewhat less than their mean and median efficiencies under Model EX. The first moment of efficiency distribution may provide a useful comparison between different efficiency measures, but having similar mean efficiency scores does not necessarily imply that the efficiency estimates from different estimators are similar. Hence, we compare the distribution of efficiency estimates from Model EX and Model EN using a test for equality of distribution. In particular, a

[5] This measure is based on raw data collected from the website of Geospatial Information Authority of Japan at http://www.gsi.go.jp/

Kolmogorov-Smirnov test shows that the distributions of technical efficiencies in Model EX and Model EN are significantly different at 0.01% level. Finally, we find that using a count of all mountains by prefectures as an IV generates quantitatively similar outcomes to the results in Table 1.

[Table-1]

To give some examples, Osaka Prefecture has the most competitive cotton spinning industry in the data with an average HHI of 0.075. We find that the average efficiency of the plants in Osaka Prefecture is about 3 percentage points less in Model EX than that in Model EN. Shizuoka Prefecture, on the other hand, has the most concentrated cotton spinning industry in the data with an average HHI if 0.938. Our results show that the average efficiency of the plants in Shizuoka Prefecture is 5 percentage points more in Model EX than that in Model EN.

Kinugawa (1964) reports examples of inefficient cotton spinning plants from the 1890s era. One particular example, which is also narrated in Braguinsky et al. (2015), is Onagigawa Menpu cotton spinning plant in Tokyo, where workers smoked, used portable charcoal heaters in the plant, cooked and ate on the floor, and gambled in the inventory room, while raw cotton and other flammables were all over the place, and managerial staff were out fishing. The cotton spinning industry in Tokyo Prefecture was relatively concentrated (HHI = 0.430) during that period, and the company's efficiency in 1898 is 73% according to Model EN while it is 77% in Model EX. Similarly, Kyushu cotton plant in Kumamoto Prefecture (HHI = 1) is 63% efficient in Model EN while it is 83% in Model EX.

Figure 1 plots the linear relationship between HHI and the efficiencies in Model EX and

Model EN. That is, in the figure, predicted efficiency values from the sample are regressed on the constant and HHI. Both lines are downward sloping indicating that higher market concentration leads to higher inefficiency. The difference between predicted efficiencies of Model EX and Model EN is increasing over HHI and positive in relatively concentrated markets (HHI > 0.2). This difference reaches to 14 percentage points at HHI = 1. Hence, plants in relatively more concentrated markets would appear to be substantially more efficient in Model EX than they are according to Model EN. This result has serious policy implications as in general, the most central firms for policy-makers are exactly those with higher market power. The efficiency estimates for these firms are even more contaminated/biased in our data if the endogeneity is not handled. While being cautious not to overly extrapolate our results, we also note that our findings may signal potential dangers of ignoring endogeneity in other sectors and frameworks.

[Figure-1]

4. Conclusion

We presented a maximum likelihood based panel stochastic frontier model that can handle and test the endogeneity problem in stochastic frontier estimation. This method allows for endogeneity of both frontier and efficiency variables. One of the advantages of this method is that it is a single stage method. Hence, unlike two-stage methods, our method doesn't need a bootstrap correction for the standard errors. Moreover, only one prediction equation, i.e., instrument, is needed for an endogenous variable and its functions. This is particularly useful in translog settings where an endogenous variable and its cross products are involved in estimation.

We applied our panel stochastic frontier model to Japanese cotton spinning industry and

estimated the firms' technical efficiencies. The distribution of efficiency is assumed to be a function of market concentration measured by HHI. We considered two models: The first model assumes that the market concentration is exogenous, and the second model assumes that the market concentration is endogenous. Our test results indicated that the market concentration is endogenously determined with production, and thus endogeneity should be handled or otherwise the parameter efficiency estimates would be inconsistent. The average of efficiency estimates from these models are 76.80% and 75.25%, which are close. However, a closer look into the efficiency estimates indicated that in concentrated markets the efficiency values may differ substantially. Since concentrated markets take more attention by policy-makers due to market power considerations, our result is particularly important. We consider this outcome as a warning for those models that ignore endogeneity in stochastic frontier models.

References

Amsler, C., A. Prokhorov, P. Schmidt, 2016. Endogenous Stochastic Frontier Models. Journal of Econometrics 190, 280-288.

Braguinsky, S., A. Ohyama, T. Okazaki, C. Syverson, 2015. Acquisitions, Productivity, and Profitability: Evidence from the Japanese Cotton Spinning Industry. American Economic Review 105, 2086-2119.

Hicks, J.R., 1935. Annual Survey of Economic Theory: The Theory of Monopoly. Econometrica 3, 1-20.

Hoxby, C.M., 2000. Does Competition among Public Schools Benefit Students and Taxpayers? American Economic Review 90, 1209-1238.

Karakaplan, M.U., L. Kutlu, 2017a. Handling Endogeneity in Stochastic Frontier Analysis. Economics Bulletin 37, 889-891.

Karakaplan, M.U., L. Kutlu, 2017b. School District Consolidation Policies: Endogenous Cost Inefficiency and Saving Reversals, Unpublished manuscript.

Kinugawa, T., 1964. Hompo Menshi Boseki Shi (History of the Japanese Cotton Spinning Industry) in 7 Volumes. Tokyo: Hara Shobo.

Kutlu, L., 2010. Battese-Coelli Estimator with Endogenous Regressors. Economics Letters 109, 79-81.

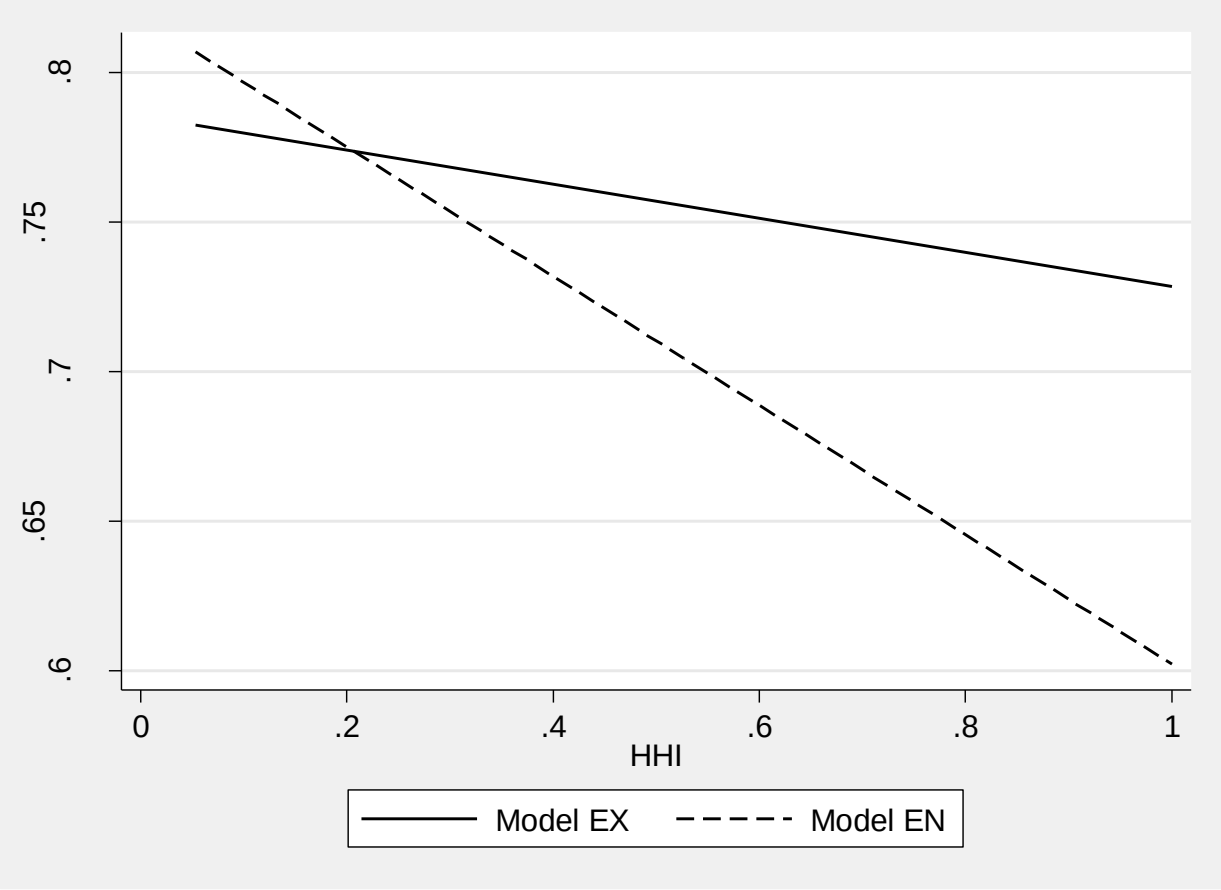
Tran, K.C., E.G. Tsionas, 2013. GMM Estimation of Stochastic Frontier Model with Endogenous Regressors. Economics Letters 118, 233-236.

Table 1. Estimation Results

Dependent variable: ln(Y)	Model EX		Model EN	
Constant	13.098***	(0.035)	13.125***	(0.036)
ln(L)	0.091**	(0.032)	0.081*	(0.034)
ln(K)	0.304***	(0.038)	0.282***	(0.040)
ln(M)	0.625***	(0.031)	0.627***	(0.032)
$0.5 \times \ln(L)^2$	-0.089*	(0.042)	-0.092*	(0.045)
$0.5 \times \ln(K)^2$	0.447***	(0.059)	0.456***	(0.061)
$0.5 \times \ln(M)^2$	0.279***	(0.026)	0.260***	(0.028)
$\ln(L) \times \ln(K)$	0.059	(0.039)	0.052	(0.041)
$\ln(L) \times \ln(M)$	0.027	(0.031)	0.054	(0.034)
$\ln(K) \times \ln(M)$	-0.413***	(0.037)	-0.429***	(0.039)
t	0.013***	(0.003)	0.014***	(0.003)
t^2	-0.001**	(0.000)	-0.001***	(0.000)
$\ln(L) \times t$	-0.002	(0.002)	-0.001	(0.002)
$\ln(K) \times t$	0.007***	(0.002)	0.008***	(0.002)
$\ln(M) \times t$	-0.003*	(0.001)	-0.004*	(0.002)
ln(AGE)	0.016	(0.011)	0.015	(0.011)
Dependent variable: $\ln(\sigma_u^2)$				
Constant	-2.850***	(0.277)	-3.053***	(0.252)
HHI	2.197***	(0.363)	2.918***	(0.306)
Dependent variable: $\ln(\sigma_v^2)$				
Constant	-3.497***	(0.033)		
Dependent variable: $\ln(\sigma_w^2)$				
Constant			-3.514***	(0.033)
η			0.360***	(0.074)
η endogeneity test (χ^2 = 23.7)			$P > \chi^2$ = 0.000	
Observations	2,049		2,049	
Log Likelihood	496.08		716.80	
Mean Technical Efficiency	0.7680		0.7525	
Median Technical Efficiency	0.7501		0.7476	

Notes: Standard errors are in parentheses. Asterisks indicate significance at the 0.1% (***), 1% (**) and 5% (*) levels. All inputs are demeaned.

Figure 1. Linear Relationship between HHI and Efficiencies in Model EX and Model EN



Notes: The lines represent the regression results of predicted efficiency values from Model EX and Model EN regressed on the constant and HHI.