Economics Bulletin

Volume 37, Issue 2

Handling Endogeneity in Stochastic Frontier Analysis

Mustafa U. Karakaplan
Georgetown University

Levent Kutlu
Georgia Institute of Technology

Abstract

We present a general maximum likelihood based framework to handle the endogeneity problem in the stochastic frontier models. We implement Monte Carlo experiments to analyze the performance of our estimator. Our findings show that our estimator outperforms standard estimators that ignore endogeneity.

Citation: Mustafa U. Karakaplan and Levent Kutlu, (2017) "Handling Endogeneity in Stochastic Frontier Analysis", *Economics Bulletin*, Volume 37, Issue 2, pages 889-901
Contact: Mustafa U. Karakaplan - mukarakaplan@yahoo.com, Levent Kutlu - levent.kutlu@gatech.edu.
Submitted: August 02, 2016. **Published:** May 01, 2017.

1. Introduction

Endogeneity problems can arise in stochastic frontier models due to a couple of major reasons: First, the determinants of the cost frontier and the two-sided error term can be correlated. Secondly, the inefficiency term and two-sided error term can be correlated, or in particular, the determinants of the inefficiency can cause this correlation. Endogeneity in a stochastic frontier model would lead to inconsistent parameter estimates, and hence, it would need to be addressed properly.

In the empirical literature, there is a growing concern about the endogeneity issues in the stochastic frontier models. For example, maximum likelihood estimation is probably the most widely used method in the stochastic frontier literature, but conventional maximum likelihood estimation of an endogenous stochastic frontier model would give inconsistent parameter estimates. This would necessitate a proper instrumental variable (IV) approach in order to deal with the endogeneity issue. In the maximum likelihood framework, a standard way to deal with this problem is modeling the joint distribution of the dependent variable and endogenous variables; and then maximizing the corresponding log-likelihood of this distribution. However, due to the special nature of the error term in the stochastic frontier models, this is a relatively more difficult task compared to the standard maximum likelihood models involving only two-sided error terms.

Guan et al. (2009) follow a two-step estimation methodology to handle the endogenous frontier regressors. In the first step of their methodology, they get the consistent estimates of the frontier parameters using GMM, and in the second step, they use the residuals from the first step as the dependent variable to get the maximum likelihood stochastic frontier estimates. Since the second step of this procedure uses the standard stochastic frontier estimators, the efficiency estimates would not be consistent when the two-sided and one-sided error terms are correlated. Kutlu (2010) makes an effort to address the endogeneity problem in the maximum likelihood estimation context. He describes a model that aims to solve the endogeneity problem due to the correlation between the regressors and two-sided error term. Tran and Tsionas (2013) propose a GMM variation of Kutlu (2010). The assumptions of these models are not sufficient for handling the endogeneity due to one-sided and two-sided error terms. Mutter et al. (2013) explain of why omitting the variable causing the endogeneity is not a viable solution. Shee and Stefanou (2015) extends the methodological approach in Levinsohn and Petrin (2003) to overcome the problem of endogenous input choice due to production shocks that are predictable by the productive unit but unknown to the econometrician. Unlike our study, however, Shee and Stefanou (2015) do not consider the endogeneity problem due to the correlation of one-sided error term and two-sided error term. Gronberg et al. (2015) try to solve the problem through pseudo-IV methodologies.

Amsler et al. (2016) propose a copula approach that allows more general correlation structures when modeling endogeneity. However, this method is computationally intensive and requires choosing a proper copula. Moreover, the model presented in Amsler et al. (2016) does not allow environmental variables that affect inefficiency, which makes it less applicable when trying to understand the factors that affect inefficiency. Griffiths and Hajargasht (2016) present a Bayesian stochastic frontier model, which allows environmental variables but their model is very different from ours.[1] Overall, one of the main strengths of our model is that it is easier to apply

[1] Amsler et al. (2016), Griffiths and Hajargasht (2016), and Tran and Tsionas (2015) are papers with alternative econometric approaches that are contemporary with a previous version of our very paper and the econometric methodology presented here. In fact, these three papers did not exist when we originally finished and submitted our first draft, and they do cite our working papers and methods.

compared to its copula or Bayesian counterparts, and our model is a direct generalization of one of the most widely used stochastic frontier models, i.e. Battese and Coelli (1995) type estimators.

2. A Practical Econometric Approach to Handle Endogeneity

We consider the following stochastic frontier model with endogenous explanatory variables:

$$y_i = x_{1i}'\beta + v_i - su_i \tag{1}$$
$$x_i = Z_i\delta + \varepsilon_i$$
$$\begin{bmatrix} \tilde{\varepsilon}_i \\ v_i \end{bmatrix} \equiv \begin{bmatrix} \Omega^{-1/2}\varepsilon_i \\ v_i \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_p & \sigma_{vi}\rho \\ \sigma_{vi}\rho' & \sigma_{vi}^2 \end{bmatrix} \right)$$
$$s = -1 \text{ for cost functions} \quad (\text{or } s = 1 \text{ for production functions})$$

where y_i is the logarithm of the expenditure (or output) of the i^{th} unit; x_{1i} is a vector of exogenous and endogenous variables; x_i is a $p\times1$ vector of all endogenous variables (excluding y_i), $Z_i = I_p \otimes z_i'$ where z_i is a $q\times1$ vector of all exogenous variables, v_i and ε_i are two-sided error terms, and $u_i \geq 0$ is a one-sided error term capturing the inefficiency. In our framework, a variable is endogenous if it is not independent from v_i. Finally, Ω is the variance-covariance matrix of ε_i, σ_{vi}^2 is the variance of v_i, and ρ is the vector representing the correlation between $\tilde{\varepsilon}_i$ and v_i.

The applicability and implications of our model is much more comprehensive than that of Kutlu (2010) who proposes a model that enables estimation of efficiency when some of the regressors are correlated with the v_i term.[2] He does not provide a solution for a potential correlation between v_i and u_i terms. In particular, the assumptions of his model do not assure consistency of parameter estimates when v_i and u_i terms are correlated, and hence, he does not mention the case. Indeed, his model does not consider heteroskedasticity in either component of the composed error term. On the other hand, our model specifications provide a methodology to deal with the endogeneity issues in stochastic frontier models in a more general setting.

The assumption that v_i and u_i are independent is dominantly made in the stochastic frontier literature. We address this issue by allowing v_i and u_i to be dependent through observables that shape both distributions. Let x_{2i} be a vector of exogenous and endogenous variables. We assume that the inefficiency term, u_i, is a function of x_{2i} and an observation unit specific random component, u_i^*. More precisely,

$$u_i = \sigma_u(x_{2i};\varphi_u)u_i^* \tag{2}$$

where $\sigma_{ui} = \sigma_u(x_{2i};\varphi_u) > 0$ and $u_i^* \geq 0$ is independent from v_i and ε_i conditional on x_i and z_i. Hence, u_i is not independent from x_i, yet u_i and v_i are conditionally independent given x_i and z_i. Similarly, u_i and ε_i are conditionally independent given x_i and z_i. Our view is that if the model is well-specified in the sense that it includes proper variables that affect efficiency, then the conditional correlation of u_i^* and v_i can be eliminated (at least in most realistic scenarios). Hence, in practice, most of the time this is not an issue unless there are omitted variables when modelling inefficiency.

By a Cholesky decomposition of the variance-covariance matrix of $(\tilde{\varepsilon}_i', v_i)'$, we can represent $(\tilde{\varepsilon}_i', v_i)'$ as follows:

[2] Also see Kutlu and Sickles (2012) for similar ideas in the Kalman filter framework to measure market powers of firms.

$$\begin{bmatrix} \tilde{\varepsilon}_i \\ v_i \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ \sigma_{vi}\rho' & \sigma_{vi}\sqrt{1-\rho'\rho} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_i \\ \widetilde{w}_i \end{bmatrix} \quad (3)$$

where $\tilde{\varepsilon}_i$ and $\widetilde{w}_i \sim \mathbf{N}(0,1)$ are independent. Hence, we can write the frontier equation as follows:

$$\begin{aligned} y_i &= x_{1i}'\beta + \sigma_{vi}\rho'\tilde{\varepsilon}_i + w_i - su_i \quad (4) \\ &= x_{1i}'\beta + \frac{\sigma_{wi}}{\sigma_{cw}}\eta'(x_i - Z_i\delta) + e_i \end{aligned}$$

where $e_i = w_i - su_i$ $w_i = \sigma_{vi}\sqrt{1-\rho'\rho}\widetilde{w}_i = \sigma_{wi}\widetilde{w}_i$, $\sigma_{wi} = \sigma_{cw}\sigma_w(.;\varphi_w)$ is separable so that $\sigma_{cw} > 0$ is a function of the constant term, $\sigma_w(.;\varphi_w)$ is a function of all variables affecting σ_{wi} except the constant term so that $\sigma_w(.;\varphi_w) = 1$ when $\varphi_w = 0$, and $\eta = \sigma_{cw}\Omega^{-\frac{1}{2}}\rho/\sqrt{1-\rho'\rho}$. For example, if $\sigma_{wi}^2 = \exp(x_{3i}'\varphi_w)$, then $\sigma_{cw} = \exp(\varphi_{cw})$ where φ_{cw} is the constant term in $x_{3i}'\varphi_w$. Hence, when there is no heteroskedasticity in w_i, we have $\sigma_{wi} = \sigma_{cw}$ so that:

$$y_i = x_{1i}'\beta + \eta'(x_i - Z_i\delta) + e_i. \quad (5)$$

Note that e_i is conditionally independent from the regressors given x_i and z_i. Hence, conditional on x_i and z_i, the distribution of e_i and $(u_i|e_i)$ are exactly the same as their traditional counterparts from the stochastic frontier literature. We can also directly assume that the conditional distribution of v_i given x_i (and exogenous variables) is a normal distribution with mean equal to $(\sigma_{wi}/\sigma_{cw})\,\eta'(x_i - Z_i\delta)$. Hence, rather than assuming that $(\tilde{\varepsilon}_i', v_i)'$ is jointly normally distributed and using this to derive the conditional distribution of v_i, we can directly assume that v_i is normally distributed with mean $(\sigma_{wi}/\sigma_{cw})\,\eta'(x_i - Z_i\delta)$ given x_i (and exogenous variables). This approach is commonly used to solve the endogeneity problem in models with intrinsic non-linearity such as choice models.[3] According to this approach $(\sigma_{wi}/\sigma_{cw})\,\eta'(x_i - Z_i\delta)$ is a correction term for bias. Hence, this approach treats endogeneity as an omitted variable problem. In what follows, we base our analysis on this assumption. We assume that:[4]

$$\begin{aligned} &u_i^* \sim \mathbf{N}^+(0,1) \quad (6) \\ &\sigma_{ui}^2 = \exp(x_{2i}'\varphi_u) \\ &\sigma_{wi}^2 = \exp(x_{3i}'\varphi_w). \end{aligned}$$

where $\varphi = (\varphi_u', \varphi_w')'$ is the vector of parameters capturing heteroskedasticity and x_{3i} is a vector of exogenous and endogenous variables which can share the same variables with x_{1i} and x_{2i}. Here, $\sigma_{cw}^2 = \exp(\varphi_{cw})$ where φ_{cw} is the coefficient of constant term for $x_{3i}'\varphi_w$. This implies that $u_i \sim \mathbf{N}^+(0, \sigma_{ui}^2)$.[5] Note that $Cov(u_i, \varepsilon_i) = \sqrt{2/\pi}\,Cov(\sigma_{ui}, \varepsilon_i) \neq 0$ in general. This is one of the important features of our model. The conventional stochastic frontier models do not allow such correlations. Let $\lambda_i = \sigma_{ui}/\sigma_{wi}$ and $\sigma_i^2 = \sigma_{wi}^2 + \sigma_{ui}^2$. Then, the probability density function of e_i is given by:

$$f_e(e_i) = \frac{2}{\sigma_i}\phi\left(\frac{e_i}{\sigma_i}\right)\Phi\left(\frac{-s\lambda_i e_i}{\sigma_i}\right) \quad (7)$$

where ϕ and Φ denote the standard normal PDF and CDF, respectively. Let $y = (y_1, y_2, \ldots, y_n)'$

---

[3] For more details about this approach, see Wooldridge (2010). Also see Terza et al. (2008) for two-stage residual inclusion methods. Unlike Terza et al. (2008), our estimations are done in a single stage and deal with additional complications of stochastic frontier models, which involve composed error terms.

[4] These particular choices of half-normal distribution and exponential function are not essential for our analysis. For illustrative purposes, we chose one of the distributions that is applied relatively more commonly in the empirical studies.

[5] Note that $Var(u_i^*) = (\pi - 2)/\pi$ and $Var(u_i) = (\pi - 2)\sigma_{ui}^2/\pi$.

be the vector of dependent variable, $x = (x_1', x_2', \ldots, x_n')'$ be a matrix of endogenous variables in the model (i.e, the elements of x are the x_i's defined earlier), and $\theta = (\beta', \eta', \varphi', \delta')'$. The log-likelihood of (y, x) is given by:[6]

$$\ln L(\theta) = \ln L_{y|x}(\theta) + \ln L_x(\theta) \tag{8}$$

where

$$\begin{aligned}
\ln L_{y|x}(\theta) &= \sum_{i=1}^{n} \left(\ln 2 - \frac{1}{2}\ln\sigma_i^2 + \ln\phi\left(\frac{e_i}{\sigma_i}\right) + \ln\Phi\left(\frac{-s\lambda_i e_i}{\sigma_i}\right)\right) \\
&= \sum_{i=1}^{n} \left(\frac{\ln(2/\pi) - \ln\sigma_i^2 - (e_i^2/\sigma_i^2)}{2} + \ln\Phi\left(\frac{-s\lambda_i e_i}{\sigma_i}\right)\right) \\
\ln L_x(\theta) &= \sum_{i=1}^{n} \left(\frac{-p\cdot \ln 2\pi - \ln(|\Omega|) - \varepsilon_i' \, \Omega^{-1}\varepsilon_i}{2}\right) \\
e_i &= y_i - x_{1i}'\beta - \frac{\sigma_{wi}}{\sigma_{cw}}\eta'(x_i - Z_i\delta) \\
\varepsilon_i &= x_i - Z_i\delta \\
\sigma_i^2 &= \sigma_{wi}^2 + \sigma_{ui}^2 \\
\lambda_i &= \frac{\sigma_{ui}}{\sigma_{wi}}.
\end{aligned}$$

Even though u_i and v_i are not independent unconditionally, they are conditionally independent. Hence, this decomposition enables us to use the usual density function for the $\ln L_{y|x}(\theta)$ part of the log-likelihood function. As can be seen, this part of the log-likelihood function is almost the same as that of a traditional stochastic frontier model. However, we also add $\ln L_x$ to the log-likelihood and adjust the e_i term by the $(\sigma_{wi}/\sigma_{cw})\,\eta'(x_i - Z_i\delta)$ factor.[7] It is worth mentioning that the inclusion of the bias correction term solves the problem of inconsistent parameter estimates due to endogenous regressors in x_{1i} and due to the endogenous variables in x_{2i}. The efficiency, $EFF_i = \exp(-u_i)$, can be predicted by:

$$E[\exp(-su_i)|e_i]^s = \left(\frac{1 - \Phi(s\sigma_i^* - \mu_i^*/\sigma_i^*)}{1 - \Phi(-\mu_i^*/\sigma_i^*)} \exp\left(-s\mu_i^* + \frac{1}{2}\sigma_i^{*2}\right)\right)^s \tag{9}$$

where

$$\begin{aligned}
\mu_i^* &= \frac{-se_i\sigma_{ui}^2}{\sigma_i^2} \\
\sigma_i^{*2} &= \frac{\sigma_{wi}^2\sigma_{ui}^2}{\sigma_i^2}.
\end{aligned}$$

For computationally difficult cases, one can use a two-step maximum likelihood estimation method as in Murphy and Topel (1985).[8] In the first stage, $\ln L_x(\theta)$ is maximized with respect to the relevant parameters. In the second stage, conditional on the parameters estimated in the first

[6] For the notational simplicity, we drop the exogenous variables from the conditional density function.

[7] This approach is applicable to various maximum likelihood estimation based stochastic frontier models widely used by researchers. For example, u_i can be assumed to have a truncated normal, exponential, or gamma distribution among other distributions.

[8] The two-stage method suggested in here is different than the one that is criticized by Wang and Schmidt (2002) or the one implemented by Kutlu (2010), which requires bootstrapping. Hence, our suggestion is not subject to their criticisms.

stage, $\ln L_{y|x}(\theta)$ is maximized. In our case, the conditional second stage becomes:

$$y_i = x_{1i}'\beta + \sigma_{wi}\rho'\hat{\tilde{\varepsilon}}_i + w_i - su_i \tag{10}$$

where $\hat{\tilde{\varepsilon}}_i$ is the first stage estimate of $\tilde{\varepsilon}_i$. A simpler approach would be estimating each component of $\tilde{\varepsilon}_i$ by OLS in the first stage using the equation $\varepsilon_i = x_i - Z_i\delta$; and estimating (10) by maximum likelihood estimation method. Since the second stage uses the estimate of $\tilde{\varepsilon}_i$ instead of the variable itself, the asymptotic variance matrix should be adjusted for the second stage. Based on Murphy and Topel (1985), Greene (2008) gives a concise presentation of this two-step maximum likelihood estimation method.[9] Hence, by applying the two-step maximum likelihood estimation method, it is possible to deal with some of the computational difficulties.

2.1. Endogeneity Test

In addition to providing a way to solve the endogeneity problem, we also offer a method to test the endogeneity. For this purpose, we propose testing the joint significance of the components of the η term. If the components are jointly significant, then we would conclude that there is endogeneity in our model. When the components are not jointly significant, this would indicate that the correction term is not necessary and the efficiency can be estimated by the traditional frontier models. The significance of the k^{th} component of η indicates that x_{ik} (the k^{th} component of x_i) and v_i are correlated. Hence, a particular variable of interest is endogenous if the corresponding component of η term is significant. Essentially, our endogeneity test relies on ideas similar to the standard Durbin-Wu-Hausman test for endogeneity. Finally, note that when $\eta = 0$, the standard errors from the second stage of the two-step estimator are valid. Moreover, asymptotically, they are as efficient as the one-step version. Hence, the F-test can be applied to test the endogeneity of relevant variables by testing the joint significance of the components of η. Our model is a particularly attractive choice as it enables us to test the endogeneity of the inefficiency term, u_i.

2.2. Monte Carlo Simulations

We implement Monte Carlo simulations in order to examine the small sample performance of our estimator. We consider a Cobb-Douglas cost function model and assume that the variance term for the one-sided error, u_i, is heteroskedastic and is a function of a variable, which can be correlated with the two-sided error term, v_i. This represents the case in which the variables explaining the efficiency are simultaneously determined with cost. Until recently, the literature largely ignored the possibility of a correlation between u_i and v_i. In contrast to what is done in practice, such a correlation is likely to be more frequent than rare. We analyze both the consequences of ignoring such a correlation and the performance of our estimator in dealing with this problem.

We examine four simulation scenarios: In Scenario 1, we analyze a model in which one of the regressors is correlated with v. In Scenario 2, we analyze a model in which u is correlated with v. In Scenario 3, we analyze a model in which one of the regressors and one of the environmental variables for u are correlated with v. Finally, in Scenario 4, we analyze a model in which one of the regressors in the frontier, one of the environmental variables for u, and u^* are correlated with

[9] Hardin (2002) explains how estimation of the two-stage maximum likelihood models with robust variance can be implemented in Stata.

v. Unlike Scenario 3, Scenario 4 violates an important assumption for our model. Hence, when estimating this scenario, we estimate it as if it is Scenario 3. The data generating process (DGP) for these four scenarios are described in Appendix. Table I and Table II present the simulation results of these four scenarios with both strong IVs and weak IVs.

Table I: Simulation Results with Strong Instruments

$\rho = 0$ and $\delta = 1$

		Scenario 1		Scenario 2		Scenario 3	
	True Values	Model EX	Model EN	Model EX	Model EN	Model EX	Model EN
c	0.5000	0.5025	0.5024	0.4997	0.4980	0.5021	0.5005
α	0.5000	0.5021	0.5016	0.5015	0.5018	0.5007	0.5007
β	0.5000	0.4995	0.5004	0.5012	0.5015	0.5004	0.5014
φ_{cu}	-1.2000	-1.2526	-1.2516	-1.2366	-1.2299	-1.2362	-1.2294
φ_u	1.4000	1.4214	1.4207	1.4118	1.4075	1.4123	1.4080
MSE c		0.0301	0.0306	0.0256	0.0259	0.0244	0.0254
MSE α		0.0069	0.0093	0.0088	0.0088	0.0061	0.0088
MSE β		0.0030	0.0093	0.0089	0.0089	0.0029	0.0089
MSE φ_{cu}		0.1746	0.1770	0.1306	0.1348	0.1270	0.1317
MSE φ_u		0.0572	0.0587	0.0316	0.0350	0.0305	0.0340
MSE u		0.1751	0.1758	0.1712	0.1723	0.1709	0.1727
Bias u		-0.0065	-0.0063	-0.0040	-0.0034	-0.0039	-0.0032
Pearson		0.7716	0.7710	0.8041	0.8034	0.8043	0.8032
Spearman		0.7650	0.7644	0.7920	0.7914	0.7923	0.7912

$\rho = 0.70$ and $\delta = 1$

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	True Values	Model EX	Model EN	Model EX	Model EN	Model EX	Model EN	Model EX	Model EN
c	0.5000	0.6239	0.5003	0.5868	0.4991	0.7671	0.4972	0.5951	0.3080
α	0.5000	0.2087	0.5025	0.4369	0.5007	0.1005	0.5022	0.1436	0.5018
β	0.5000	0.9702	0.4980	0.4373	0.5017	0.9464	0.4975	0.8896	0.4935
φ_{cu}	-1.2000	-1.0069	-1.2283	-1.8429	-1.2186	-1.7998	-1.2012	-1.0925	-0.4985
φ_u	1.4000	1.2156	1.4106	1.9473	1.4037	1.9029	1.3986	1.5893	1.1348
MSE c		0.0372	0.0233	0.0294	0.0205	0.0882	0.0125	0.0291	0.0489
MSE α		0.0903	0.0083	0.0123	0.0078	0.1643	0.0054	0.1318	0.0053
MSE β		0.2235	0.0084	0.0123	0.0080	0.2015	0.0055	0.1539	0.0057
MSE φ_{cu}		0.1375	0.0922	0.4794	0.0761	0.4175	0.0175	0.0693	0.5079
MSE φ_u		0.0732	0.0381	0.3166	0.0236	0.2687	0.0090	0.0524	0.0790
MSE u		0.1341	0.1083	0.2033	0.1060	0.1507	0.0071	0.1963	0.0644
Bias u		-0.0116	-0.0036	0.1019	-0.0024	0.0855	0.0009	-0.2845	-0.1962
Pearson		0.8177	0.8462	0.7915	0.8656	0.8328	0.9880	0.8984	0.9851
Spearman		0.8139	0.8437	0.7863	0.8574	0.8270	0.9880	0.8998	0.9892

Table II: Simulation Results with Weak Instruments

$\rho = 0$ and $\delta = 0.25$

	True Values	Scenario 1 Model EX	Scenario 1 Model EN	Scenario 2 Model EX	Scenario 2 Model EN	Scenario 3 Model EX	Scenario 3 Model EN
c	0.5000	0.5021	0.4987	0.5065	0.5024	0.5084	0.5043
α	0.5000	0.5021	0.5053	0.5011	0.5016	0.5000	0.5004
β	0.5000	0.4991	0.4848	0.5001	0.5007	0.5002	0.5040
φ_{cu}	-1.2000	-1.2505	-1.2491	-1.3080	-1.2975	-1.3026	-1.2891
φ_u	1.4000	1.4199	1.4190	1.4681	1.4440	1.4658	1.4401
MSE c		0.0299	0.0807	0.0229	0.0235	0.0214	0.0358
MSE α		0.0058	0.0654	0.0061	0.0062	0.0032	0.0224
MSE β		0.0043	1.4886	0.0061	0.0062	0.0029	0.4680
MSE φ_{cu}		0.1728	0.1780	0.3057	0.6845	0.2893	0.3679
MSE φ_u		0.0564	0.0597	0.1519	0.2704	0.1439	0.1994
MSE u		0.1751	0.1759	0.1162	0.1181	0.1160	0.1182
Bias u		-0.0064	-0.0060	-0.0089	-0.0064	-0.0084	-0.0059
Pearson		0.7716	0.7708	0.6834	0.6808	0.6839	0.6809
Spearman		0.7650	0.7642	0.6206	0.6183	0.6213	0.6185

$\rho = 0.70$ and $\delta = 0.25$

	True Values	Scenario 1 Model EX	Scenario 1 Model EN	Scenario 2 Model EX	Scenario 2 Model EN	Scenario 3 Model EX	Scenario 3 Model EN	Scenario 4 Model EX	Scenario 4 Model EN
c	0.5000	0.5484	0.4955	0.4861	0.5018	0.5696	0.4936	0.4542	0.3052
α	0.5000	0.3944	0.5107	0.4701	0.5006	0.3337	0.5104	0.3522	0.5055
β	0.5000	1.1784	0.4453	0.4694	0.5007	1.1750	0.4432	1.0913	0.4624
φ_{cu}	-1.2000	-1.1476	-1.2277	-1.7192	-1.2431	-1.9270	-1.2052	-1.2040	-0.4502
φ_u	1.4000	1.3397	1.4102	2.8373	1.4188	3.1159	1.4005	2.4346	1.0184
MSE c		0.0207	0.0439	0.0139	0.0171	0.0126	0.0175	0.0127	0.0572
MSE α		0.0151	0.0376	0.0061	0.0058	0.0293	0.0162	0.0239	0.0134
MSE β		0.4632	1.0179	0.0062	0.0058	0.4571	0.4899	0.3511	0.2257
MSE φ_{cu}		0.0882	0.0929	0.3077	0.1122	0.5552	0.0091	0.0396	0.5693
MSE φ_u		0.0383	0.0386	2.0909	0.0924	2.9641	0.0185	1.1004	0.1625
MSE u		0.1101	0.1084	0.1706	0.0790	0.1262	0.0062	-0.2292	-0.1947
Bias u		-0.0070	-0.0035	0.1042	-0.0038	0.0943	-0.0002	0.1475	0.0622
Pearson		0.8461	0.8461	0.6629	0.7700	0.7393	0.9778	0.8506	0.9780
Spearman		0.8418	0.8436	0.6056	0.7147	0.6834	0.9708	0.8147	0.9806

We refer to the model that ignores endogeneity as Model EX, and our model that captures endogeneity as Model EN and present the means and mean square errors of the frontier parameters (c, α, and β) and variance parameters for σ_{ui}^2(φ_{cu} and φ_u).[10] Moreover, mean square errors for the efficiency estimates, and Pearson and Spearman correlations of efficiency estimates with the true efficiency are presented.

In the benchmark case ($\rho = 0$ and $\delta = 1$) of Scenario 1, simulation results indicate that the parameter estimates and corresponding mean square errors for Model EX and Model EN are similar. Moreover, Pearson and Spearman correlations are similar as well. Hence, Model EN performs well. However, when there is endogeneity ($\rho = 0.7$ and $\delta = 1$), frontier and variance parameter estimates for Model EX are severely biased. Model EN, on the other hand, outperforms Model EX in terms of mean squares and correlations, and parameter estimates seem to have no bias. As the extent of identification weakens ($\delta = 0.25$), the parameter estimates for Model EN start to have some bias. However, if endogeneity is present, it can still be beneficial to use the instrumental variables approach that we proposed as the bias can be lower. This is a common result of the instrumental variables methods and not specific to our methodology. Hence, the relative magnitudes of the biases for using Model EN and Model EX depend on the degree of endogeneity and identification problem.

As in Scenario 1, the results from Scenario 2 show that the benchmark case performance of Model EN is similar to that of Model EX. However, when there is endogeneity ($\rho = 0.7$ and $\delta = 1$), Model EN dominates Model EX. For the frontier parameters, the biases are not as severe as that of Scenario 1 but they are still considerably high. Moreover, as expected, the variance parameters are severely biased. For the weak identification scenario, we did not observe serious biases when Model EN is used.

In Scenario 3, we have two variables, one in frontier and one in u, that are correlated with v. That is, noise term is not only correlated with one of the explanatory variables but also correlated with the inefficiency term ($\rho_1 = \rho_2 = 0.7$ and $\delta_1 = \delta_2 = 1$). Hence, among the first three scenarios that we examine, this scenario is the most problematic and yet the most probable scenario. In Scenario 3, Model EX has all the weaknesses from Scenario 1 and Scenario 2. The results from Scenario 3 show that Model EN outperforms Model EX and all other results in Scenario 3 are in line with the findings from the first two scenarios. All in all, these three simulations indicate that ignoring endogeneity in our model would have severe consequences.

In Scenario 4, the data generating process is the same as Scenario 3 except that u^* is correlated with v as well. This violates one of our assumptions. As a consequence, the constant term of the frontier is biased, yet other frontier parameters are reasonably close to their true values. The efficiency estimates are biased but still better than their exogenous counterparts in terms of bias and MSE for u as well as correlations. Finally, note that in many empirical scenarios, if the variables that determine the inefficiency are specified properly, it may be reasonable to assume that u^* and v are conditionally independent. Hence, although we presented these simulation results for the sake illustrating the consequences of violating one of our assumptions, we believe that in a well-defined model with no omitted environmental variables, our model is expected to perform well. In a panel data extension of our model, this situation would be even less likely since the fixed effects terms would eliminate or reduce the potential conditional correlation between u^* and v. In any case, if researchers suspect that the environmental variables that they include to identify

[10] We do not directly estimate the variance parameters for v_i term. That is why we do not present their estimates in our simulations.

efficiency are not sufficient to eliminate the conditional correlation, then they can also apply a model with a more general but more complicated correlation structure such as Griffiths and Hajargasht (2016).

3. Concluding Remarks

We introduced a maximum likelihood based methodology to handle the endogeneity problems in stochastic frontier models. In addition to that, we also presented a way to test the endogeneity. We carried out Monte Carlo simulations to analyze the small sample performance of our estimator in a variety of endogeneity scenarios; and we found that when there is endogeneity in the model, our estimator outperforms the model which assumes exogeneity.

4. References

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

Battese, G.E., Coelli, T.J. (1995) "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data" *Empirical Economics* **20**, 325-332.

Greene, W.H. (2008) *Econometric Analysis*, 6th ed, Prentice Hall: Englewood Cliffs, NJ.

Griffiths, W.E., Hajargasht, G. (2016) "Some Models for Stochastic Frontiers with Endogeneity" *Journal of Econometrics* **190**, 341-348.

Gronberg, T.J., Jansen, D.W., Karakaplan, M.U., Taylor, L.L. (2015) "School District Consolidation: Market Concentration and the Scale-Efficiency Tradeoff" *Southern Economic Journal* **82**, 580-597.

Guan, Z., Kumbhakar, S.C., Myers, R.J., Lansink, A.O. (2009) "Measuring Excess Capital Capacity in Agricultural Production" *American Journal of Agricultural Economics* **91**, 765-776.

Hardin, J.W. (2002) "The Robust Variance Estimator for Two-Stage Models" *The Stata Journal* **2**, 253-256.

Kumbhakar, S.C., Wang, H.-J. (2005) "Estimation of Growth Convergence Using a Stochastic Production Frontier Approach" *Economics Letters* **88**, 300-305.

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

Kutlu, L., Sickles, C.R. (2012) "Estimation of Market Power in the Presence of Firm Level Inefficiencies" *Journal of Econometrics* **168**, 141-155.

Levinsohn, J., Petrin, A. (2003) "Estimating Production Functions Using Inputs to Control for Unobservables" *The Review of Economic Studies* **70**, 317-341.

Murphy, K.M., Topel, R.H. (1985) "Estimation and Inference in Two-Step Econometric Models" *Journal of Business and Economic Statistics* **3**, 370-379.

Shee, A., Stefanou, S.E. (2015) "Endogeneity Corrected Stochastic Production Frontier and Technical Efficiency" *American Journal of Agricultural Economics* **97**, 939-952.

Terza, J.V., Basu, A., Rathouz, P.J. (2008) "Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling" *Journal of Health Economics* **27**, 531-543.

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

Tran, K.C., Tsionas, E.G. (2015) "Endogeneity in Stochastic Frontier Models: Copula Approach without External Instruments" *Economics Letters* **133**, 85-88.

Wang, H.-J., Schmidt, P. (2002) "One-Step and Two-Step Estimation of the Effects of Exogenous Variables on Technical Efficiency Levels" *Journal of Productivity Analysis* **18**, 129-144.

Wooldridge, J.M. (2010) *Econometric Analysis of Cross Section and Panel Data*, MIT press: Cambridge, MA.

Appendix: Data Generating Processes for Monte Carlo Simulations

For Scenario 1 and 2, without loss of generality, we assume that x_{3i} is the endogenous variable that is correlated with v_i.

$$y_i = c + \alpha x_{1i} + \beta x_{ji} + v_i - u_i \tag{11}$$
$$\begin{bmatrix} x_{1i} \\ x_{2i} \\ z_{3i} \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \Omega_x \right)$$
$$x_{3i} = \delta z_{3i} + \varepsilon_i$$
$$\varepsilon_i = \sigma_\varepsilon \tilde{\varepsilon}_i$$
$$\begin{bmatrix} \tilde{\varepsilon}_i \\ v_i \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_v\rho \\ \sigma_v\rho & \sigma_v^2 \end{bmatrix} \right)$$
$$u_i^* \sim \mathbf{N}^+(0,1)$$
$$u_i = \sigma_{ui} u_i^*$$
$$\sigma_{ui}^2 = \exp(\varphi_{cu} + \varphi_u x_{ki})$$

where $(\mathrm{j}, \mathrm{k}) = (2,3)$ or $(\mathrm{j}, \mathrm{k}) = (3,2)$.

In the base scenario of their simulations, Kumbhakar and Wang (2005) pick $\sigma_{ui}^2/\sigma_v^2 = 4.2$ and $\sigma_{ui}^2 + \sigma_v^2 = 0.15$. The variance ratio of 4.2 indicates that the variance of cost efficiency is about 1.5 times the variance of the noise term. In our simulations, we choose $c = \alpha = \beta = 0.5$, $\mu_1 = 2, \mu_2 = \mu_3 = 1,\ \sigma_v^2 = 0.3, \varphi_{cu} = -1.2, \varphi_u = 1.4$, and $\mathrm{E}[x_{ki}] = 1$ (i.e., $\mu_2 = \mu_3 = 1$). This indicates that, evaluated at the mean of x_{ki}, we have $\sigma_{ui}^2/\sigma_v^2 \cong 4.071$. We consider two different values for ρ. In particular, $\rho = 0$ represents the case where there is no endogeneity and $\rho = 0.7$ represents the case where there is endogeneity. For both cases, we choose $\sigma_\varepsilon^2 = 0.3$ so that the variance of $[\varepsilon_i \quad v_i]'$ is positive definite. Moreover, we consider two different values for δ. In particular, $\delta = 0.25$ represents the case where identification is relatively weak and $\delta = 1$ represents a case where identification is fairly good. Finally, we set:

$$\Omega_x = \begin{bmatrix} 0.3 & 0.21 & 0.21 \\ 0.21 & 0.3 & 0.21 \\ 0.21 & 0.21 & 0.3 \end{bmatrix} \tag{12}$$

The choice of Ω_x implies that the correlations between each pair from x_{1i}, x_{2i}, and z_{3i} are equal to 0.7. Moreover, Ω_x is positive definite as required. As a benchmark, we run the simulations for the same parameter values except that this time, ρ is set to be equal to zero and δ is set equal to 1. Hence, under the benchmark scenario, if the heteroskedasticity is controlled for, the parameter estimates would be consistent and there would not be a weak identification problem. Simulation experiments were repeated 25,000 times for a sample size of 500.

For Scenario 3, the DGP is given by:

$$y_i = c + \alpha x_{1i} + \beta x_{2i} + v_i - u_i \tag{13}$$
$$\begin{bmatrix} x_{1i} \\ z_{2i} \\ z_{3i} \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \Omega_x \right)$$
$$\begin{bmatrix} x_{2i} \\ x_{3i} \end{bmatrix} = \begin{bmatrix} \delta_2 z_{2i} \\ \delta_3 z_{3i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{2i} \\ \varepsilon_{3i} \end{bmatrix}$$
$$\begin{bmatrix} \tilde{\varepsilon}_i \\ v_i \end{bmatrix} \equiv \begin{bmatrix} \Omega^{-\frac{1}{2}} \begin{bmatrix} \varepsilon_{2i} \\ \varepsilon_{3i} \end{bmatrix} \\ v_i \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_2 & \sigma_v\rho \\ \sigma_v\rho' & \sigma_v^2 \end{bmatrix} \right)$$

$$\Omega_x = \begin{bmatrix} 0.3 & 0.21 & 0.21 \\ 0.21 & 0.3 & 0.21 \\ 0.21 & 0.21 & 0.3 \end{bmatrix}$$

$$\Omega = \begin{bmatrix} \sigma_{\varepsilon 2}^2 & 0 \\ 0 & \sigma_{\varepsilon 3}^2 \end{bmatrix} = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}$$

$$\rho = \begin{bmatrix} 0.7 \\ 0.7 \end{bmatrix}$$

$u_i^* \sim \mathbf{N}^+(0,1)$

$u_i = \sigma_{ui} u_i^*.$

For Scenario 4, the DGP is the same as Scenario 3 but after generating v_i we replace it by $v_i + 0.5\sigma_v(u_i^* - E[u_i^*])$ after normalizing the variance to σ_v^2, which generates correlation between u_i^* and v_i. This violates our assumption that u_i^* is independent from v_i conditional on endogenous and explanatory exogenous variables.