Endogeneity in Stochastic Frontier Models: Copula Approach without O~~o~~utside ~~instruments~~Instruments

Formatted: Font: 11 pt, Bold

Formatted: Left: 1.27 cm, Right: 1.27 cm, Top: 1.27 cm, Bottom: 1.27 cm

Formatted: Font: 11 pt

Kien C. Tran[*]
Department of Economics,
University of Lethbridge,
4401 University Drive W
Lethbridge, Alberta,
T1K 3M4 CANADA, email: kien.tran@uleth.ca

~~and~~

Efthymios G. Tsionas,
Department of Economics,
Athens University of Economics and Business
76 Pattison Street,
10434 Athens, GREECE

Abstract

This papers considers an alternative estimation procedures for estimating stochastic frontier models with endogenous regressors when no external instruments are available. The approach we propose is based on copula function to directly model the correlation between the endogenous regressors and the composed errors. Estimation of model parameters is done using maximum likelihood. Monte Carlo simulations are used to assess and compare the finite sample performances of the proposed estimation procedures.

JEL Classification: C13, C14.

Keywords: Stochastic frontier model, Endogenous regressors, Copula function, Maximum likelihood.

1. Introduction

A standard approach to handle endogeneity problem in the stochastic frontier models is to use likelihood based instrumental variable estimation methods, see for example, Kutlu (2010), Tran and Tsionas (2013) and Amsler, Prokhorov and Schmidt (2014). This type of approach relies upon the availability of a set of outside information that may be used to construct instruments either in the reduced form equations or the instruments themselves. Unlike the standard linear models, the main disadvantage in the stochastic frontier setting is that a substantive assumption needs to be made regarding the correct specification of the reduced form in order to correctly predict the technical inefficiency component. In addition, the instruments, if they are available, often subject to potential pitfalls because they fail to meet the two required conditions adequately that the instruments are sufficiently correlated with the endogenous regressors, and they are uncorrelated with the composed errors term. Thus, the potential difficulty of implementing these approaches is when there is no outside information available to construct the appropriate instruments.

To alleviate these problems, this paper considers an alternative approach to handle endogeneity in stochastic frontier models, which *does not require the availability of outside information to construct the instruments*. Consequently, we can construct a flexible joint distribution of the endogenous regressor and the composed error that can accommodate any degree of dependency between them. We then use this joint distribution to derive the likelihood function and maximize it to obtain the consistent estimates of the model parameters.

2. The Model and Methodology

Consider the following stochastic frontier model:

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

where y_i is the output of firm i, z_i is a $d \times 1$ vector of exogenous input, x_i is a $p \times 1$ vector endogenous input, α and β are $d \times 1$ and $p \times 1$ vectors of unknown parameters, v_i is a symmetric random error, u_i is the one-sided random disturbance representing technical inefficiency. We assume that z_i is uncorrelated with v_i and u_i but x_i are allowed to be correlated with v_i and possibly with u_i, and this generates the endogeneity problem. We also assume that u_i and v_i are independent and leave the form of u_i unrestricted. The discussion that follows can be easily extended for the case where (exogenous) environmental variables are included in the distribution of u_i (e.g., Battese and Coelli (1995)). Following standard practice, assume that $v_i \sim i.i.d.N(0,\sigma_v^2)$ and $u_i \sim i.i.d.\left|N(0,\sigma_u^2)\right|$. Then the density of $\epsilon_i = v_i - u_i = y_i - z_i'\alpha - x_i'\beta$ is given by

$$g(\epsilon_i) = \int_0^\infty f_v(\epsilon_i + u_i) f_u(u_i)\, du_i = \frac{2}{\sigma}\phi\left(\frac{\epsilon_i}{\sigma}\right)\Phi\left(-\frac{\lambda\epsilon_i}{\sigma}\right), \tag{2}$$

where $\sigma^2 = \sigma_v^2 + \sigma_u^2$, $\lambda = \sigma_u / \sigma_v$, $\phi(.)$ and $\Phi(.)$ are the probability density function and cumulative distribution function of a standard normal random variable, respectively.

Let $F(x_1,\ldots,x_p,\epsilon)$ and $f(x_1,\ldots,x_p,\epsilon)$ be the joint distribution and the joint density of $(x_1,\ldots,x_p)$ and ϵ_i, respectively. In practice, $F(x_1,\ldots,x_p,\epsilon)$ and $f(x_1,\ldots,x_p,\epsilon)$ are typically unknown and hence need to be estimated. Following Park and Gupta (2012), we suggest a copula approach to construct and estimate this joint density. The copula essentially captures the dependence in the joint distribution of the endogenous regressors and the composed errors. For exposition purpose, suppose we have a joint distribution of $(x_1,\ldots,x_p,\epsilon)$ with joint density $f(x_1,\ldots,x_p,\epsilon)$, and let $f_j(x_j)$, $F_j(x_j)$, for $j = 1,\ldots,p$, $g(\varepsilon)$ and $G(\varepsilon)$ denote the marginal density and CDF of x_j and ε, respectively. Also let C denotes the "copula function" defined for $(x_1,\ldots,x_{p+1}) \in [0,1]^{p+1}$ by

$$C(x_1,\ldots,x_{p+1}) = P(F_1(x_1) \le x_1,\ldots,F_p(x_p) \le x_p, G(\varepsilon) \le x_{p+1}),$$

so that the copula function is itself a CDF. Moreover, since $F_j(x_j)$ and $G(.)$ are marginal distribution function, each component$U_j = F_j(x_j)$ and $U_\varepsilon = G(\varepsilon)$ has a uniform marginal distribution (see for example Li and Racine (2007, Theorem A.2)). Let $c(x_1,\ldots,x_p)$ denotes the pdf associated with $C(x_1,\ldots,x_p)$, then by Sklar's theorem (Sklar (1959)), we have

$$f(x_1,\ldots,x_p,\varepsilon) = c(F_1(x_1),\ldots,F_p(x_p),G(\varepsilon))g(\varepsilon)\prod_{j=1}^{p} f_j(x_j). \qquad (3)$$

Thus, equation (3) shows that the copula function completely characterizes the dependence structure of $(x_1,\ldots,x_p,\varepsilon)$, and $c(x_1,\ldots,x_p) = 1$ if and only if $(x_1,\ldots,x_p,\varepsilon)$ are independent of each other. For more rigorous treatment on Copula, see Nelsen (2006). To obtain the joint density in (3), we need to specify the copula function. One commonly used copula function is the Gaussian copula. Other copula functions such as Frank, Placket, Clayton, and Farlie-Gumbel-Morgenstern can also be used. The Gaussian copula is generally robust for most application and has many desirable properties (Danaher and Smith (2011)). Let $\Phi_{\Sigma,p+1}$ denote a $(p+1)$-dimensional CDF with zero mean and correlation matrix Σ. Then the $(p+1)$-dimensional CDF with correlation matrix Σ is given by $C(w;\Sigma) = \Phi_{\Sigma,p+1}(\Phi^{-1}(U_1),\ldots,\Phi^{-1}(U_p),\Phi^{-1}(U_\varepsilon))$, where $w = (U_1,\ldots,U_p,U_\varepsilon) = (F_1(x_1),\ldots,F_p(x_p),G(\varepsilon))$. The copula density is

$$\begin{aligned} c(w;\Sigma) = {} & (\det(\Sigma))^{-1/2} \times \\ & \exp\left\{-\frac{1}{2}(\Phi^{-1}(U_1),\ldots,\Phi^{-1}(U_p),\Phi^{-1}(U_\varepsilon))'(\Sigma^{-1} - I_{p+1})(\Phi^{-1}(U_1),\ldots,\Phi^{-1}(U_p),\Phi^{-1}(U_\varepsilon))\right\}. \end{aligned} \qquad (4)$$

The log-likelihood function corresponding to (5) is then

$$\ln L(\theta,\Sigma) = \sum_{i=1}^{n}\left\{\ln c(F_1(x_{1i}),\ldots,F_p(x_{pi}),G(\varepsilon_i;\theta);\Sigma) + \sum_{j=1}^{p}\ln f_j(x_{ji}) + \ln g(\varepsilon_i;\theta)\right\}, \qquad (5)$$

where $\theta = (\alpha',\beta',\lambda,\sigma^2)'$ and the form of $c(.)$ is given in (4). Notice that the first term in the summation in (5) is derived from the copula density and this term reflects the dependence between the endogenous variables and the composed errors. In addition, since the marginal density $f_j(x_j)$ does not contain any parameters of interest, the second term in the summation in (5) can be dropped from the log-likelihood function. Finally, it is clear from (5) that if there are no endogeneity problem, (5) collapses to the log-likelihood function of the standard stochastic frontier models. By maximizing the log-likelihood function in (5), consistent estimates of (θ,Σ) can be obtained, and this can be done in a two-step estimation procedure describe below.

Step 1: Estimation of $F_j(x_j),\ j=1,\ldots,p$; and $G(\varepsilon;\theta)$

Since we have observed sample of $x_{ji},\ j=1,\ldots,p;\ i=1,\ldots,n$; in the first step, we can estimate $F_j(x_{ji})$ by

$$\tilde{F}_{nj}=\frac{1}{n+1}\sum_{i=1}^{n}1(x_{ji}\le x_{0j}),\ j=1,\ldots,p, \qquad (6)$$

where $1(.)$ is an indicator function. Note that in (6), we have used the rescaling factor $1/(n+1)$ rather than $1/n$ to avoid difficulties arising from the potential unboundedness of the $\ln c(F_1(x_{1i}),\ldots,F_p(x_{pi}),G(\varepsilon_i;\theta);\Sigma)$ as some of the $F_j(x_j)$ tend to one. To estimate $G(\varepsilon_i;\theta)$, note that its density $g(\varepsilon_i;\theta)$ is given in (2) and by definition, $G(\varepsilon_i;\theta)=\int_{-\infty}^{\varepsilon_i} g(s;\theta)ds$, thus $G(\varepsilon_i;\theta)$ can be estimated using numerical integration, and let $\tilde{G}(\varepsilon_i;\theta)$ denotes the estimator of $G(\varepsilon_i;\theta)$.

Step 2: Maximize the log-likelihood function

Maximize the log-likelihood function in (5) with $F_j(x_j)$ and $G(\varepsilon_i;\theta)$ are replaced by their estimates $\tilde{F}_j(x_j)$ and $\tilde{G}(\varepsilon_i;\theta)$, respectively, i.e.,

$$(\hat{\theta},\hat{\Sigma})=\underset{\theta\in\Theta,\Sigma}{\arg\max}\sum_{i=1}^{n}\left\{\ln c(\tilde{F}_1(x_{1i}),\ldots,\tilde{F}_p(x_{pi}),\tilde{G}(\varepsilon_i;\theta);\Sigma)+\ln g(\varepsilon_i;\theta)\right\}. \qquad (7)$$

Predicting Technical Inefficiency:

Once the parameters have been estimated, technical inefficiency u_i, can be predicted based on Jondrow et al. (1982):

$$\hat{u}_i=\hat{E}(u_i\mid\varepsilon_i)=\frac{\hat{\sigma}\hat{\lambda}}{1+\hat{\lambda}^2}\left[\frac{\phi(\hat{\lambda}\hat{\varepsilon}_i/\hat{\sigma})}{1-\Phi(\hat{\lambda}\hat{\varepsilon}_i/\hat{\sigma})}-\frac{\hat{\lambda}\hat{\varepsilon}_i}{\hat{\sigma}}\right],$$

where $\hat{\varepsilon}_i=y_i-z_i'\hat{\alpha}-x_i'\hat{\beta}$ and $\hat{\alpha},\hat{\beta},\hat{\lambda}$ and $\hat{\sigma}^2$ are the parameter estimates obtained from the Copula approach discussed above.

3. Monte Carlo Simulations

To examine the finite sample performance of the proposed Copula estimator we conduct some Monte Carlo experiments. We consider the following data generating process:

$$\begin{aligned}y_i&=z_{1,i}\alpha+x_i\beta+v_i-u_i,\\ x_i&=z_{2,i}\gamma+\varepsilon_i,\end{aligned}$$

where $u_i \sim i.i.d.\ \left|N(0,\sigma_u^2)\right|$ and the random variables $z_{1,i}$ and $z_{i,2}$ are each generated independently as $\chi^2_{(2)}$. The vector of random errors $(v_i, \varepsilon_i)'$ is generated by

$$\begin{pmatrix} v_i \\ \varepsilon_i \end{pmatrix} \sim N\left(0_2, \begin{pmatrix} \sigma_v^2 & \rho\sigma_v\sigma_\varepsilon \\ \rho\sigma_v\sigma_\varepsilon & \sigma_\varepsilon^2 \end{pmatrix}\right).$$

In our experiment, we fix $\alpha = \beta = 0.5$, $\sigma_\varepsilon^2 = 1$, $\sigma_v^2 = \sigma_u^2 = 1$ and $\gamma = 1$. We set the values of $\rho = \{0.0, 0.4, 0.8\}$ and consider two sample sizes: $n = (750, 1500)$. The simulations are replicated 1,000 times.[1] Note that since our DGP contains only one endogenous regressor, the Gaussian copula function has a simple form:

$$\begin{aligned} C(U_x, U_\varepsilon) &= N_{2,\rho}(\Phi^{-1}(U_x), \Phi^{-1}(U_\varepsilon)) \\ &= \frac{1}{2\pi(1-\rho^2)^{1/2}} \int_{-\infty}^{\Phi^{-1}(U_x)} \int_{-\infty}^{\Phi^{-1}(U_\varepsilon)} \exp\left\{ \frac{-(s^2+t^2-2\rho st)}{2(1-\rho^2)} \right\} ds\,dt, \end{aligned}$$

where $N_{2,\rho}(.,.)$ denotes the standard bivariate normal distribution function with correlation ρ. The corresponding copula density is:

$$c(U_x, U_\varepsilon) = \frac{1}{(1-\rho^2)^{1/2}} \exp\left\{ \frac{-\rho^2(x^{*2}+\varepsilon^{*2}) + 2\rho x^* \varepsilon^*}{2(1-\rho^2)} \right\},$$

and the log-likelihood function associated with (5) can be expressed as:

$$\ln L(\theta) = -\frac{n}{2}\ln(1-\rho^2) - \sum_{i=1}^{n} \left\{ \frac{\rho^2(x_i^{*2}+\varepsilon_i^{*2}) - 2\rho x_i^* \varepsilon_i^*}{2(1-\rho^2)} \right\} + \sum_{i=1}^{n} \ln g(y_i - z_i\alpha - x_i\beta),$$

where $\theta = (\alpha, \beta, \rho, \lambda, \sigma^2)'$, $x_i^* = \Phi^{-1}(F(x_i))$, $\varepsilon_i^* = \Phi^{-1}(G(y_i - z_i\alpha - x_i\beta))$ and the form of $g(.)$ is given in (2). In our simulation, the numerical integration to obtain $G(.)$ is performed using Gaussi-Kronrod quadrature with 50 points.

For comparison purpose, we also compute the standard MLE and GMM estimators of Tran and Tsionas (2013). Simulation results of the parameter estimates' MSE are displayed in Tables 1-3. Our simulations show that when there is no correlation (i.e., no endogeneity in the regressor), as expected, all estimators performed as well as the standard MLE for all ranges of the parameters considered. However, as expected, when there is correlation and as the correlation increases, the MLE deteriorates quickly and becomes severely biased, regardless of the sample sizes. Comparing the performance of the proposed Copula estimator to the GMM estimator of Tran and Tsionas (2013) shows that the proposed estimator perform quite well in term of MSE. Finally, as the sample size n doubles, the estimated MSEs of the proposed Copula estimator reduces to about half of the original values regardless whether there is correlation or not; this is consistent with the fact that the Copula estimator is $\sqrt{n}-$ consistent estimator.

4. Conclusion

[1] All computations are performed in Fortran 77 using extensively IMSL routines.

In this paper, we offer an alternative approach for estimating stochastic frontier models with endogenous regressors when there is no additional data available that is based on copula method to directly construct the joint density of the endogenous regressors and the composed errors to capture their dependency. We examine the finite sample behavior of the proposed approach via Monte Carlo simulations. The results from the simulations showed that the estimators are performed very well in finite samples in term of bias and MSE.

References

Amsler, C., A. Prokhorov and P. Schmidt (2014), “Endogeneity in stochastic frontier models.” *Working Paper*, Michigan State University.

Battese. G.E. and T.J. Coelli (1995), “A model for technical inefficiency effects in a stochastic frontier production function for panel data.” *Empirical Economics*, 20, 325-332.

Danaher, P.J. and M.S. Smith (2011), “Modeling multivariate distributions using copulas: Applications in marketing.” *Marketing Science*, 30 (1), 4-21.

Jondrow, J., C.A.K. Lovell, I.S. Materov and P. Schmidt (1982). “On the estimation of technical inefficiency in the stochastic frontier production function model.” *Journal of Econometrics*, 19 (2/3), 233-238.

Kutlu, L. (2010), “Battese-Coelli estimator with endogenous regressors.” *Economics Letters*, 109, 79-81.

Li, Q. and J. Racine, (2007). *Nonparametric Econometrics*. Princeton University Press, Princeton, NJ.

Nelsen, R.B. (2006), *An Introduction to Copula*, Vol. 139 of Springer Series in Statistics, Springer.

Park, S. and S. Gupta (2012), “Handling endogenous regressors by joint estimation using Copulas.” *Marketing Science*, 31 (4), 567-586.

Sklar, A. (1959), “Fonctions de repartition a n dimensions et leurs marges.” *Publications de l’Institut de Statistique de l’Uinversite de Paris*, 8:229-231.

Tran, K.C. and E.G. Tsionas (2013), “GMM estimation of stochastic frontier models with endogenous regressors.” *Economics Letters*, 118, 233-236.

Formatted: Space After: 8 pt, Line spacing: 1.5 lines, Keep with next, Keep lines together

Table 1: Estimated MSE: $r = 0.0, s_u = 1, s_v = 1$

	$n = 750$			$n = 1500$		
	MLE	**GMM**	**Copula**	**MLE**	**GMM**	**Copula**
a	.001577	.001580	.001585	.000727	.000729	.000731
b	.000797	.000797	.000812	.000404	.000407	.000409
g	-	.001099	-	-	.000578	-
s_v	.001463	.001458	.001533	.000735	.000737	.000736
s_u	.003002	.003015	.003241	.001336	.001339	.001341
r	-	.003366	.003561	-	.001714	.001717
Total MSE*	.006839	.006850	.00717	.003202	.003212	.003217

***** Indicates the total MSE of the four parameters (as in the MLE case).

Table 2: Estimated MSE: $r = 0.4, s_u = 1, s_v = 1$

	$n = 750$			$n = 1500$		
	MLE	**GMM**	**Copula**	**MLE**	**GMM**	**Copula**
a	.013767	.001416	.001543	.007546	.000728	.000721
b	.033650	.001579	.001583	.032991	.000757	.000763
g	-	.001149	-	-	.000578	-
s_v	.003230	.001638	.001644	.002206	.000820	.000815
s_u	.002927	.002517	.002510	.001320	.001193	.001199
r	-	.002608	.002600	-	.001265	.001273
Total MSE*	.053574	.007150	.007155	.044063	.003497	.003498

***** See Table 1.

Table 3: Estimated MSE: $r = 0.8, s_u = 1, s_v = 1$

	$n = 750$			$n = 1500$		
	MLE	**GMM**	**Copula**	**MLE**	**GMM**	**Copula**
a	.016041	.001532	.001542	.008756	.000708	.000716
b	.133403	.001296	.001299	.132104	.000742	.000747
g	-	.001037	-	-	.000569	-
s_v	.026804	.006471	.006472	.025901	.001041	.001044
s_u	.008483	.001099	.001095	.001230	.000692	.000689
r	-	.000546	.000543	-	.000288	.000286
Total MSE*	0.184731	.010398	.010408	.167991	.003183	.003196

***** See Table 1.