Economics Letters 133 (2015) 85–88

Contents lists available at ScienceDirect

Economics Letters

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

Endogeneity in stochastic frontier models: Copula approach without external instruments

Kien C. Tran [a,*], Efthymios G. Tsionas [b]

[a] Department of Economics, University of Lethbridge, 4401 University Drive W, Lethbridge, Alberta, T1K 3M4, Canada

[b] Department of Economics, Lancaster University Management School, LA1 4YX, UK

HIGHLIGHTS

- We propose a Copula approach for estimating endogenous stochastic frontier models.
- We discuss the model identification strategy.
- Maximum likelihood estimation procedure is proposed.
- Monte Carlo results show that the proposed estimator performs well in finite sample.

ARTICLE INFO

Article history:
Received 15 April 2015
Received in revised form
18 May 2015
Accepted 20 May 2015
Available online 28 May 2015

JEL classification:
C13
C14

Keywords:
Stochastic frontier model
Endogenous regressors
Copula function
Maximum likelihood

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.

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 et al. (forthcoming). 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. The specification of the reduced equations has the advantage that it provides more efficient estimates of the frontier parameters as well as improvement in predicting inefficiency term. However, 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 potential problems, this paper considers an alternative approach to handle the endogeneity problem in stochastic frontier models, which *does not require the availability of outside information to construct the instruments*. The method we propose is based on Copula function to directly model the dependency of the endogenous regressors and the composed error. Specifically, copulas allow us to model the marginal distributions

* Corresponding author. Tel.: +1 403 329 2511; fax: +1 403 329 2519.
E-mail addresses: kien.tran@uleth.ca (K.C. Tran), tsionas@aueb.gr (E.G. Tsionas).

http://dx.doi.org/10.1016/j.econlet.2015.05.026
0165-1765/

of the endogenous regressors and composed error separately from their dependency. 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.

This paper is organized as follows. Section 2 presents the model and discusses the Copula approach to deal with endogeneity issues in stochastic frontier framework when outside information may not be available for use as instruments. In Section 3 we examine the finite sample performance of the proposed two approaches via Monte Carlo simulations. Section 4 concludes the paper.

2. The model and methodology

Consider the following stochastic frontier model:

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

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 the (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 $\varepsilon_i = v_i - u_i = y_i - z_i'\alpha - x_i'\beta$ is given by

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

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

Before discussing the Copula approach, we briefly describe the model identification. Under our setting, the model is identified as long as σ_u^2 is not zero or very close to zero, even if the endogenous regressors happen to be normal. However, when $\sigma_u^2 = 0$ (implying ε_i is normal) and the endogenous regressors are normal, the model identification breaks down because it is difficult to distinguish the variations as results of endogenous regressors from the variation due to the composed-error.[1] Consequently, this identification problem has important implications for testing the null hypothesis of fully efficient firms (i.e., $H_0 : \sigma_u^2 = 0$). For more details discussion on the identification issue, see Online Appendices 1 and 2 of Park and Gupta (2012) (at http://pubsonline.informs.org/doi/suppl/10.1287/mksc.1120.0718).

Let $F(x_1, \dots, x_p, \varepsilon)$ and $f(x_1, \dots, x_p, \varepsilon)$ be the joint distribution and the joint density of $(x_1, \dots, x_p)$ and ε_i, respectively. In practice, $F(x_1, \dots, x_p, \varepsilon)$ and $f(x_1, \dots, x_p, \varepsilon)$ 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 $(x_1, \dots, x_p, \varepsilon)$ with joint density $f(x_1, \dots, x_p, \varepsilon)$, and let $f_j(x_j), F_j(x_j)$, for $j = 1, \dots, p, g(\varepsilon)$ and $G(\varepsilon)$ denote the marginal density and CDF of x_j and ε, respectively. Also let C denote the "copula function" defined for $(\xi_1, \dots, \xi_{p+1}) \in [0, 1]^{p+1}$ by

$$C(\xi_1, \dots, \xi_{p+1}) = P(F_1(x_1) \le \xi_1, \dots, F_p(x_p) \le \xi_p, G(\varepsilon) \le \xi_{p+1}),$$

so that the copula function is itself a CDF. Moreover, since $F_j(x_j)$ and $G(\cdot)$ 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 $c(\xi_1, \dots, \xi_p)$ denote the pdf associated with $C(\xi_1, \dots, \xi_p)$, then by Sklar's theorem (Sklar, 1959), we have

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

Thus, Eq. (3) shows that the copula function completely characterizes the dependence structure of $(x_1, \dots, x_p, \varepsilon)$, and $c(\xi_1, \dots, \xi_p) = 1$ if and only if $(x_1, \dots, x_p, \varepsilon)$ 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 applications 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

$$C(w; \Sigma) = \Phi_{\Sigma,p+1}(\Phi^{-1}(U_1), \dots, \Phi^{-1}(U_p), \Phi^{-1}(U_\varepsilon)),$$

where $w = (U_1, \dots, U_p, U_\varepsilon) = (F_1(x_1), \dots, F_p(x_p), G(\varepsilon))$. The copula density is

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

The log-likelihood function corresponding to (4) is then

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

where $\theta = (\alpha', \beta', \lambda, \sigma^2)'$ and the form of $c(\cdot)$ 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* $F_j(x_j)$, $j = 1, \dots, p$; and $G(\varepsilon; \theta)$

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

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

[1] We would like to thank an anonymous referee for pointing this out.