Journal of the Korean Statistical Society 45 (2016) 199–209

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Semiparametric maximum likelihood estimation of stochastic frontier model with errors-in-variables[☆]

Byungtae Seo[a], Seok-Oh Jeong[b,*]

[a] Sungkyunkwan University, South Korea

[b] Hankuk University of Foreign Studies, South Korea

ARTICLE INFO

Article history:
Received 27 January 2015
Accepted 28 September 2015
Available online 17 October 2015

AMS 2000 subject classifications:
primary 62F99
secondary 62P20

Keywords:
Stochastic frontier model
Semiparametric maximum likelihood estimation
Errors-in-variables
Measurement error
Mixture model
Constrained Newton method
Productivity analysis

ABSTRACT

This paper presents a new technique to analyze a stochastic frontier model when covariates are incorporated with measurement errors. We propose a semiparametric mixture likelihood method to estimate the stochastic frontier model which is free from any erroneous specification of the distribution of latent covariates. Some numerical studies including a real data analysis were done, which highly support the proposed approach.

1. Introduction

The performance of a decision making unit (DMU) can be assessed by computing the efficiency of its production activities. Suppose we observe a firm producing an output of level $y \in \mathbb{R}$ manipulating input quantities $\boldsymbol{x} \in \mathbb{R}^p$. Then the efficiency of the firm is analyzed by comparing y with the feasible maximal output level that can be achieved by using the input amount of $\boldsymbol{x}$. Given an input level of $\boldsymbol{x}$ we may define the maximum level of y, say $g(\boldsymbol{x})$, as a function of $\boldsymbol{x}$, and we call it the *production function* or the *frontier function*. Once the production function is given, the efficiency of a production plan $(\boldsymbol{x}, y)$ can be evaluated by the distance from this point to the production frontier. However, since the production frontier is usually unknown, the production frontier should be estimated by the observed production activities of DMU's. For this, Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977) introduced the stochastic frontier (SF) model as follows:

$$Y = g(\boldsymbol{X}|\boldsymbol{\beta}) - U + V, \tag{1}$$

☆ Seok-Oh Jeong gratefully acknowledges the financial support by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2006112). The research of Byungtae Seo was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2013R1A1A2057715).

* Corresponding author.
E-mail address: seokohj@hufs.ac.kr (S.-O. Jeong).

http://dx.doi.org/10.1016/j.jkss.2015.09.006
1226-3192/

where Y is the output quantity produced using input quantities $\boldsymbol{X}$, g is the production frontier parameterized by a vector $\boldsymbol{\beta}$, U is a nonnegative inefficiency factor and V is an error variable. Let $\mathcal{X}_n = \{(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \ldots, (\boldsymbol{X}_n, Y_n)\}$ be a random sample from the distribution of $(\boldsymbol{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$. Using $\mathcal{X}_n$, one may estimate g (or $\boldsymbol{\beta}$) and other parameters involved in the model under some reasonable parametric distributional assumptions on U and V by the maximum likelihood estimation procedure.

For example, consider a linear SF model with $p = 1$:

$$g(x|\boldsymbol{\beta}) = \beta_0 + \beta_1 x, \qquad U \sim N^+(0, \sigma_u^2), \qquad V \sim N(0, \sigma_v^2) \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$, $\sigma_u > 0$ and $\sigma_v > 0$ are unknown parameters. Then, the log-likelihood function is given by

$$\ell_n(\beta_0, \beta_1, \sigma_u, \sigma_v|\mathcal{X}_n) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \epsilon_i^2 + \sum_{i=1}^{n} \log \Phi\left(-\frac{\lambda}{\sigma}\epsilon_i\right)$$

where $\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$, $\lambda = \sigma_u/\sigma_v$, $\sigma^2 = \sigma_u^2 + \sigma_v^2$, and Φ is the cumulative distribution function of $N(0, 1)$. The maximum likelihood (ML) estimator of $(\beta_0, \beta_1, \sigma_u, \sigma_v)$ is defined by the maximizer of this. We may alter the distributional assumption on the inefficiency factor U to other distributions such as exponential, gamma, and so on. See Kumbhakar and Lovell (2000) for details.

The stochastic frontier model (1) assumes that the input variable $\boldsymbol{X}$ is measured perfectly so that it is not involved with any errors, which would become easily infeasible in practice. This paper considers an errors-in-variables (EIV) version of SF model:

$$Y = g(\boldsymbol{\xi}|\boldsymbol{\beta}) - U + V, \qquad \boldsymbol{X} = \boldsymbol{\xi} + \boldsymbol{W} \tag{3}$$

where Y is the output quantity produced using input quantities, $\boldsymbol{\xi}$ is the true (but latent) measurement of input variables, $\boldsymbol{X}$ is the observed inputs, g is a parametric frontier function, U is a nonnegative inefficiency factor, and V and $\boldsymbol{W}$ are error variables. In this setting, like in usual regression problems, a naive ML estimation based on $(\boldsymbol{X}, Y)$ ignoring bW is problematic. Fig. 1 provides an empirical evidence of this. According to the linear SF model in (2), we conducted 1000 Monte Carlo experiments to obtain the deviations of the parameter estimates and their targeted true values in both no-errors-in-variable case and with-errors-in-variable case. For each experiment, the sample size was 500 and the covariates were generated by $\xi \sim \text{Exp}(1)$ and $X = \xi + W$ with $W \sim N(0, 1)$. The true parameter values used in the experiments were $\beta_0 = 1$, $\beta_1 = 1$, $\sigma_u = \sqrt{\frac{2\pi}{\pi-2}}$ and $\sigma_v = 1$. It is commonly observed in each panel of Fig. 1 that, while the ML estimation for the data without measurement errors works quite well (left), the naive ML estimation collapses when the data are incorporated with measurement errors (right).

To the best of our knowledge, Chen and Wang (2004) is almost the first paper that carefully dealt with the measurement error problem in the stochastic frontier analysis. They suggested a method of moments (MoM) estimator for the linear stochastic frontier model to find all model parameters, and it works at least better than the usual maximum likelihood method that does not consider measurement errors in covariates. However, it suffers from the limitations in application by the additional assumptions such as the statistical independence among the latent variables $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_p)$ and the asymmetry of the distributions of ξ_j's, which can hardly be verified in practice since $\boldsymbol{\xi}$ is not directly observable. Recently, Chang, Chen, and Wang (2012) proposed a Bayesian approach to improve MoM estimator. However, they assumed a normal distribution for the latent covariates which would suffer from a mis-specification problem.

In this paper, we propose a new method to estimate the model in (3) which is free from a mis-specification of the probabilistic model for the latent variable $\boldsymbol{\xi}$. The proposed approach is semiparametric in that the distribution of $\boldsymbol{\xi}$ can be considered as a nonparametric component in the model since it is completely unspecified. Also, the proposed method can be applied to the model with any parametric form of the frontier function g and with any distributional assumptions on the inefficiency factor U. The agenda of the paper is as follows. In Section 2, we propose a semiparametric mixture approach and describe the estimation procedure. Section 3 presents the results from some simulation studies and real data analysis, and Section 4 concludes the paper.

2. The method

2.1. The semiparametric mixture model

Recall the stochastic frontier model with errors-in-variables:

$$Y = g(\boldsymbol{\xi}|\boldsymbol{\beta}) - U + V, \qquad \boldsymbol{X} = \boldsymbol{\xi} + \boldsymbol{W} \tag{4}$$

where $(\boldsymbol{X}, Y) = (X_1, X_2, \ldots, X_p, Y)$ is the vector of observed p-inputs and an output, $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_p)$ is the latent input variables, g is the production frontier parameterized by $\boldsymbol{\beta} \in \mathbb{R}^d$, U is a nonnegative inefficiency factor, and V and $\boldsymbol{W} = (W_1, W_2, \ldots, W_p)$ are the error variables. For technical convenience, $\boldsymbol{W}$ is assumed to be independent of $(\boldsymbol{\xi}, U, V)$. Also, we assume the normality of the error variables, i.e., $V \sim N(0, \sigma_v^2)$ and $\boldsymbol{W} \sim N_p(\boldsymbol{0}_p, \text{diag}(\sigma_{w_1}^2, \ldots, \sigma_{w_p}^2))$,