Journal of the Korean Statistical Society 45 (2016) 422–438

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Elliptical regression models for multivariate sample-selection bias correction

Hea-Jung Kim [a], Hyoung-Moon Kim [b,*]

[a] Department of Statistics, Dongguk University—Seoul, Seoul, Republic of Korea
[b] Department of Applied Statistics, Konkuk University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 18 May 2015
Accepted 11 January 2016
Available online 3 February 2016

AMS 2000 subject classifications:
primary 62J99
secondary 60E05

Keywords:
Bias correction
Heckman model
MCECM algorithm
Multivariate sample-selection regression
Rectangle-screened scale mixture of normal

ABSTRACT

In linear regression, a multivariate sample-selection scheme often applies to the dependent variable, which results in missing observations on the variable. This induces the sample-selection bias, i.e. a standard regression analysis using only the selected cases leads to biased results. To solve the bias problem, in this paper, we propose a class of multivariate selection regression models by extending classic Heckman model to allow for multivariate sample-selection scheme and robustness against departures from normality. Necessary theories for building a formal bias correction procedure, based upon the proposed model, are obtained, and an efficient estimation method for the model is provided. Simulation results and a real data example are presented to demonstrate the performance of the estimation method and practical usefulness of the multivariate sample-selection models.

1. Introduction

Many authors working with regression analysis have considered the problem of potentially biased estimates arising from the selection process that generated the sample (e.g., Azzalini & Capitanio, 1999; Heckman, 1974; Hevia & Arrazola, 2009 and Marchenko & Genton, 2012). This problem naturally occurs when observations of the dependent variable of the regression model are missing not at random (MNAR; Rubin, 1976) owing to a sample-selection rule such as incidental truncation, hidden truncation, or censoring (e.g., Greene, 2008). That is, observations of the dependent variable, $y_i^* \in \mathbb{R}$, in the regression model can be observed as $y_i = y_i^*$ only when a corresponding latent variable $u_i^* \in \mathbb{R}$ belongs to an interval $C \subset \mathbb{R}$ of its support. For the case of an unbounded interval $C \equiv (0, \infty)$, Heckman (1974) introduced the classic sample-selection model as follows.

$$y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, N, \tag{1}$$

with a single-selection equation

$$u_i^* = \mathbf{w}_i^\top \boldsymbol{\gamma} + \eta_i, \quad i = 1, \ldots, N, \tag{2}$$

such that the observations are missing on y_i^* for all cases in which $u_i^* \leq 0$. Here the vectors $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\boldsymbol{\gamma} \in \mathbb{R}^r$ are unknown parameters, $\mathbf{x}_i \in \mathbb{R}^q$, $\mathbf{w}_i \in \mathbb{R}^r$, and the joint distribution of the error terms of ϵ_i and η_i is assumed to follow

$$\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right]. \tag{3}$$

* Corresponding author.
E-mail address: hmkim@konkuk.ac.kr (H.-M. Kim).

http://dx.doi.org/10.1016/j.jkss.2016.01.003
1226-3192/

Under the "Heckman model" (also known as "Type 2 tobit model" considered by Amemiya, 1985), our interest is to estimate $\boldsymbol{\beta}$ by using observed values of the explanatory variables for all the sample, and N_1 of the N observations for y_i^* satisfied by the univariate selection (or single-selection) rule. Since we only observe N_1 observations $y_i = y_i^*$ for which $s_i = 1$ with $s_i = I(u_i^* \in C)$, the regression function for the Heckman model reduces to

$$E[y_i|\mathbf{x}_i, \mathbf{w}_i] = \mathbf{x}_i^\top \boldsymbol{\beta} + \rho\sigma\lambda(\mathbf{w}_i^\top \boldsymbol{\gamma}), \quad i = 1, \ldots, N_1 < N, \tag{4}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ and $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal probability density and distribution function, respectively (e.g., Catsiapis & Robinson, 1982). Here, the result is a biased estimator of $\boldsymbol{\beta}$ if we perform the OLS regression for (1) on the selected sample, ignoring the extra term $\rho\sigma\lambda(\mathbf{w}_i^\top \boldsymbol{\gamma})$ in (4) for $\rho \neq 0$. As a result, various bias correction procedures for consistent estimation of $\boldsymbol{\beta}$ in the regression model (1), as well as its applications, have been dealt with extensively in econometrics literature. See, Griliches, Hall, and Hausman (1977) and Heckman (1979) for the estimation, Das, Newey, and Vella (2003) and Newey (1999) for some robust estimation methods, and Hevia and Arrazola (2008) and Hoffmann and Kassouf (2005) for some empirical applications. With regard to the variants of Heckman model, Marchenko and Genton (2012) used the heavy-tailed bivariate t-distribution for the error model (3) to relax the assumption of normality, and Catsiapis and Robinson (1982) considered a bias correction procedure for the regression model involving multiple independent selection rules. The model also has been generalized to the case of bivariate selection rules (i.e. two correlated selection rules) and applied to the empirical analysis of different phenomena. Mohanty (2001) demonstrated the bivariate selection model in analyzing male–female wage differences; Serumaga-Zake and Naudé (2003) used a bivariate selection scheme to estimate private returns to education in South Africa; and Hevia and Arrazola (2009) considered the marginal effects in the bivariate selection model on wages in Spain.

In practical situations, however, the sample selection process often comprises multiple selection rules. A typical example is the customer segmentation analysis where customers are grouped by multiple selection rules for segmenting the level of consumer attitude, motivations, patterns of usage, and preferences (Jiang & Tuzhilin, 2006; Marcus, 1998). The analysis has been considered as one of the standard techniques used by marketers of insurance agencies, credit card companies, manufacturers and so on. As seen in Section 5, its popularity comes from the fact that segmented models usually outperform aggregated models of customer behavior (Besanko, Dube, & Gupta, 2000). In the segmented model, the observed y_i's may be regarded as outcome of a p-variate selection scheme in (2) with p bounded intervals $C_j^* \equiv (a_j, b_j)$, such that we only observe $y_i = y_i^*$ for $s_i = 1$ and $s_i = I(u_{ij}^* \in C_j^*,\ j = 1, \ldots, p)$ in the sample-selection model (1) with a non-normal error distribution in (3). Unfortunately, even for the case of $p = 1$, all the bias correction procedures, i.e. the Heckman model and its variants, no longer apply to the case where the sample-selection rule is defined by a bounded interval C^* (i.e. C_1^*), because the extra term in the regression function (4) takes a different form. In particular, the first step of the two-step estimation method by Heckman (1979) cannot be applicable to the case of the bounded interval C^*, because the first step estimates $\boldsymbol{\gamma}$ in (4) by using the probit model which fits the binary response s_i's defined by the open interval C. Further, surprisingly, a sample-selection bias correction method for the model with a selection rule using C^* has not been tackled in the literature. These motivated us to consider the contents of this paper.

In this paper, we develop and study the properties of a class of sample-selection models that extends the classical Heckman model to allow for p-variate selection scheme (i.e. multivariate selection model) consisting of various bounded (C^*)/unbounded (C) intervals and robustness against departures from normality of the error distributions. We then propose a sample-selection bias correction procedure which yields a consistent estimation of $\boldsymbol{\beta}$ defined in the multivariate selection model. The rest of this paper is organized as follows. Section 2 generalizes the Heckman model to develop a class of multivariate selection models whose robust-to-normality comes from assuming a family of elliptical distributions for the error distribution in (3). We then study some interesting properties of a model belonging to the class such as the conditional distribution of selected observation, y_i; a hierarchical representation of the distribution of y_i; moments of y_i; and exact likelihood function of the model. In Section 3, we describe some special models belonging to the class and study their relationship with the Heckman model. Section 4 constructs a bias correction procedure which uses an extended EM algorithm for estimating the multivariate selection models. A test procedure for testing existence of the selection bias in the proposed model is also given. The finite-sample performance of the algorithm is examined using a limited, but informative simulation study in Section 5. This section also gives a real data example to demonstrate the performance of the sample-selection bias correction procedure under the multivariate selection models which were proposed in this paper. The paper concludes with a discussion in Section 6.

2. Multivariate selection regression model

In this section, we extend the classical Heckman regression model to a class of multivariate selection regression models (MSRM) whose selection rules are composed of various bounded or unbounded intervals. The primary interest of the regression model is the same as in (1), but now the selection equation (2) becomes a multivariate version. In addition, the distributional assumption of the normal error model in (3) is relaxed by using a flexible elliptically contoured error model.