Economics Letters 141 (2016) 147–150

Contents lists available at ScienceDirect

Economics Letters

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

The equivalence of three latent class models and ML estimators

Vidhura S. Tennekoon [*]

Department of Economics, Indiana University Purdue University Indianapolis (IUPUI), School of Liberal Arts, Cavanaugh Hall Room 524, 425 University Boulevard, Indianapolis, IN 46202, USA

HIGHLIGHTS

- We establish the equivalence of three latent class models.
- They are the binary Roy model, the probit model with a misclassified dependent variable and a trivariate probit model with partial observability.
- The probit model with measurement error is an enhanced version of existing models.
- A researcher working on one of these estimators may benefit from the literature and software related to others.

ARTICLE INFO

Article history:
Received 1 February 2016
Received in revised form
22 February 2016
Accepted 23 February 2016
Available online 2 March 2016

JEL classification:
C01
C10
C30
C40

Keywords:
Binary Roy model
Partial observability model
Misclassified data
Probit model
Switching regression

ABSTRACT

The purpose of this letter is to show the equivalence of three latent class models; the switching regression model with endogenous switching and a latent outcome (the binary Roy model), the probit model with a systematically misclassified dependent variable, and a trivariate probit model with partial observability. The probit model with measurement error is an enhanced version of existing models which allows for the potential correlation between error terms. Establishing this connection, we hope, will help a researcher working on one of these classes of estimators to benefit from the literature and software related to other families.

1. Introduction

The seminal paper published by Andrew D. Roy in 1951 motivated the idea of self-selectivity, which was later developed by several others (see Maddala, 1983). Several econometric models including the switching regression model of Goldfeld and Quandt (1973) and the switching regression model with endogenous switching discussed in Maddala (1983) fall to the broad category of models now referred as Roy models. The switching regression model with endogenous switching is a hierarchical model in which there are two potential states each with an outcome that has a continuous distribution (log wage, for example) while the assignment to one of these two states is determined by a latent variable. When all three outcomes are specified as linear index functions and the errors are assumed to be potentially correlated and distributed normally, the parameters of this model can be estimated using popular statistical packages (Lokshin and Sajaia, 2004). In binary Roy models, the two potential outcomes also are determined by latent variables as in Heckman and Vytlacil (1999).

Poirier (1980) presented and discussed a bivariate probit model with partial observability in which only two of the four potential outcomes are observed. He showed that the usual parameters of the bivariate probit model can also be identified with partial observability under certain conditions. Abowd and Farber (1982) discussed a variant of this partial observability model. These versions of the bivariate probit model are supported in popular statistical packages. Poirier (2014) extends this analysis to multivariate probit models with partial observability and *multivariate pairwise partial observability*. According to Poirier (2014), which cites applications from eleven different fields, there

[*] Tel.: +1 317 278 2845; fax: +1 317 274 0097.
E-mail address: vtenneko@iupui.edu.

http://dx.doi.org/10.1016/j.econlet.2016.02.028
0165-1765/© 2016 Elsevier B.V. All rights reserved.

have been over hundred applications of these models. That paper presents the conditions for identifying multivariate probit models with partial observability and provides an example of a trivariate probit model with partial observability.

A third class of models is the probit model with misclassified dependent variables as discussed in Hausman et al. (1998). Lewbel (2000) showed that the parameters of this model can be identified even when the misclassification probabilities depend on one or more covariates. Tennekoon and Rosenman (2016) presented a model to identify the parameters of this model under parametric assumptions. The applications of this model include Murphy et al. (2015) and Tennekoon and Rosenman (2015). These models, however, do not consider the potential correlation between the error terms. We enhance these existing measurement error models here by introducing potentially correlated errors.

The purpose of this letter is to show the equivalence of the three latent class models discussed above, the switching regression model with endogenous switching and a latent outcome (the binary Roy model), the probit model with a systematically misclassified dependent variable and a trivariate model with partial observability, under potentially correlated trivariate normal errors. Establishing this connection, we hope, will help a researcher working on one of these classes of estimators to benefit from the literature and software related to other families. In addition, the enhanced version of the probit model with a mismeasured dependent variable we present here has not been used elsewhere.

2. The three models

2.1. Binary Roy model

The switching regression model with endogenous switching, widely known as the Roy model (Roy, 1951), assigns a given individual to one of two potential outcome regions using an endogenous switching mechanism. The special case that we discuss here is the binary Roy model in which the two potential outcomes also are observed as binary variables (Heckman and Vytlacil, 1999).

For each individual i, we assume two potential outcomes (Y_{2i}, Y_{3i}) which corresponds to treated and untreated states, respectively. Unlike in the prototypical Roy model, Y_{2i} and Y_{3i} are not continuous variables here. They are binary indicator variables generated by the latent variables Y^*_{2i} and Y^*_{3i}, respectively. The indicator variable Y_{1i} determines the receipt or non-receipt of treatment. Y_{1i} is generated by another latent variable Y^*_{1i}. Y_{1i}, Y_{2i} and Y_{3i} are generated as, $Y_{ji} = 1.\,(Y^*_{ji} > 0)$ and $Y^*_{ji} = \beta_j X_{ji} + \varepsilon_{ji}$ for $j = 1, 2, 3$. The error terms are assumed to be jointly trivariate normally distributed with potentially correlated errors as,

$$\begin{bmatrix}\varepsilon_{1i}\\ \varepsilon_{2i}\\ \varepsilon_{3i}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho_{12} & \rho_{13}\\ \rho_{12} & 1 & \rho_{23}\\ \rho_{13} & \rho_{23} & 1\end{bmatrix}\right).$$

Following the potential outcome model of Rubin (1978), we can write, $Y_i = (1 - Y_{1i})\,.Y_{2i} + Y_{1i}.Y_{3i}$ where Y_i is the observed outcome of individual i. Since $E[Y_{ji}] = Pr[Y_{ji} = 1]$ for $j = 1, 2, 3$,

$$E[Y_i] = \Pr[Y_i = 1] = \Pr[Y_{1i} = 0 \text{ and } Y_{2i} = 1] + \Pr[Y_{1i} = 1 \text{ and } Y_{3i} = 1].$$

Therefore,

$$\begin{aligned} E[Y_i|\,X_{1i}, X_{2i}, X_{3i}] &= \Pr[Y = 1_i|\,X_{1i}, X_{2i}, X_{3i}] \\ &= \Phi_2\left(-\beta_1 X_{1i}, \beta_2 X_{2i}, -\rho_{12}\right) \\ &\quad + \Phi_2\left(\beta_1 X_{1i}, \beta_3 X_{3i}, \rho_{13}\right). \end{aligned} \tag{1}$$

2.2. Binary choice model with misclassification

The binary indicator variable is misclassified when some of the true '1's are recorded as '0's and vice versa. Hausman et al. (1998) show that the usual parameters of the binary choice model can be identified consistently together with the two types of misclassification probabilities using MLE if the dependent variable is misclassified randomly. They assume a latent relationship that generates the true indicator variable, $Y_{1i} = 1.\,(Y^*_{1i} > 0)$ where $Y^*_{1i} = \beta_1 X_{1i} + \varepsilon_{1i}$, $\varepsilon_{1i} \sim N(0, 1)$ and two types of (constant) misclassification probabilities, $\alpha_0 = \Pr[(Y_i = 1)\,|\,(Y_{1i} = 0)]$ and $\alpha_1 = \Pr[(Y_i = 0)\,|\,(Y_{1i} = 1)]$ that generates the observed indicator variable, Y_i. As Hausman and colleagues show,

$$E[Y_i|\,X_{1i}] = \Pr[Y = 1_i|\,X_{1i}] = \alpha_0 + (1 - \alpha_0 - \alpha_1)\Phi\left(\beta_1 X_{1i}\right). \tag{2}$$

If we slightly change the notation of Hausman et al. (1998) without changing its structure by defining $\alpha_0 = \Pr[(Y_i = 1)\,|\,(Y_{1i} = 0)]$ and $\alpha_1 = \Pr[(Y_i = 1)\,|\,(Y_{1i} = 1)]$,

$$E[Y_i|\,X_{1i}] = \Pr[Y = 1|\,X_{1i}] = \alpha_0 + (\alpha_1 - \alpha_0)\Phi\left(\beta_1 X_{1i}\right). \tag{3}$$

When two types of misclassification probabilities are covariant-dependent as in Lewbel (2000) and Tennekoon and Rosenman (2016), using a normal link function we can write,

$$\begin{aligned} &E[Y_i|\,X_{1i}, X_{2i}, X_{3i}] \\ &\quad = \Pr[Y = 1_i|\,X_{1i}, X_{2i}, X_{3i}] \\ &\quad = \Phi(\beta_2 X_{2i}) + (\Phi(\beta_3 X_{3i}) - \Phi(\beta_2 X_{2i}))\Phi\left(\beta_1 X_{1i}\right) \\ &\quad = \Phi(\beta_2 X_{2i}) + \Phi(\beta_3 X_{3i})\Phi\left(\beta_1 X_{1i}\right) \\ &\qquad - \Phi(\beta_2 X_{2i})\Phi\left(\beta_1 X_{1i}\right) \\ &\quad = \Phi(\beta_2 X_{2i})\Phi\left(-\beta_1 X_{1i}\right) + \Phi\left(\beta_1 X_{1i}\right)\Phi(\beta_3 X_{3i}). \end{aligned} \tag{4}$$

The two misclassification probabilities in this model can be thought to be generated by a latent process where $\alpha_0 = \Pr[(Y_i = 1)\,|\,(Y_{1i} = 0)] = \Pr\,(\beta_2 X_{2i} + \varepsilon_{2i} > 0)$ and $\alpha_0 = \Pr[(Y_i = 1)\,|\,(Y_{1i}] = 1) = \Pr\,(\beta_3 X_{3i} + \varepsilon_{3i} > 0)$. Allowing the two error terms ε_{2i} and ε_{3i} to be correlated,

$$\begin{aligned} E[Y_i|\,X_{1i}, X_{2i}, X_{3i}] &= \Pr[Y = 1_i|\,X_{1i}, X_{2i}, X_{3i}] \\ &= \Phi_2\left(-\beta_1 X_{1i}, \beta_2 X_{2i}, -\rho_{12}\right) \\ &\quad + \Phi_2\left(\beta_1 X_{1i}, \beta_3 X_{3i}, \rho_{13}\right). \end{aligned} \tag{5}$$

Note that the Lewbel (2000) and Tennekoon and Rosenman (2016) models do not assume correlated errors as in above model. Each bivariate CDF term in (5) breaks to the product of two univariate terms in Tennekoon and Rosenman model. Here, we have enhanced that model allowing for the potential correlation between the error terms.

2.3. Trivariate probit models with partial observability

The bivariate probit model with full observability can be specified as, $Y_{ji} = 1.\,(Y^*_{ji} > 0)$ and $Y^*_{ji} = \beta_j X_{ji} + \varepsilon_{ji}$ for $j = 1,2$. The error terms are assumed to be jointly bivariate normally distributed with potentially correlated errors as, $\begin{bmatrix}\varepsilon_{1i}\\ \varepsilon_{2i}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\rho_{12} & 1\\ 1 & \rho_{12}\end{bmatrix}\right)$. There are four potential outcomes given by $(Y_{1i}, Y_{2i}) = (0, 0), (0, 1), (1, 0), (1, 1)$.

When one or more of above outcomes are not observed we have a bivariate probit model with partial observability. When the outcome $(Y_{1i}, Y_{2i}) = (0, 1)$ is not observed (or not possible), we have the bivariate selection model. When both outcomes $(Y_{1i}, Y_{2i}) = (0, 1)$ and $(Y_{1i}, Y_{2i}) = (1, 0)$ are not observed, we have the partial observability model discussed in Poirier (1980). In a similar manner, we can define trivariate probit models with partial observability.