Economic Modelling 52 (2016) 507–518

Contents lists available at ScienceDirect

Economic Modelling

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

Identification and estimation of endogenous selection models in the presence of misclassification errors☆

Ji-Liang Shiu

Hanqing Advanced Institute of Economics and Finance, Renmin University of China, Beijing 100872, PR China

ARTICLE INFO

Article history:
Accepted 23 September 2015
Available online 24 October 2015

Keywords:
Sample selection
Endogeneity
Non-classical measurement errors
Instrumental variables
Semi-parametric identification

ABSTRACT

This paper shows the semi-parametric identification and estimation of sample selection models when the primary equation contains a discrete mismeasured endogenous covariate. Assuming that appropriate instruments for the presence of endogeneity are available, I apply a control function approach to remove the possible endogeneity. Based on the conditional mean independence between the model error and the selection error, the model can be regarded as a semi-parametric regression model with a discrete mismeasured covariate, thereby permitting a non-classical measurement error. Additional identification assumptions include monotonicity restrictions on the regression function and an empirical testable rank condition. I then use the identification result to construct a sieve maximum likelihood estimation estimator to estimate the model parameters consistently and recover the selection rule and joint probabilities of the accurately measured endogenous variable and the mismeasured observed variable. The proposed estimation method allows for a rather flexible functional form of the mismeasured endogenous covariate, requires only one valid instrument to control for both endogeneity and measurement errors for the variable of interest, and imposes no distribution assumptions on the selection rule.

1. Introduction

Sample selection models are useful in applied research that must account for non-random sampling. These models not only allow researchers to investigate the selection decisions made by individuals, but they also enable researchers to estimate interesting economics quantities while controlling for sample selectivity.[1] However, the parameter estimators of the sample selection model often lead to inconsistent results when researchers make incorrect distributional assumptions about a selection rule.[2] Therefore, the semi-parametric estimation of sample selection models has received considerable attention. Researchers have presented a number of studies related to the nonparametric and semi-parametric estimation of selection models when the joint distribution of the error terms is of an unknown form (e.g., Das et al., 2003; Lewbel, 2007a; Newey, 2009). See Powell (1994), and Vella (1998) for a review of the estimation of selection models and for more references.

This paper presents the identification and estimation of the semi-parametric models of a discrete mismeasured endogenous covariate subject applied to a semi-parametric selection mechanism and measurement error under weak semi-parametric restrictions on the form of the endogeneity. Specifically, the proposed selection rule depends on the unknown conditional mean of the model error in the selection error, and is restricted to a suitably smooth functional form. Suppose $m(\mathbf{x}_1, y_2^*)$ is an unknown function and $1(\cdot)$ is set as an indicator function. The endogenous sample selection model with a binary selection rule under consideration exhibits the form

$$y_1^* = m(\mathbf{x}_1, y_2^*) + u, \quad \text{(Latent model)} \tag{1}$$

$$y_1 = s_1 y_1^*, \quad \text{(Offer equation)} \tag{2}$$

$$s_1 = 1(\mathbf{x}_d' \theta_d + v > 0) \quad \text{(Participation equation)} \tag{3}$$

where y_1^* is a latent variable that is observed only if $s_1 = 1$. In the model given by Eqs. (1)–(3), $\mathbf{x}_1$ and $\mathbf{x}_d$ are vectors of strictly exogenous explanatory variables, (u,v) are idiosyncratic errors, and y_2^* is a latent discrete mismeasured endogenous variable such that $\mathrm{Cov}(y_2^*, u) \neq 0$. The variable of interest is the latent true discrete variable y_2^* which is subject to misclassification error and endogeneity in the outcome equation. In this case, there are J alternatives, and the endogenous variable y_2^* takes the value $j \in \{1, \ldots, J\}$ if the j-th alternative is taken. Denote a proxy or measure of y_2^* by y_2. The variable $y_2 \in \{1, \ldots, J\}$ is observed and y_2 may be correlated with y_2^*.

☆ Ji-Liang Shiu acknowledges the support from the National Science Council of Taiwan via Grant NSC 100-2410-H-194-070. The author is solely responsible for any remaining error. The author can be reached at jishiu@ruc.edu.cn.

E-mail address: jishiu@ruc.edu.cn.

[1] For a detailed discussion, see Heckman (1974, 1979).

[2] See Arabmazar and Schmidt (1982), and Arabmazar and Schmidt (1981).

http://dx.doi.org/10.1016/j.econmod.2015.09.031
0264-9993/© 2015 Elsevier B.V. All rights reserved.

Similar to the Heckman two-step method, the identifying model given by Eqs. (1)–(3) requires an exclusion restriction on the set of regressors. Thus, at least one component of the regressors $\mathbf{x}_d$ in the participation Eq. (3) does not appear in $\mathbf{x}_1$. The settings in Eqs. (1)–(3) generally include flexible functional forms of the mismeasured endogenous covariate y_2^* in both Latent model (1) and participation Eq. (3).[3] In an empirical application, the specifications of m can contain non-linear relationships such as quadratics or polynomial terms.

1.1. Examples

The setup in Eqs.(1)–(3) is applicable to many empirical cases. I provide four examples in different areas of economics where non-random sampling is relevant and there may exist a discrete mismeasured endogenous covariate. The first example occurs when y_2^* is not always observed and is mismeasured with endogeneity.

Example 1. Female labor supply

Consider an empirical example for estimating the impact of education on the female labor supply, of which the hours worked are observed only for women who participate in the labor force. However, education is available to a woman regardless of whether she is in the workforce. In this case, although endogeneity arises in the presence of an "ability bias," the measurement error indicates the classification errors of education. Borjas (2009) reviews the literature on the estimation of the labor supply elasticity and also discusses the problems caused by measurement error.

Example 2. Wage offer equation

Consider estimating a wage offer equation for married women, accounting for selectivity bias into the workforce. The proposed model is applicable when y_1 is $\log(wage)$ and y_2^* represents the discrete health status or life expectancy quantile. When a self-reported health status is used as the health status, the health status is correlated with unobserved genetic diseases, and the variable is not an objective indicator of health, which leads to a measurement error problem. Contoyannis and Rice (2001) consider the effects of self-assessed health on hourly wage by employing fixed effects and random effect instrumental variable estimators to panel data. However, the fixed-effects estimator does not address an important selection problem in the wage equation, using wage observations for workers but do not use the potential wages of non-workers. The proposed model can take into account the selection bias with concerns about the endogeneity and measurement errors in health status.

Example 3. Health expenditure

The proposed model can be applied to investigate the effect of individual deductible plan on annual health expenditures. While the dependent variable y_1 is log medical expenditures for positive medical expenditures, y_2^* is an indicator for individual deductible plan, one if an individual has a deductible plan and zero otherwise. The model not only takes account of a significant portion of zero health expenditure in the sample but also potential possible endogeneity and measurement error of the discrete self-reported variable y_2^*. Although there is a debate between using sample selection and two-part models to health care expenditure, the method in the paper provides more modeling strategies when a sample selection model is adopted. See Jones (2000) and Madden (2008) for a summary and overview of the debate.

Example 4. R&D subsidies

Consider an empirical application in estimating the effect of government subsidies on R&D investment on firms. While the dependent variable is the amount of spending on R&D at the firm level, the discrete endogenous mismeasured variable is a dummy variable indicating whether a firm received government subsidies. There are sample firms with a net R&D investment equal to zero, so the sample is selected on the basis of R&D investment. The measures of government subsidies may contain measurement errors because it may be hard to summarize as a dummy variable when each firm receives different amounts of subsidies. On the other hand, unobserved factors of receiving government subsidies may cause the dummy variable endogenous. Almus and Czarnitzki (2003) and Busom (2000) investigate the effects of public R&D subsidies by considering potential selection biases coming from the public institutions that decide the recipients of the public funding. The studies do not pay attention to the possible selection bias with endogeneity and measurement error problems.

1.2. Theoretic discussion

Econometric techniques for addressing the problems of both endogeneity and measurement errors in linear errors-in-variables (EIV) models with measurement errors are clearly understood and are widely applied in empirical economics.[4] However, a vast amount of econometric literature has been devoted to the identification and estimation of non-linear EIV models. Reviews of developments on this subject can be found in Carroll et al. (1995), Bound et al. (2001), Ridder and Moffitt (2007), and Chen et al. (2011). Important work on the measurement error in one of the discrete variables in the outcome equation related to the work here is for example, Lewbel (2000), Mahajan (2006), and Lewbel (2007b). Non-linear EIV models are often considered based on additional information, assuming either instruments (Hausman et al., 1991; Hu and Schennach, 2008; Newey, 2001; Schennach, 2007) or repeated measurements (Carroll et al., 2004; Li, 2002; Schennach, 2004), or validation sample (Ridder and Hu, 2004).

However, it is not clear how to extend the existing results to a sample selection model when the primary equation contains a discrete mismeasured endogenous covariate. There has been few studies contended with the identification and estimation of non-linear models with a mismeasured endogenous covariate. One of the challenges to solving this problem is to simultaneously control both endogeneity and measurement errors. Song et al. (2011) studied the identification and estimation of the average marginal effects of endogenous regressors in non-separable models when the regressors are mismeasured using instruments for endogeneity and repeated measurement of the regressors. Hu et al. (forthcoming) and Hu et al. (2015b) provide identification results for the ratio of partial effects in different types of models with measurement error and endogeneity. The identification restrictions include the existence of instruments and independence of covariates and error terms. Without sample restrictions on measurement errors, the proposed approach requires only one set of instruments to control for the endogeneity to obtain identification. The reduction of the sample requirement allows the proposed approach to be employed in more empirical applications.

I use the results presented by Chen et al. (2009), who provided the identification and estimation of a semi-parametric regression model with a mismeasured discrete covariate without an additional data requirement. Assuming that appropriate instruments for the present endogeneity are available, I apply a control function approach to remove the possible endogeneity. Based on the conditional mean independence between the model error and the selection error, the proposed model can be regarded as a semi-parametric regression model with a discrete

[3] The model contains non-linear cases other than a linear case with $m(\mathrm{x}_1, y_2^*) = m(\mathrm{x}_1, y_2^*; \theta) = \mathrm{x}_1\beta_1 + \alpha_1 y_2^*$, and $\mathrm{x}_d'\theta_d = \mathrm{x}_d\beta_d$.

[4] A linear EIV model means it is linear in both the mismeasured variables and parameters of interest. In the measurement error case, a composite error can be formed by collecting the measurement error and model error. Hence, the model can be regarded as a special case of models with endogenous covariates and the method of instrumental variables (IV) provides a consistent estimator.