Economics Letters 117 (2012) 753–757

Contents lists available at SciVerse ScienceDirect

Economics Letters

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

Estimation of a heteroscedastic binary choice model with an endogenous dummy regressor☆

Zhengyu Zhang [a,*], Xiaobo He [b]

[a] School of Public Economics and Administration, Shanghai University of Finance and Economics, China
[b] School of Economics, The University of Adelaide, SA, Australia

ARTICLE INFO

Article history:
Received 1 June 2012
Received in revised form
18 July 2012
Accepted 15 August 2012
Available online 21 August 2012

JEL classification:
C13
C14
C35

Keywords:
Binary choice model
Dummy endogenous variable
Conditional symmetry
Heteroscedasticity

ABSTRACT

Estimating binary choice models with endogeneity is of considerable importance in microeconometrics. The leading control function approach does not apply when the endogenous variable is binary. We propose a multi-stage estimation procedure for a heteroscedastic binary choice model with an endogenous dummy under a joint conditional symmetry restriction, which allows us to overcome several drawbacks associated with the existing estimators.

1. Introduction

Estimating binary choice models with endogeneity is of considerable importance in applied microeconometrics. For example, evaluating the impact of job training upon subsequent employment status often involves estimating a model like

$$y = I\{c_0 + x'\beta_0 + d\alpha_0 > u_1\}, \tag{1.1}$$

$$d = I\{z'\gamma_0 > u_2\}. \tag{1.2}$$

In (1.1)–(1.2), y is the binary indicator of employment status, which relies on a p_x-dimensional vector x of observable characteristics and a binary indicator d of training reception. d is further determined by a p_z-dimensional vector of instruments z through a linear index $z'\gamma_0$. d is generally endogenous due to possible correlation between unobserved error terms u_1 and u_2. c_0 is the intercept term and the slope coefficients (β_0, α_0) are the parameters of central interest.[1] There also exist many other empirical fields where model (1.1)–(1.2) may be useful. For example, the model is used by Loureiro et al. (2010) to examine intergenerational transmission of smoking habits, where y represents the smoking indicator of a teenager and d represents the smoking indicator of his/her father or mother. In their study, parental smoking indicator d is likely to correlate with the error term u_1 because both teens and their parents share unobservable preferences such as attitudes towards risk, health consciousness and genetic traits.

As the maximum likelihood-based method is notorious for typically delivering inconsistent estimate if the parametric form of the error distribution is misspecified, recent literature on semiparametric method focuses on $\sqrt{n}$-consistent estimation of the structural parameters without requiring parametric specification of the error distribution. When the endogenous variable is continuous, one may follow Blundell and Powell (2004) and Rothe (2009) to employ a control function approach to estimate the model. However, the leading control function approach is no longer applicable

☆ This research is supported by National Social Science Foundation (Grant No. 11CJY011) from National Planning Office of Philosophy and Social Science of China.
* Correspondence to: No. 777, Guoding Road, 200433, Shanghai, China. Tel.: +86 021 53060606.
E-mail address: zyzhang@sass.org.cn (Z. Zhang).

[1] Throughout the paper, any parameter with subscript zero denotes the true parameter that generates the data.

0165-1765/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2012.08.023

when the endogenous variable is discrete, thus complicating the estimation to a great degree. Recently, Yildiz (forthcoming) proposes for model (1.1)–(1.2) a multi-stage estimation procedure, that does not rely on any parametric assumption of the error distributions. To our best knowledge, this seems to be the only distribution free estimator that can deal with the current model to date. However, Yildiz's estimation strategy, which employs Powell et al. (1989)'s density weighted average derivative estimate in constructing his first and second stage estimator, heavily relies on the independence between error term and the exogenous regressors, thus essentially ruling out the case with heteroscedastic error terms. While deviation from normality may have serious consequence for commonly used parametric estimators such as MLE, evidence also shows that these estimators are more severely affected by heteroscedasticity of unknown form than by nonnormality.

In this note, we propose an alternative multi-stage estimation procedure that is not only robust against deviation from normality of error distributions, but also allows for general forms of heteroscedasticity. Specifically, instead of assuming that the exogenous regressors are independent of the error terms, we impose a joint conditional symmetry restriction on the error distribution. As will be shown in this note, our estimator enjoys several additional advantages. First, whereas the intercept term is not identified under the independence restriction, our conditional symmetry restriction permits $\sqrt{n}$-consistency estimation not only for the slope coefficients on both exogenous and dummy endogenous regressors, but also for the intercept term as well.[2] Second, except for a mild exclusion restriction, our estimator allows for a general form of heteroscedasticity in the error term.

2. The estimator

Let w be the vector of distinct components in the regressor vector (x, z). Instead of assuming the independence between (u_1, u_2) and w, we assume that the distribution of (u_1, u_2) depends on w only through w_2, a proper subvector of w. We mean by w_2 being a proper subvector of w that there is at least one component of x and z that is not contained in w_2. Here we have imposed a mild exclusion restriction on the form of heteroscedasticity in that heteroscedasticity is only associated with a subset of the regressor vector w. Chamberlain (1992), Chen and Zhou (2010) have adopted a similar heteroscedasticity exclusion restriction. Throughout the paper, we make the following assumption.

Assumption 1a. The joint distribution of (u_1, u_2) relies on w only through w_2, and is assumed to be symmetric around the origin conditional on w_2, that is,

$$f_{u_1u_2}(v_1, v_2|w) = f_{u_1u_2}(v_1, v_2|w_2) = f_{u_1u_2}(-v_1, -v_2|w_2), \qquad (2.1)$$

where $f_{u_1u_2}(\cdot, \cdot|w)$ is the probability density function of (u_1, u_2) conditional on w.

This symmetry condition has been maintained as a common semiparametric restriction by Chen (1999a,b), Honore et al. (1997), Newey (1991) and Powell (1986), to name only a few. As an example, assume for the moment that w_2 is scalar and ϵ is any symmetric distribution independent of w_2. Then (u_1, u_2) with $u_1 = w_2\epsilon$ and $u_2 = 0.5 \cdot u_1$ satisfy (2.1).[3] Note that although (2.1) has imposed a shape restriction on the error distributions relative to the independence assumption required by Yildiz (forthcoming), neither assumption is stronger than the other, as (2.1) does allow the joint distribution of (u_1, u_2) to depend on w_2. Also, there is some evidence (see, e.g. Powell, 1986; Honore et al., 1997) that symmetry-based estimators possess certain robustness to the violation of the symmetry assumption. Overall, we make this tradeoff to free ourselves from the strong requirement of the independence between error term and regressors.

2.1. Stage 1: estimation of γ

The $\sqrt{n}$-consistent estimate of the index coefficients in a single equation binary choice model (e.g. (1.2)) under conditional symmetry restriction is available from Chen (2005), Chen and Zhou (2010). Without loss of generality, we make the following assumption in parallel with Assumption 6 in Chen and Zhou (2010):

Assumption 2. There exists some $\sqrt{n}$-consistent estimator $\widehat{\gamma}_n$ of γ_0, which permits the following asymptotic linear representation:

$$\sqrt{n}\,(\widehat{\gamma}_n - \gamma_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i + o_p(1),$$

for some ψ_i with $E\psi_i = 0$ and $E\|\psi_i\|^2 < \infty$.

2.2. Stage 2: estimation of β

It is well known that for binary choice models the coefficients can be identified only up to scale and the regressors should contain at least one component whose probability distribution conditional on the remaining components is absolutely continuous with respect to the Lebesgue measure. For these reasons, rewrite the outcome Eq. (1.1) with a minor abuse of notation as

$$y = I\{x_0 + c_0 + x_1'\beta_{10} + x_2'\beta_{20} + d\alpha_0 > u_1\}, \qquad (2.2)$$

where x_0 has a continuous probability distribution conditional on the remaining components (x_1, x_2) and the coefficient on x_0 has been normalized to unity for convenience. c_0 is the constant term, x_1 and x_2 are p_{x_1} and p_{x_2}-dimensional regressor vectors respectively such that w_2 contains x_2 as its components with $p_{x_1} + p_{x_2} + 1 = p_x$. Let $\beta_0 = (1, \beta_{10}', \beta_{20}')'$. In the presence of heteroscedasticity in an unknown form, it is straightforward to see that Powell et al. (1989)'s method no longer gives a consistent estimate of the slope coefficients β_0. To overcome this difficulty, we build our identification strategy upon two rank conditions, that generalize the insight from Abrevaya et al. (2010) and exploit the conditional symmetry restriction (2.1). Denote $\lambda_1(a, b, w_2) = Pr(a > u_1, b > u_2|w_2)$ and $\lambda_2(a, b, w_2) = Pr(a > u_1, b \le u_2|w_2)$.

Assumption 1b. Both $\lambda_1(a, b, w_2)$ and $\lambda_2(a, b, w_2)$ are strictly increasing in a for any given b and w_2.

Lemma 1 (*Rank Condition 1*)**.** *Under Assumption* 1b*, for any pair of observations indexed i and j,*

$$E\left(y_i - y_j|w_i, w_j, w_{2i} = w_{2j}, z_i'\gamma_0 = z_j'\gamma_0\right) > 0$$
$$\text{if and only if} \quad \left(x_i - x_j\right)'\beta_0 > 0.$$

Proof. It follows from (1.1)–(1.2) and the assumption that the distribution of (u_1, u_2) relies on w only through w_2 that

$$E(y|w) = Pr\left(x'\beta_0 + c_0 + \alpha_0 > u_1, z'\gamma_0 > u_2|w_2\right) + Pr\left(x'\beta_0 + c_0 > u_1, z'\gamma_0 \le u_2|w_2\right). \qquad (2.3)$$

[2] Identification of the intercept term under symmetry has been mentioned in Section 4.3 of Yildiz (forthcoming). However, Yildiz didn't consider the symmetry conditional on a set of exogenous regressors as in this note.

[3] However, our conditional symmetry assumption is not satisfied when, e.g., (u_1, u_2, w_2) are jointly normal.