Economics Letters 107 (2010) 33–35

Contents lists available at ScienceDirect

Economics Letters

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

Endogenous regressor binary choice models without instruments, with an application to migration

Yingying Dong *

Department of Economics, California State University Fullerton, Fullerton, CA 92834-6848, USA

ARTICLE INFO

Article history:
Received 11 September 2009
Received in revised form 5 December 2009
Accepted 14 December 2009
Available online 22 December 2009

Keywords:
Binary choice model
Endogeneity
Identification
Migration

JEL classification:
C35
J61

ABSTRACT

This paper shows identification of a semiparametric binary choice model containing an endogenous regressor, when no outside instrumental variable is available. A simple estimator, an easy test for endogeneity, and an empirical application to US migration data are provided.

1. Introduction

This paper shows semiparametric identification of a binary choice model having an endogenous regressor without outside instruments. A simple estimator and a test for endogeneity are provided. These results are applied to analyze working age male's migration within the US, where labor income is potentially endogenous. Identification relies on the fact that workers' migration probability is close to linear in age while income is nonlinear. With PSID data I find that income is endogenous and ignoring this endogeneity leads to downward bias in the estimated effect of income on migration probabilities.

2. The Model

Consider a binary choice model

$$D = I(\alpha + X'\beta + Y\gamma + \varepsilon \geq 0), \tag{1}$$

where $I(\cdot)$ is one if its argument is true and zero otherwise; D is a dummy dependent variable; ε is a mean zero error with a possibly unknown distribution; X is a vector of exogenous regressors; and Y is an endogenous or mismeasured regressor. How Y is determined is unknown, so the Y model is nonparametric. Let $G(X) = E(Y|X)$ for some unknown function G and define the error $U = Y - G(X)$, which has an unknown distribution. Then

$$Y = G(X) + U. \tag{2}$$

U could be heteroscedastic or otherwise depend on X in unknown ways. Endogeneity of Y comes from correlation between ε and U.

If an element of β is zero (an exclusion restriction), the corresponding covariate in X would be an instrument. Special cases of this model where G is parametric with an instrument include Newey (1987) and Rivers and Vuong (1988). Newey, Powell, and Vella (1999), Blundell and Powell (2004), and Rothe (forthcoming) are more general, except that they require an instrument.

This paper generalizes identification based on functional form, by showing that model (1) is identified without an exclusion, even if the function G and the distributions of ε and U are unknown. Identification arises from nonlinearity in the unknown function G.

Identification based on outside instruments is generally preferable. However, instruments are sometimes difficult to find, so it is useful to know when identification is possible without instruments and to be able to test for endogeneity in the absence of instruments.

3. Identification

Assume

$$\varepsilon = \lambda U + V, \tag{3}$$

* Tel.: +1 657 278 2053.
E-mail address: ydong@fullerton.edu.

0165-1765/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2009.12.017

where λ is some unknown constant and the error V is independent of U and X. Equation (3) holds when ε and U are jointly normal with $\lambda = E(\varepsilon U)/E(U^2)$ and $V = \varepsilon - \lambda U$. It could also follow from economic theory, e.g., a decision D that follows a decision Y depends on the unobservables determining Y plus new shocks.

Theorem. *Assume n independently, identically distributed observations of Y, D, and X, with $n \to \infty$. Equations.* (1), (2), and (3) *hold. The function G and the distribution functions of V and U given X may be unknown. E(Y|X) exists. U|X has a continuous mean zero distribution with whole real line support. V has a continuous mean zero distribution independent of U and X. $E(\tilde{X}\tilde{X}')$ exists and is nonsingular for $\tilde{X} = [1, X', G(X)]'$. Either $\lambda + \gamma \neq 0$ or the distribution of V is known. Then α, β, γ, the function G(X), and the distributions of U, V and ε are identified.*

See the Appendix for a proof. Nonsingular $E(\tilde{X}\tilde{X}')$ requires $G(X)$ to be nonlinear in X. The assumption $\lambda + \gamma \neq 0$ is testable because $\lambda + \gamma = 0$ if and only if $E(D|X,Y) = E(D|X)$, which is easily tested. This theorem identifies the entire model; therefore, any features of the model, for example, choice probabilities and marginal effects of X and Y, are also identified.

To see that identification fails when $G(X)$ is linear in X, and U and ε are normal, substitute $G(X) + U$ for Y in equation (1) and rewrite it as $D = I(\alpha + X'\beta + G(x)\gamma + U\gamma + \varepsilon \geq 0)$. When $G(X)$ is a linear function, for any value of γ, there are always corresponding α and β that give the same index function plus a standard normal error.

4. Estimation and Testing

I adopt control function based estimators to show the application of these identification results. First estimate G using a kernel regression and obtain $\hat{U}$ by,

$$\hat{U}_i = Y_i - \frac{\sum_{j=1}^{n} K\left(\frac{X_j - X_i}{h}\right) Y_j}{\sum_{j=1}^{n} K\left(\frac{X_j - X_i}{h}\right)} \quad \text{for } i = 1, \ldots, n, \tag{4}$$

where K is a kernel function and h is a bandwidth parameter. Then substitute $\hat{U}$ into the D equation, and semiparametrically estimate the endogeneity corrected binary choice model,

$$D = I(\alpha + X'\beta + Y\gamma + \hat{U}\lambda + V \geq 0). \tag{5}$$

Any estimator that would be consistent for a binary choice model under the assumption that the error V is independent of the covariates X, Y, and $\hat{U}$ can be applied here. I use Klein and Spady (1993) (hereafter KS) with covariates 1, X, Y, and $\hat{U}$. For comparison I also estimate Eq. (5) as an ordinary probit of D with these covariates. These estimators assume $V \perp X$, but allow higher moments of U to depend on X in unknown ways.

This is a standard semiparametric two-step estimator, so generic consistency and limiting distribution theory follow from Newey and McFadden (1994), with bootstrap theory from Theorem B in Chen, Linton, and Van Keilegom (2003).

To test for endogeneity, look at the t-statistic for λ. By Equation (3), $\lambda = 0$ under the null hypothesis of no endogeneity. One does not have to account for the first stage estimation error of $\hat{U}$ to perform this test, because under the null $\hat{U}$ drops out of the model (see Newey and McFadden, 1994, Theorem 6.2). For example, when Equation (5) is estimated using probit, the t-statistic from the probit estimation itself provides valid inference for testing if $\lambda = 0$.

5. Empirical Application

The sample, drawn from the 1990 wave of the Panel Study of Income Dynamics (PSID), consists of non-student male household heads, age 22 to 69, with positive labor income during 1989–90. The top 1% highest earning individuals are dropped to reduce the impact of outliers.

Let $D = 1$ if an individual changes residence state in the US during 1991–93, and 0 otherwise. The sample has 4582 observations, with 796 having $D = 1$. Y is logged average labor income before moving (1989–90), and X consists of age, a college dummy, log family size, and number of previously occupied states. Y is potentially endogenous in this case. However, exclusion-based instruments are hard to justify, since almost anything affecting wages may also affect expected wage gains and hence the decision to move.

Existing research shows that migration probabilities decrease nearly linearly with age among working people. For example, Burda (1993) shows "age is strongly negatively associated with the desire to migrate (quadratic terms were insignificant)." This is also consistent with the human capital theory of migration: workers migrate to maximize expected earnings; the older an individual is the shorter his remaining working life, and hence the lower the expected present value of his wage gains from moving. In contrast, income is generally found to be nonlinear in age. The underlying theory can be traced back to Mincer (1974).

To check the identifying assumptions, I non-parametrically regress the migration dummy and log labor income on all the covariates. Fig. 1 shows the nonparametric impacts of age on the migration probability and on log labor income, holding the other covariates fixed at their means. As can be seen, the age profile of migration is close to linear while the age profile of income has an inverse-U shape. This nonlinearity in G suffices for identification, even if it has unknown form, and even if the joint error distribution in the labor income and migration equations are also unknown.

Table 1 shows estimated coefficients from three different estimators for Eq. (5): probit assuming exogenous income ($\lambda = 0$), probit with $\lambda \neq 0$, and KS with $\lambda \neq 0$. KS only identifies coefficients up to location and scale, so unlike the probits, the number of states coefficient is normalized to one in KS. The last row of Table 1 reports the probability density at the index mean, $f(\overline{X}'\beta)$, which when multiplied by the coefficients gives the mean marginal effects. Marginal effects are invariant to scaling and so are comparable across specifications.

As expected, age has a significantly negative effect. Adding a quadratic term of age to the migration equation does not produce a significant coefficient. $\hat{U}$ has a positive, significant effect, showing income is endogenous. The marginal effects of $\hat{U}$ by two-step probit and KS are 0.447 and 0.582, respectively, implying that unobservables (such as personality traits), that increase earnings also increase migration propensity *ceteris paribus*. The marginal effect of labor income in the simple probit is −0.319, in contrast to −0.729 and −0.893 in the two-step probit and KS, so ignoring income endogeneity leads to underestimation of income effects on migration. The similarity between the two-step probit and KS estimates suggests that, after controlling for endogeneity,

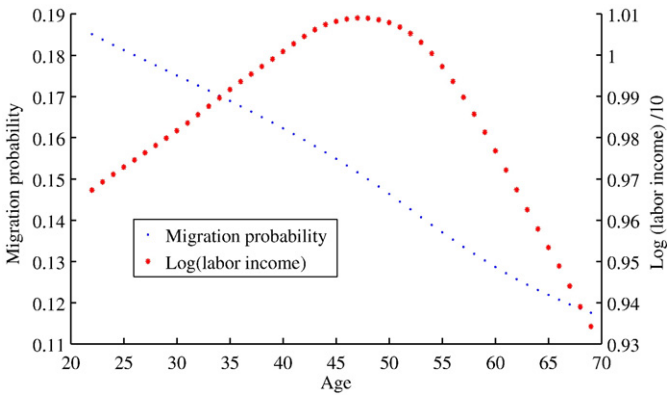

Fig. 1. Nonparametric age effects on migration probabilities and on labor income.

Table 1
Migration binary choice model estimates.

	Probit (I)	Kernel reg. — Probit (II)	Kernel reg. — KS (III)
Constant	0.720 (0.235)***	2.260 (0.973)***	
Age	−0.154 (0.021)***	−0.154 (0.021)***	−0.245 (0.061)***
College education	−0.024 (0.046)	0.062 (0.067)	0.112 (0.103)
Log (Family size)	0.013 (0.044)	0.060 (0.056)	0.131 (0.085)
# of states lived in	0.813 (0.141)***	0.801 (0.145)***	1.000† (0.000)
Log (labor income)	−1.272 (0.235)***	−2.887 (1.020)***	−5.382 (1.623)***
$\hat{U}$		1.779 (1.067)*	3.506 (1.028)***
$f(X'\beta)$	0.251 (0.006)***	0.251 (0.006)***	0.166 (0.162)

Note: †The coefficient of # of states lived in is normalized to one in the Kernel Regression – KS estimation. Bootstrapped standard errors are in parentheses. *** Significant at the 1% level; * Significant at the 10% level.

normality is a reasonable approximation for the latent error in the migration equation.

6. Conclusions

This paper shows the identification of a binary choice model having an endogenous regressor without relying on outside instruments. Based on this identification, the model is estimated using a simple control function approach, which has a nonparametric regression first step and parametric or semiparametric binary choice estimation second step. The first step error is used as an additional covariate in the second step. The ordinary t-statistic for this added covariate provides a test for the endogeneity of the suspected regressor.

I apply this estimator to analyze migration within the US among working age people. Labor income before migration is potentially endogenous and no appropriate instrument is available. Identification of this binary choice migration model relies on the fact that the migration probability among workers is close to linear in age while labor income is nonlinear in age. Reasonable estimates are obtained due to the sufficient non-linearity of the first stage income regression. Adopting both parametric estimation (probit) or semiparametric estimation (the Klein–Spady estimator) in the second stage, I find evidence that labor income is endogenous to the migration choice and that ignoring this endogeneity leads to underestimating the effect of labor income on the migration probability.

Acknowledgments

The author would like to thank Arthur Lewbel, Zhijie Xiao, Peter Gottschalk, and Shannon Seitz for very helpful advice and assistance with data.

Appendix A. Proof Sketch

The full proof of the Theorem (along with data construction and other details) can be found in a supplement to this paper at http://business.fullerton.edu/Economics/ydong/Research/BinaryendogSupplementSep11%20-%20YDong.pdf. The following is an outline of the proof.

$G(X)$, U, and $E(D|X,U)$ are identified by construction given Y, D, and X, so the function $H(U)=E(D|X=0,U)=F[\alpha+G(0)\gamma+(\lambda+\gamma)U]$ is identified, where F is the distribution function of $-V$. Define $Z=H^{-1}[E(D|X,U=0)]$. Then Z is identified and $Z=(\lambda+\gamma)^{-1}(X'\beta+G(X)\gamma-G(0)\gamma)$. Linearly projecting Z on X, $G(X)$, and 1 identifies the scaled coefficients $(\lambda+\gamma)^{-1}\beta$, $(\lambda+\gamma)^{-1}\gamma$, and $(\lambda+\gamma)^{-1}G(0)\gamma$. Plug Z into the model D to get $D=I[(\lambda+\gamma)Z+G(0)\gamma+\alpha+V\geq 0]$. $E(D|Z)$ is identified and is the distribution function of $\tilde{V}=-(\lambda+\gamma)^{-1}(G(0)\gamma+\alpha+V)$. The first two moments of this identified distribution function are $-(\lambda+\gamma)^{-1}(G(0)\gamma+\alpha)$ and $(\lambda+\gamma)^{-2}$, which along with the above scaled coefficients identifies β, γ, λ, and α. The distribution of V is then identified from the distribution of $\tilde{V}$. If instead $\lambda+\gamma=0$ then $E(D|X)=F[\alpha+X'\beta+G(X)\gamma]$ and linearly projecting $F^{-1}[E(D|X)]$ on 1, X, and $G(X)$ identifies α, β, γ, and $\lambda=-\gamma$.

References

Blundell, R.W., Powell, J.L., 2004. Endogeneity in semiparametric binary response models. Review of Economic Studies 71, 655–679.
Burda, C.M., 1993. The determinants of East–West German migration,. European Economic Review 37, 452–461.
Chen, X., Linton, O., Van Keilegom, I., 2003. Estimation of semiparametric models when the criterion function is not smooth. Econometrica 71, 1591–1608.
Klein, R., Spady, R.H., 1993. An efficient semiparametric estimator for binary response models. Econometrica 61, 387–421.
Mincer, J., 1974. Schooling, Experience and Earnings. National Bureau of Economic Research, New York.
Newey, Whitney K., 1987. Efficient estimation of limited dependent variable models with endogenous explanatory variables. Journal of Econometrics 36, 231–250.
Newey, W.K., McFadden, D., 1994. In: Engle, R.F., McFadden, D.L. (Eds.), Large Sample Estimation and Hypothesis Testing. Handbook of Econometrics, vol. iv. Elsevier, Amsterdam, pp. 2111–2245.
Newey, W.K., Powell, J.L., Vella, F., 1999. Nonparametric estimation of triangular simultaneous equations models. Econometrica 67, 565–603.
Rivers, D., Vuong, Q.H., 1988. Limited information estimators and exogeneity tests for simultaneous probit models. Journal of Econometrics 39, 347–366.
Rothe, C., 2009. "Semiparametric estimation of binary response models with endogenous regressors". Journal of Econometrics 153, 51–64.