Endogenous Regressor Binary Choice Models Without Instruments, With an Application to Migration

Yingying Dong[*][†]

Department of Economics
Boston College

March 2009

Abstract

This paper shows the semiparametric identification of a binary choice model having an endogenous regressor without relying on outside instruments. A simple estimator and a test for endogeneity are provided based on this identification. 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 the migration probability among workers is close to linear in age while labor income is nonlinear in age. Using data from the PSID, this study finds that labor income is endogenous and that ignoring this endogeneity leads to downward bias in the estimated effect of labor income on the migration probability.

JEL Codes: C35, J61
Keywords: Binary choice model, Endogeneity, Migration

[*] Correspondence: Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA. Email: dongyi@bc.edu, http://www2.bc.edu/~dongyi/

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

1 Introduction

This paper proves identification of a semiparametric binary choice model that contains an endogenous or mismeasured regressor, when no outside instrumental variable is available. A corresponding estimator, an easy test for endogeneity, and an application to migration in the United States are also provided. This paper generalizes identification based on functional form, showing identification even when the error distributions and the endogenous regressor have unknown functional forms.

In practice, identification that relies upon instruments is generally preferable to identification based on functional form. 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.

2 The Model

Consider a binary choice model

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

where $I(\cdot)$ is an indicator function that equals one if its argument is true, and zero otherwise; D is a dummy dependent variable; ε is a mean zero error term with a possibly unknown distribution; X is a vector of exogenous regressors; and Y is a potentially endogenous or mismeasured regressor. The goal is to identify the parameters α, β, and γ along with the error distribution. Any features of the model, such as choice probabilities and marginal effects of covariates, can also be calculated.

Here equation (1) is assumed to be the equation of interest, which could come from economic theory, whereas how Y is determined is unknown or left as flexible as possible.

Therefore, the model for Y is nonparametric. Let $G(X) = E(Y \mid X)$ for some possibly unknown function G and define $U = Y - G(X)$, then

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

where U is an error term with a possibly unknown distribution and $E(U \mid X) = 0$. U could be heteroscedastic or otherwise depend on X. The errors ε and U are assumed to be correlated, resulting in endogeneity of Y in the binary choice equation.

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

In this paper I show model (1) is generally semiparametrically identified without an exclusion, even if the function G and the distributions of ε and U are unknown. Identification arises primarily from nonlinearity in the unknown function G.

3 Identification

Assume the endogeneity of Y can be written as

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

where λ is some unknown constant that determines the extent of correlation between ε and Y, and V is a mean zero error term that is independent of U and X. This form of endogeneity is not uncommon. For example, equation (3) always holds when ε and U are jointly normally distributed, by letting $\lambda = E(\varepsilon U) / E(U^2)$ and $V = \varepsilon - \lambda U$. It could also follow directly from economic theories that entail triangular systems. For example, Y could represent a precondition or first-stage decision and D, a follow-up decision in which the error term equals the unobservables determining Y plus new shocks.

ASSUMPTION A1: The joint distribution of Y, D, and X is identified, and $E(Y \mid X)$ exists. Equations (1), (2), and (3) hold. Conditional on X, the error U has a continuous mean zero distribution with a support on the whole real line. The error V has a continuous mean zero distribution and is independent of U and X.

Having the joint distribution of Y, D, and X be identified is easily satisfied by assuming that we have n independently, identically distributed observations on these variables, with $n \rightarrow \infty$.

Continuity of U generally holds if Y is continuously distributed. The assumption that U has a real line support is satisfied if, for example, U is normal, and generally holds if Y can take on any value. This assumption can be relaxed, but identification will still require U to have a large support as described in the Appendix.

ASSUMPTION A2: $E\left(\widetilde{X}\widetilde{X}'\right)$ exists and is nonsingular for $\widetilde{X} = [1, X', G(X)]'$. The function G and the distribution functions of V and U given X may be unknown. Either $\lambda + \gamma \neq 0$ or the distribution function of V is known.

Assumption A2 is the main identifying assumption. The assumption that $E\left(\widetilde{X}\widetilde{X}'\right)$ exists and is nonsingular means that if we had a linear regression model where the regressors were $G(X)$ and X, then the regression would not suffer from perfect collinearity. That is, $G(X)$ must be a nonlinear function of X. Given that the index function for D is linear in X, for identification, $G(X)$ could be quadratic in one or more elements of X, say W. However, if W^2 also appears in X, then $G(X)$ would have to be some function of W other than a linear or quadratic form.

The assumption that $\lambda + \gamma \neq 0$ is testable because $\lambda + \gamma = 0$ if and only if $E(D \mid X, Y) = E(D \mid X)$. Many tests of whether a variable belongs to a nonparametric regression like this exist. A relatively early example is Lewbel (1995).

Finally, I assume V has variance one. This assumption is without loss of generality because the coefficients and error term in binary choice models are identified up to an arbitrary positive constant; i.e., the model is unchanged if the error term and all coefficients are scaled by any positive number.

THEOREM: Given Assumptions A1 and A2, α, β, γ, the function $G(X)$, and the distributions of U and V (and hence of ε) are all semiparametrically identified.

This theorem (proved in the Appendix) provides identification of the entire model. Therefore, any features of the model, for example, 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; i.e., different combinations of α, β, and γ can yield the same probit model for D, so the coefficients are not uniquely determined.

4 Estimation and Testing

Assume we have n independent and identically distributed observations of X, Y, and D. Given Theorem 1, existing estimators for binary choice models with endogenous regressors when there are exclusion restrictions can generally be applied in our case. Here I adopt control function based estimators to show the application of our identification results.

First estimate G using a kernel regression and obtain $\widehat{U}$ by,

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

where K is a kernel function and h is a bandwidth parameter. Note this kernel estimator can be used with both discrete and continuous X data (Li and Racine, 2004). Then plug $\widehat{U}$ in the D equation, and estimate the endogeneity corrected binary choice model,

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

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 $\widehat{U}$ can be applied here. This paper estimates equation (5) both parametrically and semiparametrically. The parametric approach is simply to estimate an ordinary probit of D with covariates 1, X, Y, and $\widehat{U}$. The semiparametric approach adopts the efficient estimator of Klein and Spady (1993), which makes no distribution assumption on V. The Klein and Spady estimator (KS) is essentially a maximum likelihood estimator where the probability distribution function is estimated using a nonparametric regression. KS does not identify the location α and requires a scale normalization on coefficients rather than the error variance. If desired, the location and scale based on a mean zero and variance one nonparametric error could be recovered using Lewbel (1997).

These are two-step estimators of a finite number of parameters with a nonparametric first step. Consistency of the estimates follows from identification, uniform consistency of kernel regressions, and consistency of KS or probit. Regularity conditions for root n asymptotically normal asymptotic limiting distributions with these types of estimators are provided by chapter 8 of Newey and McFadden (1994). This estimator is numerically simple and fast, so bootstrapping for standard errors is computationally feasible.

Theorem B in Chen, Linton, and Van Keilegom (2003) provides sufficient regularity conditions for bootstrapping two-step estimators with a nonparametric first step.

To test for endogeneity, one may just look at the ordinary t-statistic for λ. By equation (3), $\lambda = 0$ under the null hypothesis of no endogeneity. Unlike other inferences, one does not have to account for the first stage estimation error of $\widehat{U}$ to perform this test, because under the null hypothesis $\widehat{U}$ drops out of the model (see Newey and McFadden (1994), sections 6 and 8, specifically Theorem 6.2). For example, when we estimate equation (5) using probit, the t-statistic from the probit estimation itself provides valid inference for testing if $\lambda = 0$.

5 Empirical Application

Let D indicate if a worker moves (migrates) to another state in the United States. Labor income Y is likely to be endogenous in the migration model, and it is difficult to obtain a suitable instrument because anything that affects wages may also affect expected wage gains and hence the decision to move. Existing research shows that migration probabilities decrease with age and the effect is linear or near linear among working age 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 discounted value of his wage gains that might result from moving. In contrast, income itself is generally found to be nonlinear in age. The underlying theory can be traced back to Mincer (1974). 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.

This paper draws a sample from the 1990 wave of the Panel Study of Income Dynamics (PSID) data. The analysis focuses on male household heads who are not students and who have positive labor income during 1989 - 1990. The top 1% highest earning individuals are dropped to reduce the impact of outliers. The analysis is further restricted to those 22 to 69 years old, consistent with a downward migration-age profile. These restrictions yield a sample of 4,582 observations.

To avoid having a sample that is too unbalanced (a small share of migrants), D is based on a three-year migration probability; i.e., D = 1 if an individual changes his state of residence during 1991 - 1993, and 0 otherwise. There are 796 migrants in our sample. Y is defined as the logarithm of average annual labor income in 1990 and 1989. The vector X includes age (22 - 69), a dummy indicating college or above education (0, 1), the logarithm of family size (1 - 17), and the number of states an individual ever lived in (1 - 8). Age, the number of states, and the logarithm of annual labor income are divided by 10 to facilitate estimation. Unlike some existing studies, homeownership is not included due to its potential endogeneity in both the income and migration equations, though admittedly homeownership could be an important predictor for migration.

To check the identifying assumptions, I non-parametrically regress the migration dummy and the logarithm of labor income on all the covariates. Figure 1 shows the nonparametric impacts of age on the migration probability and on the logarithm of labor income, holding the other covariates fixed at their means. As can be seen, the migration-age profile is close to linear while the income-age profile has an inverse-U shape. This nonlinearity of the labor income in age therefore suffices for identification.

Considering that the monetary cost of migration might complicate the migration-income relationship, I experimented with limiting the sample to individuals who would not be deterred from moving by the cost, in particular, those whose household income is above the poverty threshold. This did not change the estimation results much, possibly

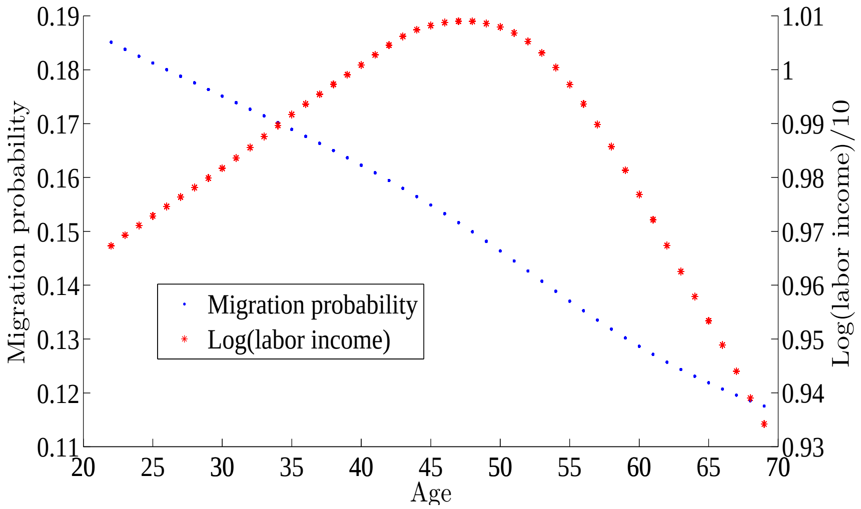

Figure 1 Nonparametric age effects on migration probabilities and on labor income

because this paper focuses on workers.

The estimated coefficients are presented in Table 1. These estimates are based on three different specifications: a simple probit assuming labor income is exogenous and two endogenous income two-step estimators (kernel regression-probit and kernel regression-KS). The bandwidth choice for the high-dimensional first stage kernel regression is obtained by cross-validation and for the one-dimensional KS estimator by Silverman's rule. Since KS can only identify coefficients up to location and scale, the coefficient of the number of states is normalized to one. Note this is different from the scaling of the probit, so the estimated coefficients in the probit (I and II in Table 1) need to be divided by the coefficient of the number of states to be comparable with the KS estimates. Further, the last row of Table 1 reports the probability density at the index mean ($f(\bar{X}'\beta)$), which when multiplied by the coefficients gives the marginal effects at the mean. 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, which further confirms that, conditional on the other covariates in the model, age has a linear or near linear impact on the migration probability.

The first step error ($\hat{U}$) has a positive and significant effect on the migration probability. Based on the estimated coefficients and density, the marginal effects of $\hat{U}$ in the two-step probit and KS are 0.447 and 0.582, respectively. Although different in size and level of significance, both estimates suggest that labor income (Y) is endogenous in the migration choice (D) equation and unobservables, such as personality traits, that result in higher earnings also increase migration propensity *ceteris paribus*. The marginal effects of labor income in the simple probit is -0.319, in contrast to -0.729 in the two-step probit and -0.893 in the KS. Therefore, ignoring the endogeneity of labor income leads to underestimation of the income effect on the migration probability. In addition, the similarity between the two-step probit and KS estimates suggest that after controlling for endogeneity, normality is a reasonable approximation for the true distribution of the latent error in the migration equation.

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)***
$f(\bar{X}'\beta)$	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 the parenthesis. *** Significant at the 1% level; * Significant at the 10% level.

6 Conclusions

This paper shows the identification of a binary choice model having an endogenous regressor without relying on outside instruments or exclusion restrictions. A simple control function type estimator is provided based on this identification, 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 nonparametric 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 downward bias in the estimated effect of labor income on the migration probability.

7 Appendix

PROOF OF THEOREM:

$G(X)$, U, and $E(D \mid X, U)$ are identified by construction. Given independence of V, $E(D \mid X, U) = F[\alpha + X'\beta + G(X)\gamma + (\lambda + \gamma)U]$, where F is the distribution function of $-V$. Define $H(U) = E(D \mid X = 0, U) = F[\alpha + G(0)\gamma + (\lambda + \gamma)U]$, then H is identified. Continuity of V implies that F and hence H are differentiable and strictly monotonic, so $dH\left(U\right)/dU$ is identified.

Case 1: $\lambda+\gamma\neq 0$. That is, $dH\left(U\right)/dU$ is not zero everywhere. Without loss of generality, assume $\lambda+\gamma$ is positive ($dH\left(U\right)/dU>0$); otherwise, one can always replace Y with $-Y$ to make $(\lambda+\gamma)>0$. Define $Z=H^{-1}\left[E\left(D\mid X,U=0\right)\right]$. By monotonicity of H and the real line support of $U\mid X$, the function H^{-1} exists and is identified over the real line; i.e., Z is identified and

$$Z=(\lambda+\gamma)^{-1}(X'\beta+G(X)\gamma-G(0)\gamma). \tag{6}$$

$(\lambda+\gamma)^{-1}\beta$, $(\lambda+\gamma)^{-1}\gamma$, and $(\lambda+\gamma)^{-1}G(0)\gamma$ in equation (6) can then be identified by linearly projecting Z on X, $G\left(X\right)$, and 1. Then plug equation (6) into the model D to get $D=I\left[\left(\lambda+\gamma\right)Z+G\left(0\right)\gamma+\alpha+V\geq 0\right]$. $E\left(D\mid Z\right)$ is identified and is the distribution function of $\widetilde{V}=-\left(\lambda+\gamma\right)^{-1}\left(G(0)\gamma+\alpha+V\right)$. The first two moments of this identified distribution function are $-\left(\lambda+\gamma\right)^{-1}\left(G(0)\gamma+\alpha\right)$ and $\left(\lambda+\gamma\right)^{-2}$, which along with the coefficients in equation (6) identify β, γ, λ, and α. The distribution of V is then identified, because $1-E\left(D\mid\alpha+X'\beta+G\left(X\right)\gamma+\left(\lambda+\gamma\right)U=-v\right)$ is the distribution function of V evaluated at v. Given a whole real line or otherwise large enough support of U so that v can take on any value of V, the distribution is fully identified.

Case 2: $\lambda+\gamma=0$. In this case, $D=I(\alpha+X'\beta+G(X)\gamma+V\geq 0)$, $E\left(D\mid X\right)=F\left[\alpha+X'\beta+G\left(X\right)\gamma\right]$, and $F^{-1}\left[E\left(D\mid X\right)\right]=\alpha+X'\beta+G\left(X\right)\gamma$. The distribution function of $-V$, F, is assumed known, so $F^{-1}\left[E\left(D\mid X\right)\right]$ is identified. Linearly projecting $F^{-1}\left[E\left(D\mid X\right)\right]$ on 1, X, and $G\left(X\right)$ then identifies α, β, γ, and $\lambda=-\gamma$.

Lastly, the distribution of ε is identified because $\varepsilon=\lambda U+V$ with $U\perp V$, and the above analysis has shown that λ as well as the distributions of U and V are identified.

References

Blundell, R.W. and J.L. Powell, 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., O. Linton, and I. Van Keilegom, 2003, Estimation of Semiparametric Models when the Criterion Function Is Not Smooth, Econometrica, 71, 1591-1608.

Klein, R. and R. H. Spady, 1993, An efficient Semiparametric Estimator for Binary Response Models, Econometrica, 61, 387-421.

Lewbel, A., 1995, Consistent nonparametric hypothesis tests with an application to Slutsky symmetry, Journal of Econometrics 67, 379-401.

Lewbel, A., 1997, Semiparametric Estimation of Location and Other Discrete Choice Moments, Econometric Theory 13, 32-51.

Li, Q. and Racine, J., 2004, Nonparametric estimation of regression functions with both categorical and continuous data, Journal of Econometrics 119, 99-130.

Mincer, J., 1974, Schooling, Experience and Earnings, New York: National Bureau of Economic Research.

Newey, Whitney K., 1987, Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables, Journal of Econometrics 36, 231-250.

Newey, W. K. and D. McFadden, 1994, Large Sample Estimation and Hypothesis Testing, in: R.F. Engle and D.L. McFadden, eds., Handbook of Econometrics, Vol. iv (Amsterdam: Elsevier) 2111-2245. .

Newey, W. K., Powell, J. L. and Vella F., 1999, Nonparametric Estimation of Triangular Simultaneous Equations Models, Econometrica 67, 565-603.

Rivers, D. and Q. H. Vuong, 1988, Limited information estimators and exogeneity tests for simultaneous probit models, Journal of Econometrics 39, 347-366.