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

Yingying Dong*

September 2009

This supplemental material includes a full proof of the paper's main theorem, details regarding the construction of data used in the paper, and additional comments and citations.

Proof of the Theorem:

Without loss of generality (since binary choice models with unknown errors are only identified up to scale), assume V has variance one. The joint distribution of Y, D, and X are observed, so their joint distribution is identified, so $G(X)$, U, and $E(D \mid X, U)$ are also identified by construction. $E(D \mid X, U) = F[\alpha + X'\beta + G(X)\gamma + (\lambda + \gamma)U]$ follows from independence of V, 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 and $dH(U)/dU$ are identified.

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

*Correspondence: Department of Economics, SGMH 3360, California State University Fullerton, Fullerton, CA 92834-6848, USA. http://business.fullerton.edu/Economics/ydong/ Email: ydong@fullerton.edu

$U \mid X$, the function H^{-1} exists and is identified over the real line; i.e., Z is identified and

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

$(\lambda + \gamma)^{-1}\beta$, $(\lambda + \gamma)^{-1}\gamma$, and $(\lambda + \gamma)^{-1}G(0)\gamma$ in equation (1) can then be identified by linearly projecting Z on X, $G(X)$, and 1. Then plug equation (1) into the model D to get $D = I[(\lambda + \gamma) Z + G(0)\gamma + \alpha + V \geq 0]$. $E(D \mid Z)$ is identified and is the distribution function of $\widetilde{V} = -(\lambda + \gamma)^{-1}(G(0)\gamma + \alpha + V)$. 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 coefficients in equation (1) identify β, γ, λ, and α. The distribution of V is then identified, because $1 - E(D \mid \alpha + X'\beta + G(X)\gamma + (\lambda + \gamma)U = -v)$ is the distribution function of V evaluated at v.

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

Notes regarding the identification and associated estimator:

The Theorem assumes that $E\left(\widetilde{X}\widetilde{X}'\right)$ exists and is nonsingular for $\widetilde{X} = [1, X', G(X)]'$. This is equivalent to saying that if we had a linear regression model where the regressors were a constant, $G(X)$, and X, then the regression would not suffer from perfect collinearity. This condition implies that $G(X)$ in nonlinear in X.

The Theorem also assumes the conditional distribution of U given X is continuous with support on the whole real line. Continuity of U holds if Y given X is continuously distributed. The assumption that U has a real line support is satisfied if, for example,

U is normal, and more 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 proof.

The Theorem assumes we have n independent and identically distributed observations of X, Y, and D. This simplifies estimation but is stronger than necessary, e.g., the identification theorem only uses the implication that the conditional distribution of Y and D given X is identified.

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).

Based on Theorem 1, estimation only requires a uniformly consistent estimator of $G(X)$ and a consistent estimator of the coefficients in $D = I(\alpha + X'\beta + Y\gamma + U\lambda + V \geq 0)$ where U is observed and V is the unobserved error. Other estimators than the ones I used could have been employed, depending on exactly what regularity or distribution assumptions one makes regarding the error terms. I chose a first step kernel regression since that is a standard nonparametric regression estimator. Note this kernel estimator can be used with both discrete and continuous X data (Li and Racine, 2004). I then estimated the second step using Klein and Spady (1993) because that is a semiparametrically efficient binary choice estimator making no parametric distribution assumptions regarding V. For comparison I used probit, because that is one of the most popular parametric models, and because the assumed equation (3) holds automatically when the errors are normal.

The Klein and Spady estimator is essentially a maximum likelihood estimator where the probability distribution function is estimated using a nonparametric regression. Klein and Spady does not identify the location constant 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).

Notes regarding the data construction and analysis:

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.

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 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.

Supplemental Appendix References

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.