Discussion Paper Series

No.106

The efficient moment estimation of the probit model with an endogenous continuous regressor

Daiji Kawaguchi and Hisahiro Naito

June 2005

Hitotsubashi University Research Unit
for Statistical Analysis in Social Sciences
A 21st-Century COE Program

Institute of Economic Research
Hitotsubashi University
Kunitachi, Tokyo, 186-8603 Japan
http://hi-stat.ier.hit-u.ac.jp/

The efficient moment estimation of the probit model with an endogenous continuous regressor

Daiji Kawaguchi[*]
Faculty of Economics, Hitotsubashi University

Hisahiro Naito[†]
Graduate School of Humanities and Social Sciences
University of Tsukuba

June 29, 2005

Abstract

We propose an efficient moment estimator for the probit model with a continuous endogenous regressor. The estimation can be readily implemented using a standard statistical package that can estimate a non-linear system two-stage least squares (instrumental variable) estimator.

JEL Classification: C25
Keywords: Probit, Continuous endogenous regressor, Moment estimation.

[*] Tel: +81-48-580-8851, Address: Naka 2-1, Kunitachi, Tokyo 186-8601, E-mail: kawaguch@econ.hit-u.ac.jp

[†] Tel: +81-29-853-7432, Address: Tennodai 1-1-1, Tsukuba, Ibaraki, 305-8573, E-mail: naito@dpipe.tsukuba.ac.jp

1 Introduction

Consider the following probit model with a continuous endogenous regressor.

$$
\begin{aligned}
y_1^* &= \alpha y_2 + z_1\beta + u \\
y_1 &= 1 \text{ if } \ y_1^* > 0 \\
y_1 &= 0 \text{ if } \ y_1^* \leq 0 \\
y_2 &= z\gamma + v \\
E(v|z) &= 0 \\
Var(v|z) &= \sigma_v^2 \\
u &= \rho v + e \\
e|z &\sim N(0,1)
\end{aligned}
$$

where z_1 is a $(1 \times k)$ vector of exogenous explanatory variables, y_2 is a continuous explanatory variable and z is a $(1 \times l)$ vector of instrumental variables that includes z_1 as a subset. The system of equations is assumed to be just or over-identified (i.e., $k + 1 \leq l$ and the part of γ that corresponds to the excluded variables includes at least a non-zero element). The probit model with a continuous endogenous regressor is typically estimated by the maximum likelihood method under the assumption of multivariate normality of u and v. However, researchers often feel that this assumption is restrictive and attempt to estimate the above model that does not impose multivariate normality of u and v. Rivers and Vuong (1988) proposed a two-step maximum likelihood estimation that estimates γ by ordinary least squares (OLS) in the first stage and introduces $y_2 - z\hat{\gamma}$ as an additional regressor in the second-stage probit. This approach is widely used by applied researchers due to its simplicity, but the standard errors are not correctly calculated when $\rho \neq 0$.[1] Although the procedure to correct standard errors

[1] Because the estimation error associated with γ does not affect the asymptotic distribution of the second-stage probit estimator under the null of $\rho = 0$, the test of endogeneity of y_2 can be implemented by testing $H_0 : \rho = 0$.

is well established, it involves rather cumbersome matrix algebra.[2]

The purpose of this note is to propose a moment estimator for the above probit model that is efficient in the class of moment estimators. A generalized moment method (GMM) estimator of a probit model with continuous endogenous regressors was originally suggested by Grogger (1990), but Dagenais (1999) and Lucchetti (2002) have shown the inconsistency of the proposed GMM estimator. The proposed estimator in this note is a consistent estimator and correct standard errors are obtained without corrections. It can be readily implemented using a standard statistical package that can estimate a system of non-linear equations by the instrumental variable method.

2 Moment conditions and the optimal instruments

By substitutions, the conditional expectation of y_1 on y_2 and z is given as:

$$E(y_1|y_2, z) = \Phi(\alpha y_2 + z_1\beta + \rho(y_2 - z\gamma)). \tag{1}$$

From this conditional expectation, the residual function

$$r_1(w; \theta) = y_1 - \Phi(\alpha y_2 + z_1\beta + \rho(y_2 - z\gamma)), \tag{2}$$

where $w = [y_1\ y_2\ z]$ and $\theta = [\alpha, \beta, \rho, \gamma]$, is orthogonal to any function of z and y_2. The conditional expectation of y_2 on z is:

$$E(y_2|z) = z\gamma. \tag{3}$$

From the above conditional expectation, the second residual function

$$r_2(w; \theta) = y_2 - z\gamma \tag{4}$$

is orthogonal to any function of z.

[2] This is probably one reason why applied researchers continue to rely on the linear probability model when regressors include an endogenous regressor.

For the above residual functions, the following moment condition holds:

$$E[r_1(w;\theta_0)|z,y_2] = E[r_2(w;\theta_0)|z] = 0, \tag{5}$$

where θ_0 is the true parameter value. Then, the natural question is what combination of z and y_2 should be used as instruments.

The optimal instrument matrix $(2 \times (2+k+l))$ that attains minimum estimator variance under the moment restriction of Eq. (5) is given as:

$$Z^*(z,y_2) \equiv \begin{pmatrix} Z_1^*(z,y_2) \\ Z_2^*(z) \end{pmatrix} \equiv \Omega_0(z,y_2)^{-1} R_0(z,y_2), \tag{6}$$

where

$$\Omega_0(z,y_2) \equiv \begin{pmatrix} Var(r_1|z,y_2) & Cov(r_1,r_2|z) \\ Cov(r_1,r_2|z) & Var(r_2|z) \end{pmatrix} = \begin{pmatrix} E[r_1^2|z,y_2] & 0 \\ 0 & E[r_2^2|z] \end{pmatrix}$$

and

$$R_0(z,y_2) \equiv \begin{pmatrix} E[\nabla_\theta r_1|z,y_2] \\ E[\nabla_\theta r_2|z] \end{pmatrix}.$$

This setting is slightly different from the usual setting of the optimal instrument. It uses different information set for different equations, contrary to the standard optimal instrument (Wooldridge (2001): pp. 439–442), but we can show that the above instrument attains the minimum estimator variance.[3] Using this optimal instrument, we can calculate the efficient moment estimator $\hat{\theta}$ by solving the following equation:

$$\sum_{i=1}^{n} Z^*(z,y_2)' r(w;\hat{\theta}) = \mathbf{0}, \tag{7}$$

where $r(w;\hat{\theta}) \equiv [r_1(w;\hat{\theta}) \;\; r_2(w;\hat{\theta})]'$. The above moment condition in our case is

$$\sum_{i=1}^{n} \begin{pmatrix} \frac{\phi(.)y_2}{\Phi(.)(1-\Phi(.))}(y_1 - \Phi(.)) \\ \frac{\phi(.)}{\Phi(.)(1-\Phi(.))} z_1 (y_1 - \Phi(.)) \\ \frac{\phi(.)}{\Phi(.)(1-\Phi(.))}(y_2 - z\gamma)(y_1 - \Phi(.)) \\ \frac{z}{\sigma_v^2}(y_2 - z\gamma) \end{pmatrix} = \mathbf{0}, \tag{8}$$

[3] The proof is in the appendix for refereeing purposes.

where the arguments of $\phi(.)$ and $\Phi(.)$ are $\alpha y_2 + z_1\beta + \rho(y_2 - z\gamma)$.[4]

The expression of the optimal instrument includes unknown parameter values and it should be estimated by the first-stage estimation. The detailed estimation procedure is explained in the following section.

3 Estimation procedure

The actual procedure for the estimation involves the following steps.

1. Run OLS regression of y_2 on z, keep the residual $\hat{v}$ and calculate $\hat{\sigma}_v^2 = (1/n)\sum_{i=1}^{n}\hat{v}_i^2$.

2. Run probit regression of y_1 on y_2, z_1, and $\hat{v}$ and keep $\hat{\alpha}, \hat{\beta}, \hat{\rho}$. Calculate the predicted value of the linear index as $\hat{ind} = \hat{\alpha}y_2 + z_1\hat{\beta} + \hat{\rho}\hat{v}_2$.

3. Calculate the optimal instruments for the first residual function as $\tilde{y}_2 = -[\phi(\hat{ind})/\{\Phi(\hat{ind})(1-\Phi(\hat{ind}))\}]y_2$, $\tilde{z}_1 = -[\phi(\hat{ind})/\{\Phi(\hat{ind})(1-\Phi(\hat{ind}))\}]z_1$, and $\tilde{v} = -[\phi(\hat{ind})/\{\Phi(\hat{ind})(1-\Phi(\hat{ind}))\}]\hat{v}$. Calculate the optimal instrument for the second residual function as $\tilde{z} = -z/\hat{\sigma}_v^2$.

4. Estimate the system of equations

$$y_1 = \Phi(\alpha y_2 + z_1\beta + \rho(y_2 - z\gamma)) + r_1 \tag{9}$$

$$y_2 = z\gamma + r_2 \tag{10}$$

   using $[\tilde{y}_2\ \tilde{z}_1\ \tilde{v}\ \tilde{z}]$ as instruments for a non-linear instrumental variable (IV) estimation procedure, assuming that r_1 and r_2 are not correlated.[5]

The instruments generated do not affect the asymptotic distribution of estimators, and thus standard errors are correctly calculated. This estimator is an efficient estimator because the optimal instruments are used for the estimation. From a practical viewpoint, all the estimation procedures used

[4] The derivation is in the appendix for refereeing purposes.

[5] This is the system 2SLS estimation in Limdep.

above are readily available in statistical packages such as Limdep, Eviews or TSP. Thus, applied researchers can easily obtain an efficient estimator with correct standard errors.

4 Example: Smoking during pregnancy and family income

We consider estimation of the effect of family income on a pregnant mother's smoking behavior. The structural model for the latent variable is:

$$smoke^* = \beta_0 + \beta_1 motheduc + \beta_2 white + \alpha \log(faminc) + u, \tag{11}$$

$smoke = 1$ if $smoke^* > 0$ and $smoke = 0$ if $smoke^* \leq 0$. The variables *smoke* takes value one if the mother smokes during her pregnancy, *motheduc* is the mother's years of education, *white* is the dummy variable that takes one if the mother is white, $\log(faminc)$ is the natural log of family income. We are interested in α. However, family income may include the mother's income during pregnancy, and $\log(faminc)$ and u may show positive correlation, given that cigarettes are normal goods. To deal with this possible endogeneity, we instrument $\log(faminc)$ by father's years of education, $fatheduc$. The equation for $\log(faminc)$ is:

$$\log(faminc) = \gamma_0 + \gamma_1 motheduc + \gamma_2 white + \gamma_3 fatheduc + v. \tag{12}$$

The above model is estimated by following three methods: (1) probit estimation assuming that family income is exogenous; (2) two-step probit estimation according to Rivers and Vuong (1988); and (3) the moment estimation proposed in this paper. The data set used for estimations is taken from Wooldrige (2001), which was originally taken from Mullahy (1997).

The descriptive statistics of the analysis sample are shown in Table 1.

Column 1 of Table 2 reports the results of probit regression, assuming the exogeneity of $\log(faminc)$. This result implies that higher family income reduces the probability of the mother smoking during her pregnancy; however, this estimate could be upward biased.

Column 2 reports the first-stage OLS result according to the Rivers and Vuong (1988) procedure. The coefficient for father's education is statistically significant and this implies that father's education serves as an instrument for family income. Column 3 reports the result for second-stage probit, which includes the residual of first-stage OLS as an additional explanatory variable. The estimated ρ is 0.61, with standard error of 0.37, which is marginally significant. This positive $\hat{\rho}$ implies that u and v are positively correlated and, after considering this correlation, the estimated coefficient for $\log(faminc)$ is -0.76, which is smaller than the coefficient estimated by probit. The standard errors reported in column 4 of Table 2 are not adjusted for the two-step estimation.

Columns 4 and 5 show results for the moment estimation of Eqs. (9) and (10) using $[\tilde{y}_2\ \tilde{z}_1\ \tilde{v}\ \tilde{z}]$ as instrumental variables. The estimated coefficients are similar to those obtained using the Rivers and Vuong (1988) procedure, but all the standard errors are lower, probably due to the efficiency gain. The estimated ρ is closer to zero and is statistically insignificant. Thus, we cannot reject the null hypothesis that $\log(faminc)$ is exogenous.

All the above estimations were calculated using Limdep 8.0 and can be similarly implemented using any statistical package that allows non-linear system 2SLS (IV) estimation.

References

[1] Dagenais, M.G., 1999. Inconsistency of a proposed nonlinear instrumental variables estimator for probit and logit models with endogenous regressors. *Economics Letters*, 63(1), pp. 19–21.

[2] Grogger, J., 1990. A simple exogeneity test for probit, logit, and Poisson regression models. *Economics Letters*, 33(4), pp. 329–332.

[3] Lucchetti, R., 2002. Inconsistency of naive GMM estimation for QR models with endogenous regressors. *Economics Letters*, 75(2), pp. 197–185.

[4] Mullahy, J., 1997. Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior. *Review of Economics and Statistics*, 79 (4), pp. 596–593.

[5] Rivers, D. and Vuong, Q.H., 1988. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39(3), pp. 347–366.

[6] Wooldridge, J., 2001. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

A Optimality of the proposed instrument

This appendix is attached for the purpose of refereeing, not for publication.

The proposed optimal instrument in (6) is not standard since different information sets are used to calculate the conditional expectation for different residual functions. A use of different information sets for different residual functions is inevitable in the presence of endogenous regressors in a nonlinear model. This note shows that our proposed instrument archives the efficiency bound within a class of GMM estimator despite the use of different information sets for different residual functions.

Let $Z_1(y_{2i}, z)$ and $Z_2(z_i)$ be arbitrary $1 \times L$, where $L \geq (2+l+k)$, instrument vectors and let Z_i' be a $L \times 2$ matrix where $Z_i' = (Z_1(y_{2i}, z_i)' \;\; Z_2(z_i)')$. Also define the 2×1 residual vector r_i as $r(y_{2i}, z_i; \theta) \equiv (r_1(y_{2i}, z_i; \theta) \; r_2(z_i; \theta))$. Let $m(y, z, \theta) = Z_i' r(y_{2_i}, z_i; \theta)$. Note that the covariance matrix of GMM estimator is $(G'\Xi G)^{-1} G'\Xi\Lambda\Xi G(G'\Xi G)^{-1}$ where Ξ is the probability limit of the weighting matrix; $G = E[\frac{\partial m}{\partial \theta}]$; $\Lambda = E[mm']$. Let s^* be $(2+l+k) \times 1$ matrix where $s^* = Z^{*\prime} r$ and $Z^{*\prime}$ is defined in (6). Let s be $(2+l+k) \times 1$

matrix where $s = G'\Xi Z'r$. If we can show that $E[ss^{*\prime}] = G'\Xi G$, then it also proves that $Z^{*\prime}$ attains the minimum variance with a a class of GMM from the Lemma 14.1 of Woodridge (2001). By using the definition of s^* and s, $E[ss^{*\prime}]$ can be calculated as follows:

$$\begin{aligned} E[ss^{*\prime}] &= E[E[G'\Xi Z'rr'Z^* |\ y_2, z]] \\ &= E[G'\Xi Z'E[rr'|y_2, z]Z^*] \\ &= G'\Xi E[Z'E[rr'|y_2, z]\Omega^{-1}R_o] \end{aligned}$$

From the the definition of Ω^{-1} and R_o, the above equations becomes

$$\begin{aligned} &= G'\Xi E[Z' \begin{pmatrix} E[r_1^2|y_2, z] & 0 \\ 0 & E[r_2^2|y_2, z] \end{pmatrix} \\ &\quad \cdot \begin{pmatrix} E[r_1^2|y_2, z] & 0 \\ 0 & E[r_2^2|z] \end{pmatrix}^{-1} \begin{pmatrix} E[\nabla_\theta r_1|z, y_2] \\ E[\nabla_\theta r_2|z] \end{pmatrix}] \\ &= G'\Xi E[Z' \begin{pmatrix} E[\nabla_\theta r_1|z, y_2] \\ E[r_2^2|y_2, z]E[r_2^2|z]^{-1}E[\nabla_\theta r_2|z] \end{pmatrix}] \\ &= G'\Xi E[Z' \begin{pmatrix} E[\nabla_\theta r_1|z, y_2] \\ E[r_2^2|z]E[r_2^2|z]^{-1}E[\nabla_\theta r_2|z] \end{pmatrix}] \\ &= G'\Xi G \end{aligned}$$

Thus, $E[ss^{*\prime}] = G'\Xi G$. From the Lemma 14.1 of Woodridge (2001), Z^* archives a minimum variance among instruments defined by $Z_i' = (Z_1(y_{2i}, z_i)'$ $Z_2(z_i)')$.

B Derivation of the optimal instrument

This appendix includes the derivation of the optimal instruments for refereeing purposes. This appendix is not for publication.

$$\begin{aligned} \Omega_0(z, y_2) \equiv &\begin{pmatrix} E[(y_1 - \Phi(.))^2|z, y_2] & E[(y_1 - \Phi(.))(y_2 - z\gamma)|z] \\ E[(y_1 - \Phi(.))(y_2 - z\gamma)|z] & E[(y_2 - z\gamma)^2|z] \end{pmatrix} \\ &= \begin{pmatrix} \Phi(.)(1 - \Phi(.)) & 0 \\ 0 & \sigma_v^2 \end{pmatrix}, \end{aligned}$$

because $E[(y_1 - \Phi(.))(y_2 - z\gamma)|z] = E[E[(y_1 - \Phi(.))(y_2 - z\gamma)|z, y_2]|z] = E[E[(y_1 - \Phi(.))|z, y_2](y_2 - z\gamma)|z] = 0$.

$$\begin{aligned} R_0(z, y_2) &\equiv \begin{pmatrix} E[\partial r_1/\partial\alpha & \partial r_1/\partial\beta & \partial r_1/\partial\rho & \partial r_1/\partial\gamma|z] \\ E[\partial r_2/\partial\alpha & \partial r_2/\partial\beta & \partial r_2/\partial\rho & \partial r_2/\partial\gamma|z] \end{pmatrix} \\ &= \begin{pmatrix} -\phi(.)y_2 & -\phi(.)z_1 & -\phi(.)(y_2 - z\gamma) & \phi(.)\rho z \\ 0 & 0 & 0 & -z \end{pmatrix}. \end{aligned}$$

Thus, the optimal instrument is:

$$\begin{aligned} Z^*(w) &= \Omega_0(z)^{-1}R_0(z) \\ &= \begin{pmatrix} -\frac{\phi(.)y_2}{\Phi(.)(1-\Phi(.))} & -\frac{\phi(.)z_1}{\Phi(.)(1-\Phi(.))} & -\frac{\phi(.)(y_2-z\gamma)}{\Phi(.)(1-\Phi(.))} & \frac{\phi(.)\rho z}{\Phi(.)(1-\Phi(.))} \\ 0 & 0 & 0 & -z/\sigma_v^2 \end{pmatrix}. \end{aligned}$$

The first, third and fourth columns in the first row are linearly dependent and we place 0 in the fourth column.

Table 1: Descriptive statistics

Variable	Mean	Standard Deviation	Minimum	Maximum
Smoking during the pregnancy	0.16	-	0	1
Family income (1988, $1000)	32.22	17.96	0.5	65
Log (Family income)	3.28	0.72	-0.69	4.17
Mother's years of education	13.13	2.42	2	18
Father's years of education	13.19	2.74	1	18
White	0.84	-	0	1

Note: N=1191.

Table 2: The effect of family income on mother's smoking during the pregnancy

	(1)	(2)	(3)	(4)	(5)
	Probit	Rivers-Voung (1988) Two Step Estimation		Efficient Moment Estimation	
Dependent Variable	Smoking	Log(family income)	Smoking	Log(family income)	Smoking
Log (family income)	-0.17	-	-0.76	-	-0.46
	(0.07)		(0.37)		(0.24)
Mother's education	-0.15	0.07	-0.08	0.07	-0.08
	(0.23)	(0.01)	(0.05)	(0.004)	(0.03)
White	0.23	0.35	0.46	0.35	0.33
	(0.14)	(0.05)	(0.20)	(0.02)	(0.14)
Father's education	-	0.06	-	0.06	-
		(0.01)		(0.003)	
Constant	1.13	1.24	1.99	1.24	1.18
	(0.30)	(0.11)	(0.60)	(0.04)	(0.40)
ρ	-	-	0.61	-	0.34
			(0.37)		(0.25)

Note: N=1191.