IMPROVING THE EFFICIENCY OF PROBIT ESTIMATORS

Andrew Chesher*

Abstract—The efficiency with which coefficients in probit models are estimated is improved by exploiting data on continuous ancillary variates. In this paper the resulting gains in efficiency are examined and illustrative calculations are provided. Extra precision is achieved at the cost of making an extra assumption but this assumption can be tested. It is shown that fully efficient maximum likelihood estimation of the probit model with a continuous ancillary variate can be achieved by a simple two step procedure involving an ordinary least squares and a probit estimation.

Received for publication March 23, 1983. Revision accepted for publication October 18, 1983.

* University of Bristol.

I am grateful to Joanna Gomulka and Tony Lancaster for helpful comments. Errors are my responsibility. The work reported here is part of a project on the Microeconometrics of Labour Market Transitions supported by the Social Science Research Council.

I. Introduction

Probit analysis of binary data is now widely practised by social scientists. Sometimes data are available on endogenous continuous variates which, given exogenous variables, are correlated with the binary variate on which probit analysis is performed. For example, when estimating models for housing tenure choice, household income or expenditure on some nondurable goods may be available. And when estimating a model for the return to work of an unemployed worker in some time interval, tenure of or wage in the previous job may be available. This paper provides a simple computational procedure for using ancillary data of this sort efficiently and a method for obtaining estimates of the asymptotic variances of the resulting estimators. Finally the gain in efficiency obtained by using the ancillary data is investigated.

Suppose that, given a vector x, two variates y_1 and y_2 are bivariate normally distributed:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \Bigg| x \sim N_2\left[\begin{bmatrix} x'\pi_1 \\ x'\pi_2 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \rho\sigma_{11}^{1/2} \\ \rho\sigma_{11}^{1/2} & 1 \end{bmatrix}\right]. \tag{1}$$

Realisations of y_1 are observed but y_2 is not observable. Instead realisations of a binary variate D are obtained[1] where $D = 1$ if $y_2 \geq 0$, $D = 0$ if $y_2 < 0$. This model provides a framework for the analysis of binary data when jointly dependent continuous variates are available and it can be regarded as the reduced form of a simultaneous equations model in binary and continuous variates (see Heckman (1978)). The model also finds application in the sample selection bias area if we attach to it a variate y_3 only observed when $D = 1$. Heckman's (1979) two-step procedure uses an estimate of π_2 to calculate the normal hazard function included as an extra "regressor" in the y_3 equation. In all these contexts more efficient estimation of π_2 is of interest.

With y_2 observed (1) can be written as a pair of seemingly unrelated regression equations with identical regressors and separate ordinary least squares (OLS) estimation of the two equations is efficient. Data on y_1 are not informative about π_2 when y_2 is observed (see Conniffe (1982)). The coarse grouping of y_2 into the two classes indicated by D destroys information but the ordinary least squares (OLS) estimators of π_1 and σ_{11} can still be calculated and are therefore still efficient. Since the data are informative about the conditional correlation of y_1 and y_2 given x, the magnitude of y_1 is generally informative about the location of y_2 within the two classes ($y_2 \geq 0$, $y_2 < 0$) into which it is coded. Consequently, with D observed in place of y_2, the "single equation" probit estimator of π_2 which ignores the y_1 data, using just D and x data, is generally inefficient.

It is shown in section II that the fully efficient maximum likelihood (ML) estimator of π_2 is obtained by a probit analysis of D on x and y_1 so that ML estimation of (1) can be achieved using standard OLS and probit analysis software with negligible increase in computational cost over separate analysis of the y_1 and D data. In section III the magnitude of the (asymptotic) efficiency gain, which is nonzero for $\rho \neq 0$, is considered.

II. Marginal and Joint Maximum Likelihood Estimators

From the marginal distributions of y_1 and D the log likelihood functions (2) and (3) are obtained, assuming n independent realisations of y_1 and D given x.

$$L_m(\pi_1, \sigma_{11}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_{11} - \frac{1}{2\sigma_{11}}\sum(y_{1i} - x_i'\pi_1)^2 \tag{2}$$

$$L_m(\pi_2) = \sum \log \bar{\Phi}\left((1 - 2D_i)x_i'\pi_2\right). \tag{3}$$

Here $\bar{\Phi}$ is the complement of the standard normal distribution function. All summations here and later are over $i = 1$ to n.

The maximum likelihood estimators from (2) and (3) are $\hat{\pi}_1$, the OLS estimator given by regressing y_1 on x, $\hat{\sigma}_{11}$, the mean squared OLS residuals from this regression and $\hat{\pi}_2$, the probit estimator obtained from probit analysis of D using x as explanatory variables. These estimators, which are called marginal maximum likelihood (MML) estimators, have the usual optimality properties under the assumption that y_1 given x and y_2 given x are marginally normally distributed.

If y_1 and y_2 are assumed to be *jointly* normally distributed given x then the joint log-likelihood is

$$\begin{aligned} L_J(\gamma_1, \theta_1, \gamma_2, \theta_2) &= \sum \log \bar{\Phi}\left((1 - 2D_i)(x_i'\gamma_2 + \theta_2 y_{1i})\right) \\ &\quad - \frac{n}{2}\log 2\pi - \frac{n}{2}\log\theta_1 - \frac{1}{2\theta_1}\sum(y_{1i} - x_i'\gamma_1)^2. \end{aligned} \tag{4}$$

The log likelihood (4) is written using the decomposition $P(y_1 \cap D|x) = P(D|y_1 \cap x)P(y_1|x)$ and the parameterisation:

$$\begin{aligned} \gamma_1 &= \pi_1 \\ \theta_1 &= \sigma_{11} \\ \gamma_2 &= (1 - \rho^2)^{-1/2}(\pi_2 - \rho\sigma_{11}^{-1/2}\pi_1) \\ \theta_2 &= \rho\sigma_{11}^{-1/2}(1 - \rho^2)^{-1/2}. \end{aligned} \tag{5}$$

Setting the first derivatives of (4) equal to zero it can be seen that the joint maximum likelihood (JML) estimators of γ_1 ($= \pi_1$) and θ_1 ($= \sigma_{11}$) are identical to the MML estimators[2] of π_1 and σ_{11} and that the JML estimators of γ_2 and θ_2 are obtained by a simple probit analysis of the D data using x *and* y_1 as explanatory variables.

The inverse of the transformation (5) is

$$\begin{aligned} \pi_1 &= \gamma_1 \\ \sigma_{11} &= \theta_1 \\ \pi_2 &= (1 + \theta_1\theta_2^2)^{-1/2}(\gamma_2 + \theta_2\gamma_1) \\ \rho &= \theta_1^{1/2}\theta_2(1 + \theta_1\theta_2^2)^{-1/2} \end{aligned} \tag{6}$$

which enables unique estimates of π_1, π_2, σ_{11} and ρ to be obtained from estimates of γ_1, γ_2, θ_1 and θ_2. The invariance properties of maximum likelihood estimators

[1] Since y_2 is not observable the normalisation $\text{var}(y_2|x) = 1$ is innocuous.

[2] This is also true when elements of π_1 are constrained to be zero.

ensure that the resulting estimators $(\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\sigma}_{11}, \tilde{\rho})$ possess the usual optimality properties.

The preceding argument demonstrates that full maximum likelihood estimation of the parameters of (1) can be achieved via a simple two-step procedure comprising OLS estimation of π_1 and σ_{11} using the y_1 and x data and probit analysis of the D, x and y_1 data. If M continuous variates $\underline{y_1}$ are available then the probit analysis uses x and the vector $\underline{y_1}$ as explanatory variables. This two-step procedure is essentially the reverse of that suggested by Heckman (1979) to correct for sample selection bias, as is expected given the decomposition of $P(y_1 \cap D|x)$ used in writing down the joint log likelihood (4).

So, joint maximum likelihood estimation is computationally straightforward. But what gains in efficiency can be expected from exploiting ancillary continuous data? This question is investigated in the next section.

III. Asymptotic Variances and Relative Efficiency

First consider the marginal maximum likelihood estimator, $\hat{\pi}_2$. The asymptotic variance matrix of $\sqrt{n}(\hat{\pi}_2 - \pi_2)$, $\overline{\text{var}}(\hat{\pi}_2)$ is:

$$\overline{\text{var}}(\hat{\pi}_2) = \left\{ \text{plim}\, n^{-1} \sum_{i=1}^{n} h(x_i'\pi_2) h(-x_i'\pi_2) x_i x_i' \right\}^{-1} \tag{7}$$

where $h(\;)$ is the standard normal hazard function. It is assumed that probability limits exist and where appropriate are non-singular.

If $\phi(w)$ and $\Phi(w)$ are the standard normal density and distribution functions then $h(w)h(-w) = \phi(w)^2/\Phi(w)\Phi(-w)$ which attains its maximum of $2/\pi = 0.637$ when $w = 0$. With y_2 observed, the asymptotic variance of the ML estimator of π_2 (the OLS estimator from regressing y_2 on x) is (7) with $h(x_i'\pi_2)h(-x_i'\pi_2)$ replaced by $\text{var}(y_2|x_i)^{-1} = 1.0$. So the grouping of y_2 into two classes coded by D results in an increase in the asymptotic variance of π_2 of at least $100\,((.637)^{-1} - 1)\% = 57\%$. Assuming *joint* normality of y_1 and y_2 given x and utilising the y_1 data as described in the previous section allows some of this lost efficiency to be regained but the y_1 data cannot be used to reduce the variance of $\tilde{\pi}_2$ below that attained when y_2 is observed.

Now consider the asymptotic variance matrix of the JML estimator. Since the log likelihood (4) is an additively separable function of (γ_1, θ_1) and (γ_2, θ_2) the asymptotic covariance matrix of the JML estimator of $(\gamma_1, \theta_1, \gamma_2, \theta_2)$ is block diagonal given by

$$\overline{\text{var}} \begin{bmatrix} \tilde{\gamma}_1 \\ \tilde{\theta}_1 \\ \tilde{\gamma}_2 \\ \tilde{\theta}_2 \end{bmatrix} = V = \left[\begin{array}{cc|cc} \theta_1 Q^{-1} & 0 & 0 & 0 \\ 0 & 2\theta_1^2 & 0 & 0 \\ \hline 0 & 0 & & \\ 0 & 0 & & V_2 \end{array}\right] \tag{8}$$

where

$$V_2^{-1} = \text{plim}\, n^{-1} \sum E h(w_i) h(-w_i) \times \left[\begin{array}{c} x_i \\ \hline \dfrac{w_i - x_i'\gamma_2}{\theta_2} \end{array}\right] \left[\begin{array}{c|c} x_i' & \dfrac{w_i - x_i'\gamma_2}{\theta_2} \end{array}\right],$$

expectation is with respect to w_i given x_i which is

$$N_1\left[x_i'(\gamma_2 + \theta_2\gamma_1), \theta_2^2\theta_1)\right]$$

and

$$Q = \text{plim}\, n^{-1} \sum_{i=1}^{n} x_i x_i'.$$

To obtain the asymptotic variance matrix of the JML estimators of $(\pi_1, \sigma_{11}, \pi_2, \rho)$, consider the Jacobian, Δ, of the transformation (5):

$$\Delta = \left[\begin{array}{cc|cc} I_m & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline \dfrac{\rho}{\sigma_{11}^{1/2}} I_m & \dfrac{-\rho^2\pi_2}{2\sigma_{11}} & (1-\rho^2)^{1/2} I_m & (1-\rho^2)^{1/2}(\pi_1 - \sigma_{11}^{1/2}\rho\pi_2) \\ 0 & \dfrac{\rho}{2\sigma_{11}}(1-\rho^2) & 0 & \sigma_{11}^{1/2}(1-\rho^2)^{3/2} \end{array}\right]$$

$$= \left[\begin{array}{c|c} I_{m+1} & 0 \\ \hline A & B \end{array}\right], \quad \text{say.} \tag{9}$$

A standard argument exploiting the local linearity of the transformation (5) gives, for the asymptotic variance matrix of $(\tilde{\pi}_1, \tilde{\sigma}_{11}, \tilde{\pi}_2, \tilde{\rho})$, $\Delta V \Delta'$, which is

$$\overline{\text{var}}\begin{bmatrix} \tilde{\pi}_1 \\ \tilde{\sigma}_{11} \\ \tilde{\pi}_2 \\ \tilde{\rho} \end{bmatrix} = \left[\begin{array}{cc|c} \begin{matrix} \sigma_{11} Q^{-1} & 0 \\ 0 & 2\sigma_{11}^2 \end{matrix} & & \begin{matrix} \rho\sigma_{11}^{1/2} Q_2^{-1} & 0 \\ -\rho^2\sigma_{11}\pi_2' & \sigma_{11}\rho(1-\rho^2) \end{matrix} \\ \hline \begin{matrix} \rho\sigma_{11}^{1/2} Q^{-1} & -\rho^2\sigma_{11}\pi_2 \\ 0 & \sigma_{11}\rho(1-\rho^2) \end{matrix} & & \begin{bmatrix} \rho^2 Q^{-1} + \dfrac{\rho^4}{2}\pi_2\pi_2' & \dfrac{-\rho^3(1-\rho^2)\pi_2}{2} \\ \dfrac{-\rho^3(1-\rho^2)\pi_2'}{2} & \dfrac{\rho^2(1-\rho^2)^2}{2} \end{bmatrix} + BV_2B' \end{array}\right]. \tag{10}$$

Some tedious algebra shows that $\overline{\text{var}}(\hat{\pi}_2)$ exceeds $\overline{\text{var}}(\tilde{\pi}_2)$ by a positive semidefinite matrix which converges to zero as ρ passes to zero.[3]

The gain in efficiency from using the y_1 data is obtained at the cost of making an extra assumption, namely, that y_1 and y_2 given x are jointly as well as marginally normally distributed. When this additional assumption is incorrect the JMLE $\tilde{\pi}_2$ may be inconsistent but while the marginal normality assumption is correct $\hat{\pi}_2$, the MMLE, is consistent. When the joint normality assumption is correct the JMLE is efficient relative to the MMLE. So $\tilde{\pi}_2 - \hat{\pi}_2$ provides the basis for a Hausman (1978) test of the additional assumption.

The variance matrix (10) can be estimated in the two-step procedure outlined earlier. Q^{-1} is consistently estimated by $(n^{-1}\Sigma x_i x_i')^{-1}$ and $\pi_1, \pi_2, \sigma_{11}, \rho$ by their JML estimators. The matrix V_2 is consistently estimated by the inverse of n^{-1} times the Hessian of the "log likelihood" used in the probit analysis of D on y_1 and x. Thus all the required elements of $\overline{\text{var}}(\tilde{\pi}_2)$ are readily available as normal outputs of the OLS and probit programmes used in the two-step procedure.

Now consider the magnitude of the efficiency gain obtained by using the ancillary continuous data. When there is no regression, so that x is scalar and equal to one, it is possible to calculate this gain for most interesting values of π_2 and ρ. The results are shown in table 1 where two variances $\overline{\text{var}}(\tilde{\pi}_2|\rho)$ and $\overline{\text{var}}(\tilde{\pi}_2)$ are given, since in this model the asymptotic variance of $\tilde{\pi}_2$ depends on whether ρ is known or estimated. The column headed $\rho^2 = 0$ gives $\overline{\text{var}}(\hat{\pi}_2)$. The three variances are invariant under changes in the signs of π_2 and ρ and are smallest when $\pi_2 = 0$, i.e., when $P[D = 1] = 0.5$. In this "no regression" case, $\hat{\pi}_2 = -\bar{\Phi}^{-1}(n_1/n)$ where n_1 is the number of observations with $D = 1$. Though $\text{var}(n_1/n)$ is at a *maximum* when $\pi_2 = 0$, the increasing flatness of the $\bar{\Phi}^{-1}$ function results in $\overline{\text{var}}(\hat{\pi}_2)$ being at a minimum when $\pi_2 = 0$. For all $\rho^2 < 1$, $\overline{\text{var}}(\hat{\pi}_2) \geq \overline{\text{var}}(\tilde{\pi}_2) \geq \overline{\text{var}}(\tilde{\pi}_2|\rho) > \overline{\text{var}}(\hat{\pi}_2|y_2 \text{ observed}) = 1$, with equalities holding when $\rho = 0$. The greatest gains in efficiency are obtained for large $|\pi_2|$, when the event $D = 1$ is either relatively rare or relatively common, and for large ρ^2.

Now suppose there is regression and write $x'\pi_2 = \pi_{20} + \pi_{21}x_1 + \pi_{22}x_2$ where x_1 and x_2 are $N(0, 1)$, $\text{cor}(x_1, x_2) = r$ and realisations of $(x_1 x_2)$ are independent. Taking expectations over x gives expressions for $\overline{\text{var}}(\hat{\pi}_{2i})$ and $\overline{\text{var}}(\tilde{\pi}_{2i}|\rho)$ which are evaluated (with $\pi_{21} = \pi_{22}$ denoted by $\pi_{2.}$) to produce table 2. Only one variance is reported since under these conditions variances of the estimators of π_{21} and π_{22} are equal. Asymptotic covariances, which are not reported, are reduced in magnitude on introducing the y_1 data by approximately the same amount as is the asymptotic variance.

The asymptotic variance of the MML estimator $\hat{\pi}_{2i}$ (see the columns headed $\rho^2 = 0$) increases with π_{20} and $\pi_{2.}$. With $\pi_{2.} = 0$ minimum asymptotic variance is obtained when x_1 and x_2 are uncorrelated but as $\pi_{2.}$ increases through positive values the value of r for which the asymptotic variance is at a minimum declines. Comparing the entries in table 2 with the last column of table 2, which gives $\overline{\text{var}}(\hat{\pi}_{2i}|y_2 \text{ observed}) = (1 - r^2)^{-1}$, it can be seen that the loss in efficiency due to the grouping of y_2 into the two classes indicated by D increases with π_{20} and with $\pi_{2.}$ and for non-zero $\pi_{2.}$ varies with r, generally increasing as the magnitude of r increases.

TABLE 1.—ASYMPTOTIC VARIANCES OF $\tilde{\pi}_2$, ρ KNOWN (UPPER ENTRIES), ρ ESTIMATED (LOWER ENTRIES)

	ρ^2			
π_2	0.0[a]	0.4	0.8	0.95
0	1.57	1.55	1.43	1.26
	1.57	1.55	1.43	1.26
1.0	2.29	2.17	1.82	1.46
	2.29	2.27	2.15	1.91
2.0	7.62	7.00	5.10	3.21
	7.62	7.52	6.48	5.03
2.5	20.09	18.58	13.46	7.81
	20.09	19.71	15.89	10.71

[a]When $\rho^2 = 0$, entries give $\overline{\text{var}}(\hat{\pi}_2) = \overline{\text{var}}(\tilde{\pi}_2)$.

[3] When $\rho^2 = 1$ and ρ is known, y_2 can be "reconstructed" from y_1 and x and all the lost efficiency due to grouping y_2 can be regained. It seems plausible that this is also true, asymptotically, when ρ is unknown and in fact $\rho^2 = 1$.

TABLE 2.—ASYMPTOTIC VARIANCES OF $\tilde{\pi}_{2.}$, ρ KNOWN, WHERE $x'\pi_2 = \pi_{20} + \pi_{2.}x + \pi_{2.}x_2$, $r = \text{cor}(x_1, x_2)$

		$\pi_{20} = 0.0$				$\pi_{20} = 2.0$				$\overline{\text{var}}(\hat{\pi}_{2.} \mid$
$\pi_{2.}$	r	ρ^2: 0[a]	0.4	0.8	0.95	0[a]	0.4	0.8	0.95	y_2 observed)
	−.9	8.27	8.18	7.55	6.64	40.1	36.8	26.8	16.9	5.26
	−.6	2.45	2.43	2.24	1.97	11.9	10.9	8.0	5.0	1.56
	−.3	1.73	1.71	1.58	1.39	8.4	7.7	5.6	3.5	1.10
0	0.0	1.57	1.55	1.43	1.26	7.6	7.0	5.1	3.2	1.00
	.3	1.73	1.71	1.58	1.39	8.4	7.7	5.6	3.5	1.10
	.6	2.45	2.43	2.24	1.97	11.9	10.9	8.0	5.0	1.56
	.9	8.27	8.18	7.55	6.64	40.1	36.8	26.8	16.9	5.26
	−.9	10.11	9.74	8.52	7.13	40.8	36.2	25.7	16.1	5.26
	−.6	4.65	4.26	3.37	2.55	12.8	10.9	7.5	4.7	1.56
	−.3	4.25	3.81	2.87	2.05	9.3	7.9	5.4	3.3	1.10
1	0.0	4.48	3.97	2.92	2.02	8.4	7.2	4.9	3.0	1.00
	.3	5.16	4.54	3.31	2.28	8.8	7.5	5.1	3.2	1.10
	.6	6.91	6.08	4.47	3.11	10.9	9.3	6.4	4.1	1.56
	.9	18.89	16.77	12.71	9.26	28.1	24.1	17.1	11.4	5.26
	−.9	16.62	15.18	11.87	8.85	45.9	39.3	26.9	16.6	5.26
	−.6	13.93	11.96	8.14	5.00	22.1	18.7	12.3	7.1	1.56
	−.3	15.71	13.31	8.76	5.08	21.2	17.8	11.5	6.5	1.10
2	0.0	17.89	15.08	9.81	5.56	22.3	18.7	12.0	6.7	1.00
	.3	20.58	17.30	11.22	6.34	24.5	20.5	13.2	7.4	1.10
	.6	25.03	21.03	13.70	7.84	28.9	24.2	15.6	8.8	1.56
	.9	47.92	40.28	26.95	16.52	54.1	45.3	29.9	18.0	5.26

[a] With $\rho^2 = 0$ entries give $\overline{\text{var}}(\hat{\pi}_{2.})$.

The gain in efficiency on introducing the y_1 data increases with ρ^2, π_{20} and $\pi_{2.}$ and varies with $r = \text{cor}(x_1, x_2)$. For models in which the exogenous variables are highly correlated and for which the event $D = 1$ is relatively rare or relatively common, considerable gains in efficiency can be achieved by using the joint maximum likelihood estimator which, as noted earlier, is simple to calculate.

REFERENCES

Conniffe, Denis, "Testing the Assumptions of Seemingly Unrelated Regressions," this REVIEW 64 (Feb. 1982), 172–174.

Hausman, Jerry A., "Specification Tests in Econometrics," *Econometrica* 46 (Nov. 1978), 1251–1271.

Heckman, James, "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica* 46 (July 1978), 931–959.

——, "Sample Selection Bias as Specification Error," *Econometrica* 47 (Jan. 1979), 153–161.