A Monte Carlo comparison of estimators for a bivariate probit model with selection

R. Belkar[1] and D.G. Fiebig[1]

[1]School of Economics, University of New South Wales, Sydney, 2052.
Email: d.fiebig@unsw.edu.au

EXTENDED ABSTRACT

The prototypical sample selection model consists of a two-equation system: one equation representing the selection mechanism and the second a continuous outcome variable that is only observed for the selected cases. A variant of this model where the outcome variable is binary leads to a bivariate probit model with sample selection.

A Monte Carlo experiment is undertaken to examine the small sample properties of three alternative estimators of a bivariate probit model with selection. The three estimators are the censored probit estimator, single-equation probit applied to the selected sub-sample and single-equation probit applied to the full sample. These estimators are compared in terms of properties of coefficient estimates and predicted probabilities.

From a methodological perspective we emphasise the need to account for possible censoring in modelling choice. In practice, one could argue that in finite samples a simple single-equation approach might be just as effective as using a censored probit estimator that relies on a large sample justification. For example, one might claim that with only minor censoring the problem could be ignored. Our Monte Carlo results do not support this argument.

While no one estimator dominates in all possible situations a clear recommendation follows from an overall evaluation of the relative performance of the three estimators. Ignoring the selection problem through use of a single-equation probit can often lead to very poor estimator and predictor performance. Both single equation probit estimators have properties that can vary dramatically over the different design points. The properties of censored probit vary much less than the two single equation estimators and this robustness characteristic tends to favour its use.

The fluctuations in relative performance are more pronounced for full-sample probit than for sub-sample probit. Obviously, in order to avoid the use of full-sample probit, it is necessary to know the sample split into censored and non-censored observations. In terms of survey design this may require the inclusion of questions that probe for reasons why people do not choose.

1. INTRODUCTION

Selection problems abound in micro-econometric modelling. The prototypical sample selection model consists of a two-equation system: one equation representing the selection mechanism and the second a continuous outcome variable that is only observed for the selected cases. A variant of this model where the outcome variable is binary leads to a bivariate probit model with sample selection. While there are various forms of censoring that could be considered as in Poirier (1980) and Meng and Schmidt (1985), we concentrate on the variant proposed by van de Ven and van Praag (1981).

While recent surveys of Vella (1998), Puhani (2001) and Lee (2001) provide useful summaries of the extensive literature on selection models they include little by way of small sample properties of alternative estimators in the particular case of a bivariate probit model with selection. Our interest was stimulated by the applied work of Belkar et al. (2005) where two issues arose that seemed to have been downplayed in the existing examination of alternative estimators in prototypical selection models and hence justified the current Monte Carlo study.

Belkar et al. (2005) were interested in determinants of whether women had ever screened for cervical cancer using a Pap test. Recent national surveys in Australia indicated that a small proportion (4% in 1995) of women had never heard of a Pap test. For these women, they have never had a test not because of any conscious choice but because they simply were not aware of its existence. There is a potential problem if the entire sample is used with no differentiation made between non-screeners who are there by choice and those that are non-screeners because of lack of awareness.

Piga and Vivarelli (2003) argue that previous studies of whether firms have undertaken cooperative R&D or not using a single equation framework have ignored the initial decision to carry out R&D in the first place. For their particular example, probit estimates for factors affecting whether firms undertake cooperative R&D appeared quite different depending on whether they were based on the selected sub-sample of firms who carried out some R&D or on the full sample of firms.

Thus, the first issue to be considered in our Monte Carlo comparisons is the relative performance of these two single-equation probit estimators that vary according to whether the selected sub-sample is used or whether the full sample is used including observations that are incorrectly coded as non-choosers.

One might expect that selection problems would be minor given only limited censoring. This would have to be a conjecture because existing Monte Carlo work typically includes much more dramatic censoring amongst its designs. This conjecture proved to be incorrect in the empirical work of Belkar et al (2005). Thus, the second issue to be addressed is whether there are situations routinely found in practice where modest censoring makes a substantial difference for inferences if ignored in the modelling.

In what follows, we consider the estimation of a bivariate probit model with selection. The finite sample properties of three estimators are compared: the censored probit estimator, the single-equation probit applied to the selected sub-sample, which we call sub-sample probit, and single-equation probit applied to the full sample which is called full-sample probit.

2. THE MODEL AND EXPERIMENTAL DESIGN

Denote the latent variable representing the outcome choice by $y_1{*}$ and that representing selection by $y_2{*}$ and assume a simple model specification of the form:

$$(1) \quad \begin{aligned} y_{1i}{*} &= \beta_{11} + \beta_{12}x_{1i} + u_{1i} \\ y_{2i}{*} &= \beta_{21} + \beta_{22}x_{2i} + u_{2i} \end{aligned}$$

where the realisation of the latent variable $y_{ji}{*}$ is defined by $y_{ji} = 1$ if $y_{ji}{*} > 0$ and $y_{ji} = 0$ otherwise and x_{ji} are exogenous variables. Under the assumption that the disturbance term (u_1, u_2) is bivariate normal with:

var $(u_j) = 1$ and cov $(u_1, u_2) = \rho$

(1) specifies a bivariate probit model. The sample selection or censored probit variant of this model occurs when both y_1 and y_2 are only observed if $y_2 = 1$.

Given this structure, univariate probit analyses of outcome choice will typically produce inconsistent estimates of the model's coefficients. If the available data allow us to separate cases into those that are selected and this sub-sample is used to estimate the outcome model then this will produce inconsistent estimates unless $\rho = 0$.

In some cases there is no information that enables this split of the sample. If non-selectors are automatically treated as non-choosers and full-sample probit is applied, inconsistent

parameter estimates will be produced because of the inflation of the number of non-choosers in the sample. The problem is similar to that which occurs with a misclassified binary dependent variable. See Hausman, Abrevaya and Scott-Morton (1998) who emphasise that small levels of misclassification can lead to substantial estimation biases.

Previous studies of sample selection with a continuous censored variable [see Nelson (1984), Zuehlke and Zeman (1991), Leung and Yu (1996) and Nawata and Nagase (1996)] have isolated correlation between the variables in the two equations, dependence between the disturbances, and the degree of censoring as key factors in any comparison of alternative estimators.

Values of the independent variables are generated as bivariate normal as follows:

$$(2) \quad \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \pi \\ \pi & 1 \end{pmatrix} \right].$$

Varying π, allows for various levels of correlation between the variables in the outcome and selection equations. $\pi = 0$ corresponds to independent regressors and is expected to be the most conducive for estimation. An intermediate value of $\pi = 0.8$ is probably more representative of situations likely to occur in practice.

Dependence between the outcome and selection equations is varied according to $\rho = -0.8, -0.4, 0, 0.4, 0.8$.

Given true regression parameter values of $\beta_{12} = \beta_{22} = 1$, the intercept parameter β_{21} governs the degree of selection (or conversely censoring) while the other intercept, β_{11}, governs the outcome penetration. Three pairs of values were chosen where $\beta_{11} = \beta_{21} = 2.33, 0.96$ and 0 corresponding to equal degrees of selection and penetration of 95%, 75% and 50% respectively. A fourth pair was chosen as $\beta_{11} = 2.33, \beta_{21} = 0.96$ in order to gauge the separate impact of higher penetration in the outcome choice holding selection or censoring fixed. The sample size was fixed at n = 400 and there were 250 replications for all cases.

3. MONTE CARLO RESULTS

The estimators are compared on the basis of four criteria, three of which refer to estimation of β_{12}, the key parameter in the outcome equation. These criteria are the percentage bias, relative efficiency as measured by the RMSE and a measure of how well the asymptotic standard errors (ASE) reflect the true variability of the estimators, represented by the ratio of the ASE and the RMSE. The fourth criterion relates to predicting probabilities and is represented by the mean probability of success in the outcome equation.

When the estimated correlation between the disturbances is close to 1 or -1 complications such as non-convergence or a non-invertible Hessian (resulting in missing standard errors) can occur. This problem most often arises when there is concurrent minor censoring and extreme penetration. Conditions for when the estimated correlation approaches the boundaries ±1 are discussed in Butler (1996). Consequently, the cases of non-convergence and missing standard errors have been excluded from the analysis. As a result some of the conclusions discussed below are based on less than 250 observations.

There are two sets of results; the first set comprising Tables 1 and 2 containing bias and RMSE and the second comprising Tables 3 and 4 where the ASE and RMSE ratios and mean probabilities are presented. Each set of results has the same structure. Each table refers to a different level of correlation between the variables in the outcome and selection equations. Within each table the results are provided for each level of censoring and for the range of disturbance correlations considered.

Consider Table 1 which refers to the case where $\pi = 0$. Here censored probit clearly dominates the other two estimators both in terms of bias and relative efficiency. As expected, the differences between censored probit and sub-sample probit are minor when $\rho = 0$, but as the disturbance correlation increases in magnitude the bias of sub-sample probit increases. For $\pi = 0$, full-sample probit performs poorly irrespective of the degree of censoring or disturbance correlation exhibiting severe downward bias in the estimation of β_{12}. Holding the degree of censoring constant, a higher outcome penetration leads to poorer small sample properties for each of the three estimators.

The case of positive correlation between the regressors in the outcome and selection equations is shown in Table 2. In this case, the relative performance of the three estimators is less clear-cut. Censored probit continues to dominate the other two estimators in terms of bias although its superiority is not uniform. However, in terms of RMSE the two single equation estimators perform well. In particular, full-sample probit, despite biases that can be quite large, performs best overall in terms of RMSE.

Obviously the variance of full-sample probit for these design points is small relative to the other two estimators. The censored observations that are being added to the selected sample and being treated as non-choosers will tend to be observations with relatively extreme x_1 (negative) values. Because of the positive correlation between the explanatory variables in the outcome and selection equations these observations will provide added leverage and hence increase estimation precision. This tendency is reinforced when there are large correlations between the disturbances.

The patterns in Tables 3 and 4 are reasonably clear. Except for a few exceptions, censored probit performs extremely well. The estimated asymptotic standard errors tend to provide an accurate picture of actual variability (measured by RMSEs), indicated by ASE/RMSE ratios that tend to be close to unity. In terms of predicted probabilities, censored probit almost always provides predictions close to the theoretical values of 50% (extensive censoring), 75% (moderate censoring), 95% (minor censoring) and 95% (high outcome penetration and moderate censoring).

Neither of the single equation probit estimators do overly well in these comparisons. As expected, sub-sample probit is fine when the disturbance correlation is small. But for larger disturbance correlations, there is a tendency for the asymptotic standard errors to understate the true variability in the estimator of β_{12} and to produce predicted probabilities that are biased down for negative disturbance correlations and biased up when the correlations are positive.

The performance of full-sample probit is even more problematic. There is a systematic tendency for the asymptotic standard errors to understate the true variability in the estimator of β_{12}. Obviously this is often due to the substantial biases in the associated coefficient estimates. The inflation of non-choosers that occurs with full-sample probit impacts on the predicted probabilities as expected, producing predictions that are systematically biased downwards.

In an extended version of the paper, that is available on request, results were also produced for when the same regressor appears in the outcome and selection equations (i.e. $\pi = 1$). As expected this situation is the most challenging for censored probit. However, the qualitative nature of the results remains the same although the censored probit estimator is much more prone to bias than in the reported situations.

4. CONCLUSION

From a methodological perspective we have emphasised the need to account for possible censoring in modelling choice. In practice, one could argue that in finite samples a simple single-equation approach might be just as effective as using a censored probit estimator that relies on a large sample justification. For example, one might claim that with only minor censoring the problem could be ignored. Our Monte Carlo results do not support this argument.

While no one estimator dominates in all possible situations a clear recommendation follows from an overall evaluation of the relative performance of the three estimators. Ignoring the selection problem through use of a single-equation probit can often lead to very poor estimator and predictor performance. Both single equation probit estimators have properties that can vary dramatically over the different design points. The properties of censored probit vary much less than the two single equation estimators and this robustness characteristic tends to favour its use.

The fluctuations in relative performance are more pronounced for full-sample probit than for sub-sample probit. Obviously, in order to avoid the use of full-sample probit, it is necessary to know the sample split into censored and non-censored observations. In terms of survey design this may require the inclusion of questions that probe for reasons why people do not choose. For the screening example that has been discussed, this involved asking a question about screening awareness.

5. ACKNOWLEDGMENTS

6. REFERENCES

Belkar, R., Fiebig, D.G., Haas, M. and Viney, R. (2005), "Why worry about awareness in choice problems? Econometric analysis of screening for cervical cancer", forthcoming *Health Economics.*

Butler, J.S. (1996), "Estimating the correlation coefficient in censored probit models", *Review of Economics and Statistics* 78, 356-358.

Hausman, J.A., Abrevaya, J. and Scott-Morton, F.M. (1998), "Misclassification of dependent variable in discrete-choice setting", *Journal of Econometrics* 87, 239-270.

Lee, L.F. (2001), "Self-selection", in *Companion in Theoretical Econometrics*, B.H. Baltagi ed., Basil Blackwell: Oxford.

Leung, S.F. and Yu, S. (1996), "On the choice between sample selection and two-part models", *Journal of Econometrics* 72, 197-229.

Meng, C.L. and Schmidt, P. (1985), "On the cost of partial observability in the bivariate probit model", *International Economic Review* 26, 71-86.

Nawata, K. and Nagase, N. (1996), "Estimation of sample selection bias models", *Econometric Reviews* 15, 387-400.

Nelson, F. (1984), "Efficiency of the two-step estimator for models with endogenous sample selection", *Journal of Econometrics* 24, 181-196.

Piga, C. and Vivarelli, M. (2003), "Sample selection in estimating the determinants of cooperative *R&D*", *Applied Economics Letters* 10, 243-246.

Puhani, P.A. (2001), "The Heckman correction for sample selection and its critique", *Journal of Economic Surveys* 14, 53-68.

van de Ven, W.P.M.M. and van Praag, B.M.S. (1981), "The demand for deductibles in private health insurance", *Journal of Econometrics* 17, 229-252.

Vella, F. (1998), "Estimating models with sample selection bias: A Survey", *Journal of Human Resources* 33, 127-169.

Zuehlke, T.W. and Zeman, A.R. (1991), "A comparison of two-stage estimators of censored regression models", *Review of Economics and Statistics* 73, 185-189.

Table 1: Bias and RMSE of alternative probit estimators of β_{12}: Independent regressors ($\pi = 0$)

ρ	% Bias			RMSE		
	Censored probit	Sub-sample probit	Full-sample probit	Censored probit	Sub-sample probit	Full-sample probit
Extensive censoring: $\beta_{11} = \beta_{21} = 0$						
-0.8	0.66	13.60	-33.86	0.147	0.206	0.349
-0.4	1.26	4.46	-42.71	0.137	0.145	0.433
0.0	2.00	3.12	-47.22	0.147	0.153	0.477
0.4	1.90	5.12	-51.22	0.146	0.154	0.517
0.8	2.65	17.20	-54.48	0.160	0.240	0.549
Moderate censoring: $\beta_{11} = \beta_{21} = 0.96$						
-0.8	2.07	9.62	-43.59	0.149	0.185	0.442
-0.4	0.80	3.38	-48.21	0.131	0.136	0.487
0.0	0.55	1.84	-51.53	0.130	0.130	0.520
0.4	-0.11	3.48	-55.28	0.135	0.140	0.557
0.8	2.45	14.93	-56.29	0.146	0.215	0.567
Minor censoring: $\beta_{11} = \beta_{21} = 2.33$						
-0.8	5.27	7.33	-54.26	0.236	0.244	0.554
-0.4	2.42	4.42	-54.70	0.229	0.234	0.557
0.0	2.45	5.21	-55.29	0.230	0.247	0.564
0.4	3.53	7.59	-56.49	0.266	0.278	0.576
0.8	6.21	13.85	-59.22	0.261	0.291	0.603
High outcome penetration and moderate censoring: $\beta_{11} = 2.33$, $\beta_{21} = 0.96$						
-0.8	8.27	12.32	-79.59	0.258	0.282	0.798
-0.4	3.75	6.26	-81.41	0.245	0.255	0.817
0.0	4.02	6.68	-83.61	0.265	0.274	0.839
0.4	8.08	13.72	-84.72	0.341	0.374	0.850
0.8	14.84	29.49	-88.08	0.584	0.685	0.883

Table 2: Bias and RMSE of alternative probit estimators of β_{12}: Correlated regressors ($\pi = 0.8$)

ρ	% Bias			RMSE		
	Censored probit	Sub-sample probit	Full-sample probit	Censored probit	Sub-sample probit	Full-sample probit
Extensive censoring: $\beta_{11} = \beta_{21} = 0$						
-0.8	2.21	47.80	26.64	0.236	0.526	0.295
-0.4	1.61	18.52	9.19	0.219	0.246	0.138
0.0	-0.53	1.39	-1.16	0.183	0.151	0.101
0.4	1.36	-6.86	-8.71	0.157	0.172	0.131
0.8	2.13	-12.46	-15.29	0.140	0.206	0.177
Moderate censoring: $\beta_{11} = \beta_{21} = 0.96$						
-0.8	4.07	34.96	18.67	0.180	0.381	0.220
-0.4	2.26	14.82	7.18	0.201	0.217	0.126
0.0	-1.29	2.02	-1.12	0.175	0.145	0.094
0.4	-0.28	-7.76	-8.50	0.150	0.168	0.125
0.8	0.07	-13.30	-14.71	0.127	0.196	0.171
Minor censoring: $\beta_{11} = \beta_{21} = 2.33$						
-0.8	9.55	19.89	8.08	0.231	0.300	0.182
-0.4	4.54	10.38	4.94	0.207	0.225	0.158
0.0	1.96	4.01	-0.96	0.209	0.208	0.153
0.4	0.36	-2.91	-6.53	0.214	0.206	0.167
0.8	1.96	-6.20	-10.86	0.235	0.253	0.174
High outcome penetration and moderate censoring: $\beta_{11} = 2.33$, $\beta_{21} = 0.96$						
-0.8	13.35	50.39	-11.58	0.326	0.608	0.149
-0.4	6.17	20.70	-17.70	0.323	0.345	0.198
0.0	3.14	5.30	-21.26	0.326	0.299	0.230
0.4	4.28	-2.44	-25.91	0.338	0.363	0.274
0.8	8.48	26.57	-30.34	0.644	2.265	0.315

Table 3: Inference and prediction with alternative probit estimators: Independent regressors ($\pi=0$)

ρ	ASE/RMSE			Predicted probabilities		
	Censored probit	Sub-sample probit	Full-sample probit	Censored probit	Sub-sample probit	Full-sample probit
Extensive censoring: $\beta_{11}=\beta_{21}=0$						
-0.8	1.010	0.774	0.281	0.507	0.369	0.176
-0.4	1.070	1.015	0.206	0.498	0.435	0.209
0.0	0.983	0.947	0.176	0.500	0.505	0.243
0.4	0.998	0.954	0.155	0.509	0.573	0.274
0.8	0.942	0.679	0.141	0.508	0.647	0.310
Moderate censoring: $\beta_{11}=\beta_{21}=0.96$						
-0.8	0.837	0.706	0.296	0.759	0.713	0.522
-0.4	0.978	0.940	0.155	0.758	0.736	0.540
0.0	1.004	1.000	0.143	0.762	0.764	0.562
0.4	1.019	0.984	0.132	0.760	0.792	0.582
0.8	1.028	0.726	0.130	0.762	0.824	0.606
Minor censoring: $\beta_{11}=\beta_{21}=2.33$						
-0.8	0.895	0.883	0.183	0.954	0.954	0.899
-0.4	0.921	0.909	0.184	0.954	0.956	0.902
0.0	0.934	0.876	0.182	0.954	0.957	0.903
0.4	0.845	0.816	0.178	0.954	0.959	0.905
0.8	0.931	0.854	0.170	0.957	0.964	0.910
High outcome penetration and moderate censoring: $\beta_{11}=2.33$, $\beta_{21}=0.96$						
-0.8	0.874	0.834	0.089	0.952	0.944	0.696
-0.4	0.935	0.923	0.087	0.953	0.950	0.700
0.0	0.919	0.903	0.084	0.952	0.956	0.703
0.4	0.841	0.788	0.084	0.953	0.966	0.712
0.8	0.606	0.576	0.080	0.956	0.977	0.721

Table 4: Inference and prediction with alternative probit estimators: Correlated regressors ($\pi=0.8$)

ρ	ASE/RMSE			Predicted probabilities		
	Censored probit	Sub-sample probit	Full-sample probit	Censored probit	Sub-sample probit	Full-sample probit
Extensive censoring: $\beta_{11}=\beta_{21}=0$						
-0.8	0.935	0.363	0.431	0.477	0.486	0.235
-0.4	0.952	0.647	0.789	0.475	0.554	0.266
0.0	0.992	0.965	0.978	0.480	0.621	0.300
0.4	0.976	0.838	0.695	0.481	0.693	0.336
0.8	0.930	0.749	0.483	0.485	0.786	0.380
Moderate censoring: $\beta_{11}=\beta_{21}=0.96$						
-0.8	0.995	0.415	0.490	0.720	0.731	0.537
-0.4	0.893	0.659	0.788	0.716	0.760	0.560
0.0	0.960	0.951	1.007	0.727	0.794	0.584
0.4	0.976	0.816	0.724	0.721	0.832	0.612
0.8	1.025	0.754	0.511	0.729	0.880	0.649
Minor censoring: $\beta_{11}=\beta_{21}=2.33$						
-0.8	0.907	0.694	0.797	0.933	0.938	0.889
-0.4	1.006	0.879	0.900	0.933	0.942	0.890
0.0	1.015	0.964	0.914	0.937	0.949	0.897
0.4	0.972	0.971	0.826	0.936	0.955	0.905
0.8	0.901	0.885	0.799	0.939	0.967	0.914
High outcome penetration and moderate censoring: $\beta_{11}=2.33$, $\beta_{21}=0.96$						
-0.8	0.942	0.486	0.616	0.930	0.940	0.692
-0.4	0.939	0.765	0.448	0.928	0.952	0.703
0.0	0.950	0.929	3.421	0.931	0.967	0.714
0.4	1.000	0.933	2.705	0.928	0.980	0.723
0.8	1.339	0.421	2.648	0.931	0.993	0.732