14th October 2011

Extended abstract:

Non-response bias in studies of survey data with an application to residential mobility

Paul Clarke, Fiona Steele and Elizabeth Washbrook
University of Bristol, United Kingdom

Summary

This paper considers the issue of non-response or attrition bias in binary outcome models when the response mechanism takes a particular form: the probability an outcome is observed depends directly on the realized outcome. Our analysis is illustrated with a model of residential mobility applied to the British Household Panel Study (BHPS), a long-running household panel dataset similar to the US PSID. In the mobility application movers are less likely to be observed post-move because of loss of contact with the survey administrators. Hence achieved samples in such studies will tend to underestimate mobility rates, and more seriously those who are observed to move may not be typical of the population of movers as a whole. The usefulness of our approach is not restricted to residential mobility and migration, however, as the same logic applies in any situation in which the outcome causally affects response. Longitudinal follow-ups of medical interventions, for example, are likely affected by the fact that poor health or quality of life outcomes (the outcome of interest) directly reduce the probability of study participation.

We begin by outlining a simple modelling framework that incorporates the response mechanism of interest – referred to as the direct dependence model – and contrast it with the more standard textbook two-equation selection model approach in the manner of Heckman (1979)[1]. We then use simulations to characterize the extent and sign of the biases that result when the selection issue is ignored, and show that a standard Heckman-type approach has drawbacks when applied to a data generating process of this form, both in terms of the detection of and correction for those biases. We explore the sensitivity of our estimates to the parametric assumption of normal errors and investigate the role of exclusion restrictions in improving the stability of the two estimators.

Preliminary findings identify a simple rule for predicting the sign of the non-response bias in the complete case estimates. We find that when the true errors are normally distributed the Heckman-type estimator tends to falsely accept the null hypothesis of ignorable non-response, leading the researcher to underestimate the potential for biased estimates. The Heckman approach, incorrectly applied, will reduce the bias relative to complete case estimates but does not eliminate it. In contrast, the direct dependence model can correctly detect and fully eliminate the biases. When the true errors are non-normal, however, estimates from both frameworks remain biased, albeit less so than the complete case estimates. The existence of exclusion restrictions – variables that predict the outcome but not response or vice versa – can help to improve the stability of estimates of the response equation but have little impact on the coefficients of interest in the outcome equation. Our results suggest that, at least in one formulation, when errors are non-normal the direct dependence and Heckman-type approaches produce similar results.

Finally we apply the two models to data on the residential mobility of a sample of single adults from the BHPS. We focus on estimates of the effect of unemployment on local moves and employ two exclusion restrictions: local house price inflation is assumed to affect mobility but not response and survey membership status is assumed to affect response but not mobility. We show that the results from the

[1] Heckman, J.J. (1979). 'Sample Selection Bias as a Specification Error', *Econometrica*, 47: 153-161

complete case model and selection models are in line with the biases predicted by our simulation analyses. Specifically, the effect of unemployment on mobility is positive but small and insignificant when only complete cases are used, but becomes larger and statistically significant when non-response bias is modelled and corrected. Estimates of the selection parameters are negative as expected and strongly statistically significant. Hence our example provides a case in which the ignoring of non-response issues could lead the researcher to conclude incorrectly that unemployment is not associated with residential mobility, and also a case in which selection model estimates can detect and correct for downward non-response bias.

Model frameworks

We begin with the simplest possible cross-sectional scenario that enables us to illustrate the key features of the models. Extensions to the panel data case are applied in our empirical analysis and do not affect the substantive conclusions from the more simple models.

Suppose that the unobserved latent tendency for an individual to move between *t* and *t*+1 is captured by the latent variable M^*:

$$M^* = \alpha_0 + \alpha_1 X + \alpha_2' Z + \varepsilon^M \quad (1)$$

where X is a binary covariate (unemployment in our example), Z is a vector of controls capturing observed heterogeneity between individuals, ε^M is a zero mean unit variance error term, α_0 is a scalar constant term, α_1 is the coefficient of primary interest and α_2 is a vector of incidental coefficients. $X$and Z are measured at t and fully observed. The individual moves if the latent propensity is sufficiently high and moves are denoted by the indicator M, with $M = 1$ if $M^* > 0$ and $M = 0$ otherwise.

The direct dependence response equation is given by

$$R^* = \beta_0 + \beta_1 X + \beta_2' Z + \gamma M + \varepsilon^R \quad (2)$$

where R^* is the unobserved latent tendency for an individual to respond at t, β_0, β_1 and β_2 are scalars or vectors of parameters as before and ε^R is a zero mean unit variance error term, uncorrelated with ε^M. The term γM captures the idea that the response propensity of an individual with given X, Z and ε^R will differ – by the amount γ - if they move relative to the case in which they do not move. In the mobility example we expect that $\gamma < 0$. M is observed if $R^* > 0$ and not observed otherwise.

It can be shown that the probability of moving among the observed sample is given by

$$\Pr(M = 1|X, Z, R^* > 0) = \frac{F_M(\alpha' W) F_R(\beta' W + \gamma)}{F_M(\alpha' W) F_R(\beta' W + \gamma) + [1 - F_M(\alpha' W)] F_R(\beta' W)}$$

where $F_j(\varepsilon^j)$ is the marginal c.d.f. of ε^j $(j = M, R)$, $\alpha' W = \alpha_0 + \alpha_1 X + \alpha_2' Z$, and $\beta' W = \beta_0 + \beta_1 X + \beta_2' Z$. This expression depends on X not only via α_1, the parameter of interest, but also via β_1, its independent association with the response probability. Hence estimates based on the observed data will in general be biased.

In the more standard selection model it is assumed that $\gamma = 0$ while the assumption that the errors in the outcome and response equations are uncorrelated is relaxed. The response equation is then given by

$$R^* = \beta_0 + \beta_1 X + \beta_2' Z + \varepsilon^R \quad (3)$$

with $E(\varepsilon^M, \varepsilon^R) = \rho$. The probability of a move among the observed sample is then

$$\Pr(M = 1 | X, Z, R^* > 0) = \frac{F(\alpha' W, \beta' W, \rho)}{F_R(\beta' W)}$$

where $F(\varepsilon^M, \varepsilon^R, \rho)$ is the joint c.d.f. of the error terms. Again estimates of the mobility equation based only on the uncensored observations will be biased.

The censoring of outcomes for identical individuals will be different under the two response mechanisms. The standard mechanism shown in (3) is more appropriate when unobserved individual characteristics influence both the outcome and response. In this situation the problem is one of omitted variables; if all relevant characteristics could be observed and controlled ρ would be driven to zero and estimation based on the observed sample would be unbiased. In contrast, the mechanism in (2) is more appropriate when the outcome exerts a causal effect on response, as in our mobility example. Complete case estimates will remain biased even if all relevant influences are observed and controlled.

Two-equation maximum likelihood models can be used to fit both the direct dependence and standard selection models. [Derivation of the likelihoods in both cases are given in the full version of the paper.]

Simulation analyses

We use simulation analyses to explore the consequences of different estimation methods when the true data generating process (DGP) is given by equations (1) and (2) above. Each model is estimated 100 times on data generated for a sample of 100,000 individuals. The analysis is repeated for a series of DGPs in which the signs of the key underlying parameters - α_1, β_1 and γ - are varied, and for two choices of the error distribution, first with $F_j(\varepsilon^j)$ as the standard normal c.d.f. and then as a right-skewed distribution based on a transformed gamma variable. (The latter distribution incorporates the idea that there exist individuals with extreme negative propensities both to move – "stayers" – and to respond.)

Table 1 summarizes the results of ignoring non-response and estimating a single-equation probit model on the observed sample only, for eight different DGPs, all with normally-distributed errors. We find estimates of $\hat{\alpha}_1$ are biased in all cases, and that the direction of the bias depends on the signs of γ and β_1. When they are of the same sign, so that $\gamma\beta_1 > 0$, $\hat{\alpha}_1$ is biased downwards. When they are of opposite sign, so that $\gamma\beta_1 < 0$, $\hat{\alpha}_1$ is biased upwards.

Applying this to our analysis of unemployment and mobility, we have already argued that in this application we expect $\gamma < 0$. Prior research suggests that the unemployed are less like to respond to surveys than the employed, so we expect $\beta_1 < 0$, and that unemployment is positively associated with mobility, so that $\alpha_1 > 0$. If this is the case (the one illustrated in row 1 of Table 1) single equation estimates that ignore selection will tend to underestimate the effect of unemployment on mobility.

Subsequent results focus on $bias(\hat{\alpha}_1)$ and the frequency the null hypothesis of independence between the outcome and response equations is correctly rejected when alternative selection models are fit to the data. We explore the sensitivity of these results to the introduction of exclusion restrictions and non-normal errors. Appendix Table A1 shows preliminary results from 10 simulations for a particular set of true coefficient values (corresponding to the first row of Table 1).

Table 1. Bias in single equation probit estimates fit to a direct dependence DGP: Summary of 100 simulations

True values			Uncensored	$bias(\hat{\alpha}_1)$	$bias(\hat{\alpha}_1)$	Proportion
α_1	β_1	γ	obs (mean)	(mean)	(SD)	$\hat{\alpha}_1$ biased*
0.5	-1	-1	82,535.5	-0.319	0.025	1.00
0.5	1	-1	85,729.5	0.140	0.017	1.00
0.5	-1	1	89,077.5	0.171	0.018	1.00
0.5	1	1	91,098.0	-0.044	0.016	0.71
-0.5	-1	-1	83,144.4	-0.219	0.038	1.00
-0.5	1	-1	85,863.4	0.142	0.023	1.00
-0.5	-1	1	88,641.9	0.121	0.025	1.00
-0.5	1	1	91,080.5	-0.043	0.021	0.45

Sample size 100,000. Each row summarizes the results of 100 estimates calculated from samples with different combinations of true parameter values. Errors simulated from the standard normal distribution. Probit models (probit command in Stata11) fit to observations where $R^* > 0$. $bias(\hat{\alpha}_1) = \hat{\alpha}_1 - \alpha_1$.
*$\hat{\alpha}_1$ is biased if the estimated p-value of the z-statistic on $bias(\hat{\alpha}_1)$ is less than .05.

Empirical application

We use data from the BHPS, which has interviewed a representative sample of about 5,500 households covering more than 10,000 individuals annually from 1991. We use data from Waves 5 to 18 (1996 to 2008) on individuals resident in England and Wales at *t*. We limit our analysis to the sample of single individuals (i.e. without a live-in partner) aged between 18 and 60, thereby precluding the need to consider interactions between the characteristics of one's partner and one's own mobility decision. The estimation sample consists of 25,332 person-year observations on 6,489 individuals (mean 6.5 year observations per person). The outcome variable is equal to 1 if the individual moved to a different residence within the same local authority between *t* and *t+1* and zero otherwise (longer distance moves are coded as zero). The outcome is observed for 86.9% of observations, among which the mobility rate is 11.6%. Covariates (measured at *t*) included in both equations include labor market status (with unemployment as one category), educational qualifications, age and its square, housing tenure, gender, and year and local authority dummies. We employ two exclusion restrictions. First we assume what the rate of house price inflation in the local authority of residence at *t* predicts mobility but not response. Second we assume that an individual's sample membership status affects response but not mobility. Sample membership reflects whether the individual was recruited to the panel as part of the original sample or whether they joined later for one of several different reasons, and has been used for this purpose by other researchers[2]. A number of one- and two-equation models, each allowing for a time-invariant person-specific random effect, are fit to the data using maximum likelihood and the results are compared. Preliminary results suggest that single equation methods tend to underestimate the effect of unemployment on mobility.

[2] E.g. Cappellari, L. and Jenkins, S.P. (2008). 'Estimating low pay transition probabilities accounting for endogenous selection mechanisms', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2:165-186.

Appendix Table A1. Summary of selected statistics from 10 simulations of alternative estimation approaches

		$bias(\hat{\alpha}_1)$ Mean (SD)	Proportion $\hat{\alpha}_1$ biased*	$\hat{\gamma}$ Mean (SD)	Proportion $\hat{\gamma}$ sig**	$\hat{\rho}$ Mean (SD)	Proportion $\hat{\rho}$ sig***
Normal errors							
1	1-eq Probit	-0.319	1.00	-	-	-	-
		(0.025)					
2	2-eq DD	0.010	0.00	-1.041	1.00	-	-
		(0.042)		(0.150)			
3	2-eq DD w restrictions	-0.007	0.10	-0.987	1.00	-	-
		(0.026)		(0.037)			
4	2-eq Heckprobit	-0.149	0.00	-	-	-0.311	0.20
		(0.083)				(0.143)	
5	2-eq Heckprobit w restrictions	-0.045	0.30	-	-	-0.495	1.00
		(0.032)				(0.032)	
Skewed errors							
1	1-eq Probit	-0.222	1.00				
		(0.023)					
2	2-eq DD	-0.074	0.70	-0.547	0.60	-	-
		(0.119)		(0.561)			
3	2-eq DD w restrictions	0.041	0.40	-0.879	1.00	-	-
		(0.030)		(0.037)			
4	2-eq Heckprobit	0.241	1.00	-	-	-0.864	1.00
		(0.021)				(0.016)	
5	2-eq Heckprobit w restrictions	0.098	0.80	-	-	-0.625	1.00
		(0.032)				(0.027)	

Sample size 100,000. Each row summarizes the results of 10 estimates calculated from samples with true parameter values $\alpha_1 = 0.5, \beta_1 = -1, \gamma = -1$. DD estimates use a user-specified likelihood command; Heckprobit estimates use the heckprob command in Stata11. Models with restrictions include a predictor in each equation that is excluded from the other equation. $bias(\hat{\alpha}_1) = \hat{\alpha}_1 - \alpha_1$

*$\hat{\alpha}_1$ is biased if the estimated p-value of the z-statistic on $bias(\hat{\alpha}_1)$ is less than .05.

** $\hat{\gamma}$ is significant if the estimated p-value of the z-statistic on $\hat{\gamma}$ is less than .05.

*** $\hat{\rho}$ is significant if the estimated p-value of the χ^2 test of independent equations is less than .05.