AUXILIARY VARIABLES AND ADJUSTMENTS FOR MISSINGNESS IN LONGITUDINAL STUDIES

IAN PLEWIS

SOCIAL STATISTICS, UNIVERSITY OF MANCHESTER,MANCHESTER M13 9PL,UK.

ian.plewis@manchester.ac.uk

Abstract

Different sorts of auxiliary variables – variables measured at previous waves, frame variables and paradata - can be used to improve the accuracy of response propensity models, and to enhance adjustments for missing longitudinal data. All these variables are used in this paper when constructing iterative probability weights, carrying out multiple imputations, and specifying models that jointly model a substantive process and the missingness mechanism. Data from the first two waves of the UK Millennium Cohort Study are used to illustrate the potential value of auxiliary variables. We find that the accuracy of response probability models – as measured by the area under the Receiver Operating Characteristic curve – is improved by the inclusion of frame variables and paradata but these variables have rather little effect when adjusting the chosen longitudinal estimates. There is, however, evidence to suggest that unobserved variables are correlated with the outcome of interest and with the probability of being a respondent at wave two.

Keywords:

Auxiliary variables; multiple imputation; paradata; response propensity models; ROC curves; selection models.

AUXILIARY VARIABLES AND ADJUSTMENTS FOR MISSINGNESS IN LONGITUDINAL STUDIES

The mid twentieth century pioneers of the longitudinal method in the social sciences would be impressed by the rising popularity of this kind of study. There is, however, one aspect of the longitudinal approach that exercises current practitioners just as much as it did the early researchers: the fact that cases are lost from selected samples over time and this loss is both cumulative and systematic. There is an increasing recognition in the social sciences, to some extent following a trend set in epidemiology (Sterne et al. 2009), that the problems of potential bias and loss of efficiency that are associated with missingness should be routinely addressed by analysts. The intention of this paper is to make some progress towards a resolution of the problems of bias and loss of precision that arise when longitudinal data are missing, by drawing on the contribution that different kinds of auxiliary variables can make to predicting and adjusting for missingness.

Apart from the problems of unit and item non-response that are common to all observational studies, data can be missing from a longitudinal study for a number of reasons:

1. Some units drop out after the first or subsequent waves, never to return. This is attrition; the extent of this absorbing state depends on the resources put into tracking, and on procedures relating to which cases are issued to the field at each wave. It is often only possible to ascertain attrition cases with certainty at the end of a study.

2. Some units move in and out of the sample over time. This is wave non-response which will only be observed if non-responding cases at wave t are reissued to the field at wave t + k ($k \geq 1$).

3. Both the attrition cases and the wave non-respondents can be sub-divided into three main categories: (i) not located, (ii) not contacted, conditional on being located, and (iii) refusing to cooperate, conditional on being contacted (Lepkowski and Couper 2002).

It is often argued (by, for example, Groves 2006) that the keys to unlocking missingness problems of bias are to find those variables that predict whether a piece of data is missing, and which of those variables that predict missingness are also related to at least one out of possibly many outcomes of interest. In this sense, longitudinal researchers are at an advantage compared with analysts of cross-sectional data because they can draw on a wider range of potential predictors:

A. Variables of substantive interest that are measured on all the responding cases in the first wave of the study (although item non-response might affect some of these variables).
B. As in cross-sectional studies, some information is available from the sampling frame for all sampled cases at the first wave.
C. Variables that are related to aspects of data collection, either derived from administrative procedures used, for example, to track sample members over time or,

as in cross-sectional studies, from data collected from respondents and interviewers during fieldwork.

These three groups of variables are sometimes referred to collectively as auxiliary variables although sampling statisticians tend to reserve this term for the variables derived from sampling frames and population registers (i.e. group B). Variables in group C are sometimes (as here) labelled paradata (Couper 1998), but are also referred to as instruments especially in the econometric literature on adjusting for non-response (e.g. Fitzgerald, Gottschalk and Moffitt 1998).

Response propensity models, often based on the variables in group A, are widely employed to adjust for longitudinal non-response, using inverse probability weights (IPW) derived from the estimated probabilities of response from the model. These models can also be used to improve applications of procedures that impute missing data, in particular multiple imputation. This paper considers the potential added value of using variables from groups B and C when adjusting for missingness. It is common to use the frame variables to adjust for unit non-response but they are less often used to adjust for missingness after the first wave. If we find that the variables in group C, the paradata, appear to improve our adjustment methods then this has implications for the kinds of data that might routinely be collected in longitudinal studies.

The approach taken in this paper is related to the one taken by Kreuter et al. (2010) who focused on cross-sectional survey data. They examine auxiliary variables obtained from sampling frames and different kinds of interviewer observations in five surveys and show

that very few of these auxiliary variables are strongly associated with both the propensity to respond and their outcome variables of interest. Nevertheless, their findings indicate that mean-square error in measures of central tendency can be reduced by using IPW derived from response propensity models that include these auxiliary variables. Analysts of cross-sectional surveys are often most interested in the distributions of survey variables but longitudinal surveys are designed primarily to measure and model change and so we need to determine whether there is any added value from paradata and frame variables when applied to the adjustment of measures of change. Kreuter et al. (2010) focus on the use of IPW to reduce non-response bias whereas more attention is given here to different kinds of models that adjust for missingness.

The rest of the paper is organised as follows. The next section outlines the approach taken to predicting non-response and Section 3 describes methods of adjusting for missingness. This is followed by a brief description of the study - the Millennium Cohort Study - and the underlying research question used to illustrate the ideas in this paper. Section 5 presents analyses that show (i) to what extent prediction of non-response is improved by including frame variables and paradata and (ii) how these variables can be used to adjust for missingness, and what effects they have on estimates of interest. The concluding section includes some discussion of data collection issues.

Predicting non-response

There are many instances in the literature of studies that have modelled the predictors of non-response in longitudinal surveys: for example, Behr, Bellgardt and Rendtel (2005); Hawkes and Plewis (2006); Watson and Wooden (2009) and, for the Millennium Cohort

Study, Plewis (2007a) and Plewis et al. (2008). A defining characteristic of these response propensity models is that a binary or categorical outcome is linked to a set of explanatory variables, using either a logit or probit link (or their multivariate equivalents). An important, but somewhat neglected topic is how the accuracy of these predictive models should be assessed. This issue is considered in Plewis, Ketende and Calderwood (2012) and so only a précis of their discussion is presented here.

A widely used method of assessing accuracy of models for binary or categorical outcomes is to estimate the goodness-of-fit by using one of several possible pseudo-R^2 statistics. Apart from their rather arbitrary nature, which thus makes comparisons across datasets difficult, estimates of pseudo-R^2 are not especially useful in this context because they assess the overall fit of the model rather than distinguishing between the accuracy of the model for discriminating between non-respondents (the true positive rate) and respondents (the true negative rate) separately. Consequently, we use a measure based on Receiver Operating Characteristic (ROC) curves, illustrated in Fig. 1 as the plot of the sensitivity or true positive rate against the false positive rate. Krzanowski and Hand (2009) give a detailed discussion of how to estimate ROC curves.

The area enclosed by the ROC curve and the x-axis in Fig. 1, known as the AUC (area under the curve), is of particular interest and this can vary from 1 (when the model for predicting non-response perfectly discriminates between respondents and non-respondents) down to 0.5, the area below the diagonal (when there is no discrimination between the two categories). The AUC can be interpreted as the probability of assigning a pair of cases, one respondent and one non-respondent, to their correct categories, bearing in mind that

guessing would correspond to a probability of 0.5. A linear transformation of AUC (= 2*AUC – 1), often referred to as a Gini coefficient, is commonly used as a more natural measure than AUC because it varies from 0 to 1. Our interest here is in determining to what extent the inclusion of frame variables and paradata in response propensity models increases the AUC.

Adjusting for missingness

Carpenter and Plewis (2011) set out the three main ways of adjusting for missingness in longitudinal studies that are considered here: (i) inverse probability weighting (IPW); (ii) multiple imputation (MI); (iii) jointly modelling the substantive process and the missingness mechanism. As we shall see, each of these three approaches addresses somewhat different combinations of missingness.[(1)]

The roots of IPW go back to the application of survey weights to correct for unequal selection probabilities as represented by the Horvitz-Thompson estimator. The method is based on a model that predicts non-response as just described. The approach is widely used to adjust for attrition and wave non-response; it is easily understood and straightforward to apply across the board. There are, however, disadvantages to IPW. It is a 'one size fits all' approach based on adjusting for differential probabilities of responding but ignoring any association between the predictors of response and the outcome (and model) of interest and consequently has the potential to introduce inefficiencies into the analysis (Little and Vartivarian 2005). In addition, it is only possible to estimate weights for cases with complete data on the predictors of non-response.[(2)] Moreover, IPW does not adjust for item non-response in the outcome and explanatory variables of interest. The fact that weights are

estimated (and therefore subject to sampling error) should also be incorporated into analyses when the sample size for the response propensity model is not large.

Imputation methods also have a long history but it is now generally recognised that computer-intensive methods of MI are the most satisfactory. Historically, imputation methods have been seen as ways of adjusting for item non-response but current thinking, especially in a longitudinal context (e.g. Goldstein 2009), indicates that they can be used to adjust for attrition and wave non-response as well, given the availability of information from previous (and future) waves. The strength of MI is its focus on a model of interest and the link between this model and models for missingness. More details of the approach to MI taken in this paper are given later after introducing the model of interest.

Both IPW and MI assume that data are missing at random (MAR), i.e. that missingness is ignorable conditional on the chosen set of predictors. This is not an assumption that can be supported by the data and there can be grounds for assuming that, in the longitudinal context, missingness depends on the wave t value of the variable, even after conditioning on measured wave t-k variables. In other words, the missingness mechanism can be informative or non-ignorable and then the data are 'missing not at random' (MNAR). One way of dealing with the problem of non-ignorable missingness is to jointly model the model of interest and the missingness mechanism and then either assume something about the joint distribution of the residuals (as in Heckman-type models) or introduce further information in the form of a prior distribution for the missingness as in the Bayesian models used by, for example, Mason et al. (2012) but not considered here. Again, more details are given later.

The Millennium Cohort Study

The wave one sample of the UK Millennium Cohort Study (MCS) includes 18,818 babies in 18,552 families born in the UK over a 12-month period during the years 2000 and 2001, and living in selected UK electoral wards at age nine months. As practically all mothers of new-born babies in the UK were, at that time, eligible to receive Child Benefit, the Child Benefit register was used as the sampling frame. The initial response rate was 72%. Areas with high proportions of Black and Asian families, disadvantaged areas and the three smaller UK countries are all over-represented in the sample which is disproportionately stratified and clustered by electoral ward as described in Plewis (2007b). The design weights vary from 2.0 (England advantaged stratum) to 0.23 (Wales disadvantaged stratum). The first four waves took place when the cohort members were (approximately) nine months, 3, 5 and 7 years old. Partners were interviewed whenever possible and data were also collected from the cohort members themselves and from their older siblings.

The first two waves of MCS provide the model of interest for this paper, based on work by Plewis and Kallis (2008) on the effect of family income on children's educational attainment. The aim here is to establish whether using the extra variables from groups B and C to adjust for missingness changes (i) estimates of the mean of the outcome variable of interest, the Bracken test of educational preparedness; (ii) estimates of the regression coefficients in models that relate this outcome to change in family income between waves one and two.

Analysing the effects of auxiliary variables

Predictive models for non-response in the second wave of MCS

The variables that are potentially predictive of non-response at wave two of MCS are divided into three groups corresponding to groups A, B and C described earlier:

1. Variables of substantive interest that were measured at wave one and shown to be predictive of non-response at wave two by Plewis (2007a). These are described in more detail in Appendix 1.
2. Variables that define the sample design: (i) the nine strata and (ii) the 398 primary sampling units (electoral wards).
3. Paradata:

   (a) Variables collected at wave one but unlikely to be of any substantive interest – (i) giving consent to link health records to the interview data (although the records themselves will be of substantive interest); (ii) providing a stable address for tracking purposes; (iii) refusing to answer the question about family income; (iv) either no response or a proxy response from the partner of the main respondent. More details can be found in Appendix 1.

   (b) Two variables collected after wave one and substantively important in some contexts: (i) whether the family containing the cohort child changed address between waves one and two, generated from data collected by the survey administration team for tracking purposes and described in more detail in Plewis et al. (2008); (ii) an assessment by the survey interviewers of neighbourhood conditions at wave two – see Appendix 2 for more details.

The first aim of the analysis is to show how the accuracy of prediction of (i) overall non-response; (ii) non-response separated into wave non-response and attrition; (iii) refusal and other non-productive (i.e. not located combined with not contacted) separately, changes as the variables from groups 2 and 3 above are added to the baseline logistic and multiple

logistic models using the variables in group 1. The results are summarised in Table 1 where the first row shows the prevalence of overall and different categories of non-response.

The results for model A are based on binary and multinomial logistic regression models using the explanatory variables in group 1 and excluding all aspects of the sample design. We see that the overall levels of predictive accuracy are not high generally, although with a marked contrast between predicting refusal (Gini estimate = 0.28) and predicting other non-productives (Gini = 0.47).

The primary sampling units are introduced into models B to D as a random effect (i.e. a random intercept) so that we have a two level model (cohort members within electoral wards) with two or three outcome categories. The model for three categories is essentially the same as that used in a related context by Durrant and Steele (2009):

$$\log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(0)}}\right) = \sum_{k=1}^{K} \beta_k^{(s)} x_{kij} + \sum_{p=1}^{P} \gamma_p^{(s)} z_{pj} + u_j^{(s)} \quad (1)$$

where:

$\pi_{ij}^{(s)}$ = P($y_{ij} = s$); s = 1, 2 (refusal, other non-productive *or* wave non-response, attrition) and $\pi_{ij}^{(0)}$ is the probability of responding, for household i ($i = i..n_j$) in cluster j ($j = 1..J$). Note that s = 1 (not responding) for the binary model.

x_{kij} are individual level explanatory variables;

z_{pj} are dummy variables defining the nine strata;

$u_j^{(s)}$ are random effects at level two representing residual variability between clusters and following a bivariate Normal distribution (or a univariate Normal when the outcome is binary).

All the models were estimated using Markov chain Monte Carlo methods available in the *MLwiN* software (Rasbash et al. 2009; Browne 2009), based on 40000 chains with non-informative priors throughout and supplemented by orthogonal parameterisation and, for the binary models, parameter expansion as described by Browne et al. (2009) to improve convergence.

The predicted values $p_{ij}^{(s)}$ from equation (1) (used to estimate the AUC and Gini coefficients) include the empirical Bayes estimates $\hat{u}_j^{(s)}$ of the proportion of non-response for each cluster as follows:

$$p_{ij}^{(s)} = \exp\left(\sum_{k=1}^{K} b_k^{(s)} x_{kij} + \hat{u}_j^{(s)}\right) / 1 + \exp\left[\sum_{r=1}^{2} \left(\sum_{k=1}^{K} b_k^{(r)} x_{kij} + \sum_{p=1}^{P} c_p^{(r)} z_{pj} + \hat{u}_j^{(r)}\right)\right] \quad (2)$$

with a simpler version of (2) for the binary model, and where b_k and c_p are estimates of β_k and γ_p.

The results are given for model B in Table 1 and indicate that including the primary sampling units explicitly in the model does noticeably improve the accuracy of prediction for the different types of non-response (although less so for overall non-response). This improvement might have arisen because these second level units carry information about the interviewers assigned to them although we do not have enough information about interviewer assignment to test this. The estimated between cluster variances and covariances are given in Table 2; we see that the cluster level proportions of refusals and

other non-productives are associated (r = 0.36) and this association is more marked for attrition and wave non-response (r = 0.76).

A set of eight dummy variables for the strata is added to these models (model C); the stratification is related to all types of non-response but including this part of the sample design does not improve predictive accuracy.

Turning to model D in Table 1 then, from the variables collected at wave one, 'consent' and 'no partner response' predict all types of non-response; 'stable address' predicts only attrition and refusal; 'refusing income' predicts only wave non-response. Considering the two variables collected after wave one, refusals and attrition are less likely when households move, and other non-productives and wave non-respondents are less likely if households were living in flats at wave one and then move address. The scale generated from the neighbourhood observations predicts all types of non-response but predicts wave non-respondents and other non-productives better than it does attrition and refusal.[(3)] The addition of the paradata variables (model D) further improves the discrimination of the model with estimates of Gini coefficients ranging from 0.45 to 0.61 in Table 1 although there are two caveats to these results. The first is that there is a degree of circularity when using the residential mobility variable in that not located households are bound to be mobile. When the mobility variable is excluded from the 'other non-productive' part of the model, the Gini coefficient falls, but only from 0.61 to 0.60. Second, the interviewers collected observations about the neighbourhood for all cases apart from those that were not located at wave two and this limits the applicability of this variable both for prediction and for adjusting for missingness. [(4)]

Using frame variables and paradata to adjust for missingness

The question of the effectiveness of frame variables and paradata when adjusting for missingness is best considered within the context of a substantive model of interest; here the effect of family income on young children's educational attainment as discussed by Plewis and Kallis (2008). We consider data from two waves and regression models of the form:

$$y_{ti} = f\left(x_{ti}, x_{t-1,i}\right) + e_i \quad (3)$$

The wave two (i.e. $t = 2$) outcome variable is child i's score on the Bracken test, the explanatory variables are log family income at waves one (x_1) and two (x_2). Three sub-models are considered:

(i) The two income variables enter the model separately.

(ii) The difference between log income at wave two and log income at wave one, i.e. the log of the income ratio, enters the model on its own.

(iii) The difference in the untransformed incomes enters the model on its own.

Having established that the frame variables and paradata are associated with the propensity to respond, we now need to consider whether they are associated with the variables in the model of interest.[5] Table 3 shows that the frame variables and the neighbourhood assessment score are associated with the variables in the model of interest but the correlations for the other paradata are small (≤ 0.10) and these variables are therefore unlikely to make a substantial contribution to adjustments that use either weighting or multiple imputation.

Applying inverse probability weights

The design weights are combined with the inverse probability weights to generate overall weights whose effect can be compared with using just the design weights. We see from the first row of Table 4 that the overall mean of the outcome variable is lower once we adjust for the estimated probabilities of responding, indicating that children with lower attainments are more likely to be missing at wave two. There is a more marked reduction when using the weights from model D but these weights are only adjusting for non-response due to refusals and not contacted given the restriction in the collection of the neighbourhood observations.[(6)] The estimated regression coefficients for sub-model (i) are essentially unchanged by the application of non-response weights. The application of weight A to sub-models (ii) and (iii) does lead to changes in the estimated coefficients but once the effects of the clusters are introduced into the weights, the estimates become closer to the estimates based just on the design weights. The differences between the estimates using weight B and those using weights C and D for sub-models (ii) and (iii) are small. The variability in the overall weights A to D is greater than the variability in the design weights although not substantially so, as indicated by the weight range.

Multiple imputation

We consider two approaches to MI applied to the model of interest (MoI) previously introduced. First, we assume multivariate Normality for the variables in the MoI. We then apply the data augmentation method, a Bayesian procedure described by Little and Rubin (2002: Ch. 10), consisting of two steps that are repeated until convergence is achieved: an imputation step in which the missing data is imputed by drawing from a distribution generated by the observed data and the model parameters from the previous iteration, and

a posterior step in which the model parameters are drawn from a distribution generated by the observed data and the missing data imputed in the previous step. This procedure is replicated at least five times and the estimates from each replication are combined using Rubin's rules as described in Carpenter and Plewis (2011). Data augmentation is implemented in the *mi* suite of programs in STATA11 and is used here with uniform prior distributions on all the parameters thus making the procedure equivalent to maximum likelihood.

Table 5 presents the patterns of missing data for the MoI after eliminating the very few cases (n = 237) with missing data (i.e. item non-response) on one or more of the auxiliary variables in groups 1 to 3 (with the exception of the neighbourhood assessment score). We see that the complete cases (CC) comprise only 59% of all cases and that 12% of the cases do not confirm to a monotonic pattern of non-response.

MI is applied to missingness scenarios as represented by models A, C and D in Table 1 but without allowing for any influence of the clusters in models C and D. The neighbourhood assessment score is included in model D although it becomes part of the imputation process because of the missing data for this variable. This does imply that MI is adjusting not only for attrition and wave non-response as IPW does but also for item non-response in the explanatory variables in the MoI and, if necessary, in the auxiliary variables. However, as Little and Zhang (2011) point out, if item non-response in the explanatory variables is informative then the assumptions underpinning MI break down and CC analysis might be preferred.

Table 6 gives the results and shows that, for sub-model (i), there is a small change in the estimated coefficient for log family income at wave two and a reduction in the estimated standard errors compared with the CC analysis in the first column of Table 4 so that the confidence intervals are about 15% narrower. These small gains in efficiency are obtained from the variables in group 1; the inclusion of the paradata does not have any effect on the point estimates based on MI although the standard errors are slightly higher. For sub-models (ii) and (iii), however, we find that MI generates estimates that are different from those based on complete cases in Table 4 with the estimate for log income difference being reduced whereas the estimate for income difference is substantially higher. Again, including the paradata in the model does not change the point estimates but it does reduce the standard errors. We return to the possible implications of these results in the concluding section.

There are, however, two drawbacks to the data augmentation approach just described. The first is the assumption of multivariate Normality that might not always be sustainable. The second is that it cannot easily incorporate the effects of the clusters in the data generated by the sample design. Reiter, Raghunathan and Kinney (2006) identify this as a potentially serious problem for complex designs. It is, however, possible to deal with the second problem using a two level MI method illustrated in Goldstein (2009) that incorporates a random effect into both the model of interest and the imputation equations in such a way that there is a set of underlying equations for all the variables with missing data, each of which includes a random intercept to account for the clustering. Reiter et al. (2006) suggest representing the clusters as fixed effects but this is not attractive when there are large numbers of clusters. Thus:

$$\boldsymbol{y}_{ij} = \boldsymbol{\beta_{00}} + \sum_{k=1}^{K} \boldsymbol{\beta_k} x_{kij} + \boldsymbol{u_j} + \boldsymbol{e_{ij}} \quad (4)$$

where **y** is a vector of p variables measured on case *i* (*i = 1..n*$_j$) in primary sampling unit *j* (*j = 1..J*) with, in general, different numbers of cases missing for each y_p; x_k are the k auxiliary variables with no missing data; the **β** are vectors of regression coefficients; **e** and **u** are the sets of level one and level two residuals respectively arranged as vectors and uncorrelated across levels with **e** ~ MVN(**0, Φ**) and **u** ~ MVN(**0, Σ**).

It is important to note that the MoI includes a random effect for the sample clusters and so the estimated regression coefficients of interest are now pooled within cluster effects rather than the combination of within and between cluster effects that characterises the earlier models. Estimation proceeds using MCMC as built into the REALCOM-IMPUTE software (Goldstein et al. 2008). The results are given in Table 7 and show small effects of applying MI in sub-model (i); rather more marked effects of change in income on the education outcome compared with the CC analysis and slight changes when the paradata are used.

Joint models

We can jointly model the substantive model of interest and the selection equation, i.e. the probability of observing the outcome *y,* as follows:

$$y_i = \sum_{k=0}^{K} \beta_k x_{ki} + e_{1i} \quad (5a)$$

$$P\left(\sum_{q=1}^{Q} \alpha_q z_{qi} + e_{2i} > 0\right) = \Phi\left(\sum_{q=1}^{Q} \alpha_q z_{qi}\right) \quad (5b)$$

where (5a) is a multiple regression model with e_1 ~ N(0, σ^2); Φ is the standard Normal distribution function in 5(b) so e_2 ~ N(0,1) and (5b) is a probit model; e_1 and e_2 are bivariate

Normal with corr(e_1e_2) = ρ implying that unobserved variables that are correlated with being missing are also correlated with the outcome via e_1 and its correlation with e_2. The auxiliary variables are contained within ***z*** and are needed to identify the joint model but ***z*** can also contain *x* variables from (5a). Note that, in terms of our example, (5) models both missingness due to attrition and wave non-response, and also item non-response in the Bracken test, and allows all that missingness to be associated with *y* conditional on *x,* i.e. MNAR. Table 5 shows that there are 1077 productive interviews at wave two with a missing outcome, information that is also used in MI but not when inverse probability weights are applied. In contrast to MI, however, no adjustment is made here for missingness in *x* and *z* variables.

Models like (5) are discussed in detail by Vella (1998) and maximum likelihood estimates can be obtained using the *heckman* procedure in STATA. The estimates are, however, sensitive both to the assumption of bivariate Normality and to the specification of the selection model. Table 8 gives the results from models whose specifications are the same as those used earlier except that log family income at wave one is included in the selection equation. The main points to emerge from the estimates are:

(i) There are substantial unobserved selection effects as indicated by the high negative correlations between the estimated residuals, with unobserved variables that increase the probability of not responding and reduce the score on the Bracken test. These correlations are not reduced, contrary to expectations, by the introduction of the paradata (model D).

(ii) The estimates for income at wave two conditional on income at wave one are lower than those from the CC analysis and do not vary as the specification of the

selection equation changes. The estimate for log family income at wave one in the selection equation is consistently small.

(iii) The estimates for log income difference and income difference are higher than they are when based on CC but reduced by the inclusion of the paradata. The estimates for log family income at wave one in the selection equations are now consistently about six times greater than their standard errors.

Discussion

We have seen that when variables of substantive interest that are measured at the first wave of a longitudinal study are used in response propensity models, they enable us to distinguish between respondents and non-respondents although the degree of discrimination is not high. The level of discrimination improves when at least some aspects of a stratified and clustered design are included in the model. Moreover, our example illustrates that those variables often referred to as paradata can further improve discrimination so that the probability of distinguishing a respondent from non-respondents of different kinds rises from less than 0.7 to close to 0.8. Thus, if our interest is in preventing non-response after the first wave of a longitudinal study, frame variables and paradata could help to target more effectively procedures and interventions designed to improve response rates.

If, however, our interest is in adjusting for longitudinal missingness both in terms of reducing bias and increasing precision in estimates that are specifically concerned with measuring and modelling change then our example casts some doubt about the usefulness of frame variables and paradata. This is partly because the paradata in particular are not

strongly associated with the outcome in the MoI and cannot therefore reduce any bias that might be present in the estimates as a result of selective non-response. On the other hand, we do see that adjustments for missingness vary in magnitude according to how the model relating change to the outcome is specified. When the focus is on the regression coefficient of income at wave two conditional on income at wave one then we find only marginal changes from the CC estimates in both the point estimates and the standard errors regardless of whether weights or MI are used, and irrespective of which variables are used in the weighting scheme and the imputations. But when change is represented by a ratio of two variables or the difference between them then adjustments for missingness are greater and do vary according to which variables are included in the adjustment process. One possible explanation for this is that the missingness mechanism for the outcome is close to being MAR conditional on income at wave one and no further adjustment is required for models like sub-model (i) whereas there is no explicit conditioning on the wave one measure when ratios and differences are used. Against that, we do see that the estimate of the coefficient of wave two income conditional on wave one income is about 25% lower when we allow for unobserved selection effects that are correlated with the outcome, suggesting that if the assumptions of the selection model hold then there are aspects of the response process which we do not understand and which the available paradata do not appear to capture. Paradata do, however, have a potentially important part to play if the MAR assumption is not satisfied because their association with response behaviour and their irrelevance to the MoI means that they can be useful instruments that help to identify the joint model and thus improve the robustness of estimates from these models. This can be particularly important when the MoI includes more explanatory variables as controls – as

a fuller investigation of the relation between income and educational attainment would do - and these explanatory variables also predict non-response.

This paper only considers data and models from two waves of a longitudinal study. A more complete understanding of the relation between, say, education and income would come from using data from several waves and then modelling the relation both within and between individuals as described by Plewis (2001). It would be interesting to explore the value of paradata in adjustments for missingness in these circumstances.

It is, of course, possible that the inclusion of other paradata in the adjustment processes considered here might have made more of a difference. Candidate variables include the length of the interview at wave one, the proportion of questions answered at wave one (a variable found to be predictive of non-response by a number of researchers), the reluctance of the cohort child to attempt the educational tests, and variables like interviewer gender, ethnic group, age and experience. It is also possible to ask the interviewers to record observations about their contact with respondents, and to ask respondents to describe their experience of the interview and interviewer. Both these approaches were used in wave four of MCS and can be used to predict and adjust for non-response at wave five. On the other hand, there is a cost attached to collecting paradata. The question of whether it is cost-effective to collect more variables of this kind must depend in part on an assessment of the possibilities and benefits of reducing the deleterious effects of non-response that they bring. One important issue that is widely ignored when considering longitudinal missingness is unit non-response at wave one of a longitudinal study. Frame variables can be used to make adjustments but are not likely to be adequate in this case. Data from administrative

registers are not always available at a case level and good quality paradata generated from interviewer observations and interactions with potential respondents might, as Kreuter et al. (2010) indicate, play a useful part in adjusting measures and models for change for this initial non-response.

Footnotes

(1) A fourth method – calibration as described by, for example, Sikkel, Hox and deLeeuw (2009) – is not used in this paper.

(2) Carpenter and Plewis (2011, p.533) propose a way of reducing this problem by replacing missing weights by estimated weights from calibrations.

(3) Detailed results are available on request.

(4) Interviewer observations of the neighbourhood were not collected at wave one.

(5) It is only possible to do this for the observed cases.

(6) Weights generated from response propensity models should be applied only if the models exhibit 'common support' in the sense that there are respondents and non-respondents across the distribution of propensity scores. All the response propensity models referred to in Table 1 satisfy this condition, verified by dividing the distribution into twenty bands of equal sample size and tabulating the numbers of respondents and non-respondents in each band.

Figure 1: ROC curve

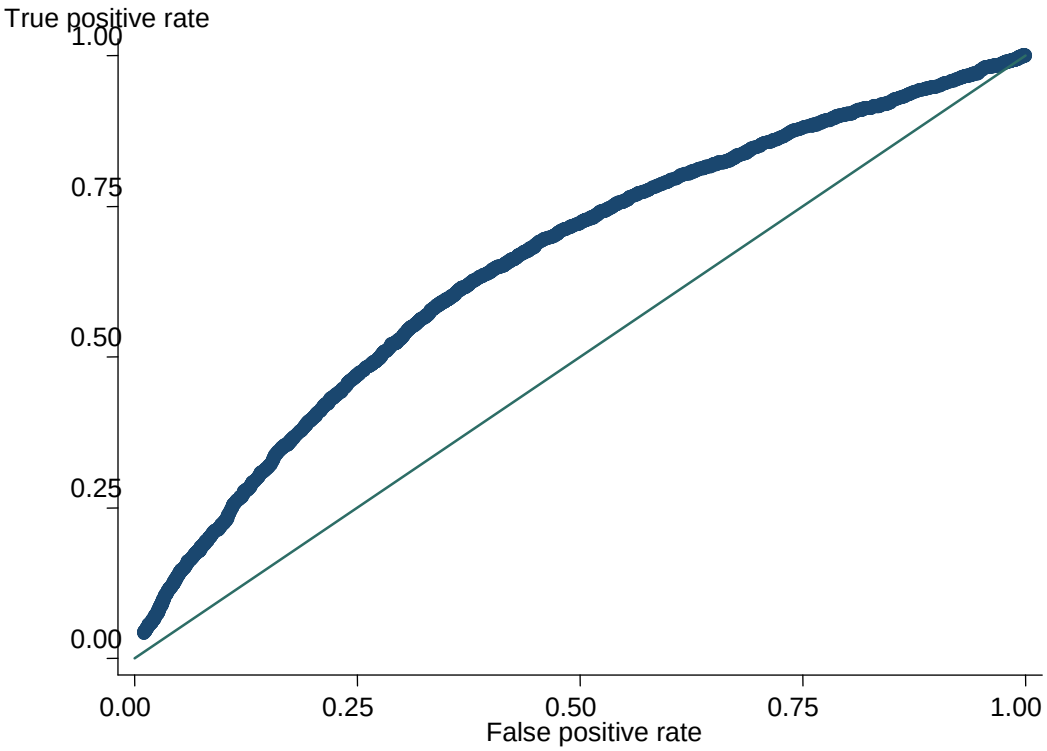

Table 1: AUC and Gini coefficients

Model	Overall non-response (19%)	Non-response (type 1)		Non-response (type 2)	
		Wave non-response (8%)	Attrition (11%)	Refusal (10%)	Other non-productive (10%)
A	0.68 (0.0049) 0.36 (16675)	0.69 (0.0075) 0.38 (14871)	0.67 (0.0067) 0.35 (15411)	0.64 (0.0071) 0.28 (15130)	0.73 (0.0066) 0.47 (15152)
B	0.69 (0.0058) 0.37 (16675)	0.75 (0.0062) 0.50 (14871)	0.70 (0.0069) 0.40 (15411)	0.70 (0.0073) 0.39 (15130)	0.77 (0.0059) 0.54 (15152)
C	0.69 (0.0058) 0.38 (16675)	0.74 (0.0061) 0.49 (14871)	0.70 (0.0068) 0.40 (15411)	0.69 (0.0074) 0.38 (15130)	0.76 (0.0059) 0.53 (15152)
D	0.72 (0.0056) 0.45 (16401)	0.78 (0.0091) 0.56 (15079)	0.75 (0.0068) 0.51 (15591)	0.75 (0.0063) 0.51 (15622)	0.80 (0.0077) 0.61 (15048)

Notes

1. Each cell gives the AUC (s.e.) followed by the Gini coefficient (sample size).
2. Standard errors based on 100 bootstrap replications.

Table 2: Estimated between cluster variances and covariances (with 95% credible intervals)

		Variance	Covariance
Unproductive		0.142 (0.095, 0.196)	n.a.
Non-response (type 1)	Wave non-response	0.288 (0.190, 0.404)	0.123 (0.071, 0.182); r = 0.76
	Attrition	0.091 (0.049, 0.143)	
Non-response (type 2)	Refusal	0.191 (0.125, 0.274)	0.078 (0.017, 0.145); r = 0.36
	Other non-productive	0.247 (0.166, 0.347)	

Note

Based on model B, Table 1.

Table 3: Associations with variables in model of interest

	Bracken test[3]	Log income, wave one	Log income, wave two
Cluster[1]	0.12	0.20	0.16
Stratum[2]	0.89	0.91	0.80
Consent[4]	-0.026	-0.036	-0.040
Stable address[4]	0.071	0.048	0.046
Refusing income question[4]	-0.015	n.a.	-0.023
No partner response[4]	-0.050	0.001	-0.030
Changed address[4]	0.042	-0.10	-0.071
Neighbourhood assessment[4]	-0.23	-0.40	-0.37

Notes

[1] Intra-cluster correlation

[2] Range (SD units)

[3] Square root transformation of standardised score

[4] Pearson correlations

Table 4: Applying overall weights

		Design weights	Weight from model A	Weight from model B	Weight from model C	Weight from model D
Bracken score (mean (s.e.); n)		0.10 (0.023); 13294	0.080 (0.023); 12199	0.082 (0.023); 12199	0.082 (0.023); 12199	0.075 (0.023); 12759
Sub-model (i)	Log family income, wave 1 (estimate (s.e.); n)	0.21 (0.019); 10708	0.21 (0.019); 10650	0.21 (0.019); 10650	0.21 (0.019); 10650	0.21 (0.020); 10315
	Log family income, wave 2 (estimate (s.e.))	0.19 (0.017)	0.19 (0.017)	0.19 (0.017)	0.19 (0.017)	0.19 (0.017)
Sub-model (ii)	Log income difference	0.045 (0.017)	0.040 (0.015)	0.047 (0.018)	0.047 (0.017)	0.048 (0.017)
Sub-model (iii)	Income difference	0.0025 (0.00077)	0.0037 (0.00074)	0.0028 (0.00079)	0.0028 (0.00079)	0.0027 (0.00079)
Weight range		0.23 – 2.0	0.20 – 3.7	0.20 – 3.4	0.20 – 3.1	0.21 – 3.7

Table 5: Missing value patterns

Log family income, wave 1	Bracken score	Log family income, wave 2	Frequency (%)
1	1	1	10628 (59)
1	1	0	1544 (8.5)
1	0	0	3390 (19)
1	0	1	1077 (5.9)
0	1	1	686 (3.8)
0	1	0	295 (1.6)
0	0	1	98 (0.54)
0	0	0	430 (2.4)

Note
1 = present; 0 = missing.

Table 6: Estimates from MI (1)

		Model A	Model C	Model D
Sub-model (i)	Log family income, wave 1 (estimate (s.e.))	0.21 (0.014)	0.21 (0.014)	0.21 (0.016)
	Log family income, wave 2 (estimate (s.e.))	0.20 (0.014)	0.20 (0.013)	0.20 (0.015)
Sub-model (ii)	Log income difference	0.041 (0.015)	0.040 (0.015)	0.041 (0.014)
Sub-model (iii)	Income difference	0.0035 (0.00076)	0.0035 (0.00076)	0.0035 (0.00074)

Notes

n = 18148; based on 20 imputations; MCMC burn in = 1000; 10 iterations between draws.

Table 7: Estimates from MI (2)

		No imputation (CC)	Model A	Model C	Model D
Sub-model (i)	Log family income, wave 1 (estimate (s.e.))	0.18 (0.015)	0.16 (0.014)	0.16 (0.014)	0.17 (0.015)
	Log family income, wave 2 (estimate (s.e.))	0.16 (0.013)	0.17 (0.014)	0.17 (0.014)	0.17 (0.013)
Sub-model (ii)	Log income difference	0.038 (0.013)	0.039 (0.012)	0.045 (0.014)	0.040 (0.013)
Sub-model (iii)	Income difference	0.0023 (0.0007)	0.0024 (0.0008)	0.0024 (0.0006)	0.0025 (0.0007)

Table 8: Estimates for MoI from joint model

		Model A	Model C	Model D
Sub-model (i)	Log family income, wave 1 (estimate (s.e.))	0.14 (0.016)	0.15 (0.016)	0.17 (0.016)
	Log family income, wave 2 (estimate (s.e.))	0.14 (0.013)	0.14 (0.013)	0.14 (0.014)
Estimated correlation		-0.69 (0.019)	-0.69 (0.019)	-0.71 (0.018)
Sub-model (ii)	Log income difference	0.056 (0.013)	0.055 (0.013)	0.048 (0.013)
Estimated correlation		-0.81 (0.0098)	-0.80 (0.0099)	-0.81 (0.010)
Sub-model (iii)	Income difference	0.0030 (0.00067)	0.0030 (0.00067)	0.0028 (0.00068)
Estimated correlation		-0.80 (0.0099)	-0.80 (0.010)	-0.81 (0.010)
Sample sizes (response; non-resp.)		10650; 4476	10650; 4476	10315; 3253

References

Behr, A., E. Bellgardt, and U. Rendtel. 2005. "Extent and Determinants of Panel Attrition in the European Community Household Panel." *European Sociological Review* 21: 489-512.

Browne, W. J. 2009. *MCMC Estimation in MLwiN, v2.10*. Centre for Multilevel Modelling, University of Bristol.

Browne, W. J., F. Steele, M. Golalizadeh, and M. J. Green. 2009. "The Use of Simple Reparameterizations to Improve the Efficiency of Markov chain Monte Carlo Estimation for Multilevel Models with Applications to Discrete Time Survival Models." *Journal of Royal Statistical Society, Series A* 172: 579-98.

Carpenter, J. and I. Plewis. 2011. *Analysing Longitudinal Studies with Non-response: Issues and Statistical Methods*. Pp 518-43 in *Handbook of Methodological Innovations* (edited by M. Williams and W. P. Vogt), Newbury Park, Ca.: Sage.

Couper, M. P. 1998. "Measuring Survey Quality in a CASIC Environment." *Proceedings Survey Research Methods Section, American Statistical Association*: 41-9.

Durrant, G. B. and F. Steele. 2009. "Multilevel Modelling of Refusal and Non-contact in Household Surveys: Evidence from six UK Government Surveys." *Journal of the Royal Statistical Society, Series A* 172: 361-82.

Fitzgerald, J., P. Gottschalk, and R. Moffitt. 1998. "An Analysis of Sample Attrition in Panel Data." *Journal of Human Resources* 33: 251-99.

Goldstein, H. 2009. “Handling Attrition and Non-response in Longitudinal Data.” *Longitudinal and Life Course Studies* 1: 63-72.

Goldstein, H., F. Steele, J. Rasbash, and C. Charlton. 2008. *REALCOM: Methodology for Realistically Complex Multilevel Modelling.* Centre for Multilevel Modelling, University of Bristol.

Groves, R. M. 2006. “Nonresponse Rates and Non-response Bias in Household Surveys.” *Public Opinion Quarterly* 70: 646-75.

Hawkes, D. and I. Plewis. 2006. “Modelling Non-response in the National Child Development Study.” *Journal of the Royal Statistical Society, Series A* 169: 479–91.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T. M. Ezzati-Rice, C. Casas-Cordero, M. Lemay, A. Peytchev, R. M. Groves, and T. E. Raghunathan. 2010. “Using Proxy Measures and other Correlates of Survey Outcomes to Adjust for Non-response: Examples from Multiple Surveys.” *Journal of the Royal Statistical Society, Series A* 173: 389-408.

Krzanowski, W. J. and D. J. Hand. 2009 *ROC Curves for Continuous Data*. Boca Raton, Fl.: Chapman and Hall/CRC.

Lepkowski, J. M., and M. P. Couper. 2002. *Nonresponse in the second wave of longitudinal household surveys.* In *Survey Nonresponse* (edited by R. M. Groves *et al*.), New York: John Wiley.

Little, R. J. A. and D. B. Rubin. 2002. *Statistical Analysis with Missing Data (2nd. ed.).* New York: John Wiley.

Little, R. and S. Vartivarian. 2005. “Does Weighting for Nonresponse Increase the Variance of Survey Means?” *Survey Methodology* 31: 161-68.

Little, R. J. and N. Zhang. 2011. “Subsample Ignorable Likelihood for Regression Analysis with Missing Data.” *Applied Statistics* 60: 591-606.

Mason, A., S. Richardson, I. Plewis, and N. Best. 2012. “Strategy for Modelling Non-random Missing Data Mechanisms in Observational Studies using Bayesian Methods.” *Journal of Official Statistics* (forthcoming).

Plewis, I. 2001. “Explanatory Models for Relating Growth Processes.” *Multivariate Behavioral Research* 36: 207-26.

Plewis, I. 2007a. “Non-response in a Birth Cohort Study: The case of the Millennium Cohort Study.” *International Journal of Social Research Methodology* 10: 325-34.

Plewis, I., ed. 2007b. *The Millennium Cohort Study: Technical Report on Sampling (4th. ed.)*. London: Institute of Education, University of London.

Plewis, I. and C. Kallis. 2008 *Changing Economic Circumstances in Childhood and their Effects on Subsequent Educational and other Outcomes.* DWP Working Paper No. 49. London: Department for Work and Pensions.

Plewis, I., S. C. Ketende, H. Joshi, and G. Hughes, G. 2008. “The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First two Waves of the Millennium Cohort Study.” *Journal of Official Statistics* 24: 365-85.

Plewis, I., S. C. Ketende, and L. Calderwood. 2012. “Assessing the Accuracy of Response Propensities in Longitudinal Studies.” *Survey Methodology* (forthcoming).

Rasbash, J., C. Charlton, W. J. Browne, M. Healy, and B. Cameron. 2009. *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol.

Reiter, J. P., T. E. Raghunathan, and S. K. Kinney. 2006. "The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data." *Survey Methodology* 32: 143-49.

Sikkel, D., J. Hox, and E. de Leeuw. 2009. *Using Auxiliary Data for Adjustment in Longitudinal Research.* Pp 141-156 in *Methodology of Longitudinal Surveys* (edited by P. Lynn). Chichester: John Wiley.

Sterne, J. A. C., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. Wood, and J. R. Carpenter. 2009. "Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls." *British Medical Journal* 338: b2393.

Vella, F. 1998. "Estimating Models with Sample Selection Bias: A Survey." *Journal of Human Resources* 33: 127-69.

Watson, N. and M. Wooden. 2009. *Identifying Factors Affecting Longitudinal Survey Response.* Pp 157-82 in *Methodology of Longitudinal Surveys* (edited by P. Lynn). Chichester: John Wiley.

Appendix 1

Potential predictors of non-response (group 1)

1. Family income (6 ordered categories: 1.8%; 26%; 33%; 20%; 14%; 5.0%)
2. Ethnic group of cohort child (White British (83%); Mixed (3.0%); Indian (2.5%); Pakistani/Bangladeshi (6.9%); Black/Black British (3.6%); Other (1.4%))
3. Accommodation type (House (85%); other (15%))
4. Tenure (Own (58%); rent (36%); other (6.4%))
5. Main respondent's age (< 30 (50%); 30+ (50%))
6. Main respondent's educational qualifications (None (20%); NVQ1-5 (3.3%; 12%; 8.4%; 9.3%; 44%); other/overseas (2.8%))
7. Cohort child breast fed (Yes: 67%)
8. Longstanding illness, main respondent (Yes: 21%)
9. Parental status (2 (83%) or 1 parent family (17%))
10. Main respondent voted in last general election (Yes: 51%)

Potential predictors of non-response (group 3)

1. Gave consent to record linkage (Yes: 93%)
2. Provided a stable address at wave one (Yes: 82%)
3. Missing family income, wave one (8.7%)
4. No or proxy partner response (11%)

Appendix 2

Interviewer assessments of the neighbourhood, MCS wave 2.

For each visit they made to the household, the wave two interviewers responded to 11 questions about the general state of the neighbourhood and on whether they felt safe or unsafe when they visited the household. This information was gathered for both responding and non-responding households across the UK. Up to 15 visits were made in some cases. In most cases, however, the interviewer gave the same answer regardless of how many times they visited the property and so there was no evidence that interviewers' perceptions changed according to the time of day or day of the week that they were in the area. Consequently, the data used here come from the first visit to each household.

The scoring for the summary score is as follows:

Assessment item	**Category**	**Score**
1. How would you rate the general condition of most of the residences or other buildings in the street?	Well kept, good repair and exterior surfaces Fair condition Poor condition, peeling paint, broken windows Badly deteriorated	0 1 2 2
2. Do any of the fronts of residential or commercial units have metal security blinds, gates or iron bars & grilles?	None Some Most	0 1 2
3. Are there any traffic calming measures in place on the street?	No traffic permitted Light traffic Calming + moderate traffic	0 0 0
4. How would you rate the volume of traffic on the street?	No calming+ moderate Calming + heavy traffic No calming +heavy	1 1 2

5. Are there any burnt-out cars on the street?	No Yes	0 2
6. Is there any of the following: rubbish, litter, broken glass, drug related items, beer cans etc, cigarette ends or discarded packs - in the street or on the pavement?	None or almost none Yes, some Yes, just about everywhere you look	0 1 2
7. Is there any graffiti on walls or on public spaces like bus shelters, telephone boxes or notice boards?	No A little A lot	0 1 2
8. Is there dog mess on the pavement?	None Some A lot	0 1 2
9. Is there any evidence of vandalism such as broken glass from car windows, bus shelters or telephone boxes?	No Yes	0 2
10. Are there any adults or teenagers in the street or on the pavements arguing, fighting, drinking or behaving in any kind of hostile or threatening way?	No-one seen in the street or pavement None observed behaving in hostile ways Yes, one or two arguing etc. Yes, at least one group of three or more	0 0 1 2
11. How did you feel parking/walking /waiting at the door in the street?	Very comfortable, can imagine living/ working/shopping here Comfortable - a safe and friendly place Fairly safe and comfortable I would be uncomfortable living/working/shopping here I felt like an outsider, looked on suspiciously I felt afraid for my personal safety	0 0 1 2 2 2

The summary score can vary from zero to 20 but very few scores over 10 were obtained as shown in Table A3.1.

Table A3.1: Distribution of neighbourhood assessment score (n = 16594)

Score	0	1 - 3	4 - 6	7 – 10	>10
%	34	42	16	6	2