The Cathie Marsh Centre for Census and Survey Research

Assessing the accuracy of response propensities in longitudinal studies

CCSR Working Paper 2010-08

Ian Plewis, Sosthenes Ketende, Lisa Calderwood

Ian Plewis, Social Statistics, University of Manchester, Manchester M13 9PL, U.K. E-mail: ian.plewis@manchester.ac.uk. Sosthenes Ketende and Lisa Calderwood, Centre for Longitudinal Studies, Institute of Education, London WC1H 0AL, U.K.

The omnipresence of non-response in longitudinal studies is addressed by assessing the accuracy of statistical models constructed to predict different types of non-response. Particular attention is paid to summary measures derived from receiver operating characteristic curves and logit rank plots as ways of assessing accuracy. The ideas are applied to data from the first four waves of the UK Millennium Cohort Study and the results suggest that our ability to discriminate and predict non-response is not high. Changes in socio-economic circumstances do predict wave non-response with implications for the underlying missingness mechanism. Conclusions are drawn in terms of the potential of interventions to prevent non-response and methods of adjusting for it..

www.ccsr.ac.uk

Abstract

The omnipresence of non-response in longitudinal studies is addressed by assessing the accuracy of statistical models constructed to predict different types of non-response. Particular attention is paid to summary measures derived from receiver operating characteristic curves and logit rank plots as ways of assessing accuracy. The ideas are applied to data from the first four waves of the UK Millennium Cohort Study and the results suggest that our ability to discriminate and predict non-response is not high. Changes in socio-economic circumstances do predict wave non-response with implications for the underlying missingness mechanism. Conclusions are drawn in terms of the potential of interventions to prevent non-response and methods of adjusting for it.

Key words: Longitudinal studies; missing data; attrition; propensity scores; ROC curves; Millennium Cohort Study.

1. Introduction

Designers and managers of longitudinal studies have to put into operation strategies for preventing sample loss over time. Despite the designers' often heroic efforts, however, analysts of longitudinal data must deal with the problem of missingness. Ideally, they do this by generating information about why data are missing and then combining this information with statistical techniques that adjust for the missingness. The focus of this paper is on how we can learn more about missingness by assessing the accuracy of models that predict the different kinds of, and different reasons for non-response that affect longitudinal studies. Knowledge from these models – and from estimates of their accuracy - can then be exploited in three ways. First, in the construction and evaluation of weighting schemes designed to remove biases from estimates for variables of interest, variables that are often associated with the systematic non-response usually found in these studies. Second, they can be used to generate imputations both to remove bias and also to improve the precision of estimates of interest. Third, the models can be used to predict who might be responders and non-responders at future waves of a study and thus to consider targeting or tailoring fieldwork resources to those respondents who might otherwise be lost from the study.

This paper is built around a framework for assessing the accuracy of models that account for variability in non-response outcomes, i.e. non-response propensity models. This framework is widely used in epidemiology and criminology to generate risk scores but has not, to our knowledge, been used in survey research before. We apply it to address the following three questions:

1) How is the accuracy of non-response propensity models best assessed?

2) Can the accuracy of non-response propensity models at a particular wave be enhanced by using variables measured at later waves?

3) How accurate are the propensity models at an early wave if they are applied to non-response at later waves?

There are many instances in the literature of studies that have modelled the predictors of non-response in longitudinal surveys, stimulated by the fact that these models can draw on measures obtained from sample members before (and, as we shall see, after) the occasions at which they are non-respondents. See, from many possible examples, Lepkowski and Couper (2002) for an analysis that separates refusals from not being located or contacted; Hawkes and Plewis (2006) who analyse data from the UK National Child Development Study and who separate wave non-respondents from attrition cases; and, of particular relevance here, Plewis (2007a) and Plewis et al. (2008) who consider non-response in the first two waves of the UK Millennium Cohort Study, described in more detail below. The accuracy of models of this kind for prediction has not, however, been given the amount of attention it warrants in terms of their

ability to discriminate between respondents and non-respondents, and to predict future non-response.

The paper is organised as follows. The framework for assessing accuracy is set out in the next section. Section 3 introduces the UK Millennium Cohort Study and propensity score methods are illustrated using data from this study in Section 4. Implications of the findings for preventing non-response and for statistical adjustment for missingness are then considered; Section 6 concludes.

2. Models for predicting non-response

It is relatively straightforward to specify and estimate models for explaining both overall non-response and also different kinds of non-response, i.e. wave non-response and attrition; and failure to locate, to contact (conditional on location) and to cooperate (conditional on contact). A typical model for a binary outcome is the one proposed by Hawkes and Plewis (2006):

$$f(\pi_{it}) = \sum_p \beta_p x_{pi} + \sum_q \sum_k \gamma_q x^*_{qi,t-k} + \sum_r \sum_k \delta_r z_{ri,t-k} \quad (1)$$

where:

$\pi_{it} = E(r_{it})$ is the probability of not responding for subject *i* at wave *t* with r_{it} = 0 for a response and 1 for non-response, and *f* is an appropriate function such as logit or probit.

i = 1..n where *n* is the observed sample size at wave one.

t = 1..T_i where T_i is the number of waves for which r_{it} is recorded for subject *i*.

x_{pi} are fixed characteristics of subject *i* measured at wave one, *p = 0..P*; x_0 = *1* for all *i*.

$x^*_{qi,t-k}$ are time-varying characteristics of subject *i*, measured at waves *t-k, q = 1..Q, k = 1,2…* Often *k* will be 1.

$z_{ri,t-k}$ are time-varying characteristics of the data collection process, measured for subject *i* at waves *t-k, r = 1..R, k = 0,1...* Often *k* will be 1 but can be 0 for variables such as number of contacts before a response is obtained.

This model can easily be extended to more than two response categories such as {response, wave non-response, attrition}. Other approaches are also possible. For example, it is often more convenient to model the probability of not responding just at wave $t = t^*$ in terms of variables measured at earlier waves $t^* - k$, $k \geq 1$ or, when non-response is monotonic (implying there is no wave non-response), to model time to attrition as a survival process.

The estimated probabilities (p_{it}) from equation (1) can be used to generate inverse probability weights w_{it} (=$1/p_{it}$) and these are widely applied to try to adjust for biases arising from non-response under the assumption that data are missing at random (MAR) as defined by Little and Rubin (2002).

2.1 Assessing the accuracy of predictions

Regardless of the method that is used to construct a function estimated from a generalised linear model like (1) that links the response categories to the explanatory variables, a key question remains: how accurate is the model? We can think of these functions as risk scores (Copas, 1999) or propensity scores (Little and Rubin, 2002) and we can then ask about the accuracy of these scores. A widely used method of assessing accuracy is to estimate the goodness-of-fit of models for binary or categorical outcomes by using one of several possible pseudo-R^2 statistics. Apart from their rather arbitrary nature, which thus makes comparisons across datasets difficult, pseudo-R^2 are not especially useful in this context because they assess the overall fit of the model and do not distinguish between the accuracy of the model for the respondents and non-respondents separately.

As the epidemiological literature emphasises (e.g. Pepe, 2003), there are two related components of accuracy: classification (or discrimination) and prediction. Classification refers to the conditional probabilities of having a propensity score (s) above a chosen threshold (h) given that a person either is or is not a non-respondent. Prediction, on the other hand, refers to the conditional probabilities of being or becoming a non-respondent given a propensity score above or below the threshold.

More formally, let D and $\bar{D}$ refer to the presence and absence of the poor outcome (i.e. non-response) and define + ($s > h$) and – ($s \leq h$) as positive and negative tests derived from the propensity score and its threshold. Then, for classification, we are interested in P(+|D), the true positive fraction (TPF) or sensitivity of the test, and P(-|$\bar{D}$), its specificity, equal to one minus the false positive fraction (1 – FPF). For prediction, however, we are interested in P(D|+), the positive predictive value (PPV) and P($\bar{D}$|-), the negative predictive value (NPV). If the probability of a positive test (P(+) = τ) is the same as the prevalence of the poor outcome (P(D) = ρ) then inferences about classification and prediction are essentially the same. With $\tau = \rho$, sensitivity equals PPV and specificity equals NPV. Generally, however, {TPF, FPF, ρ} and {PPV, NPV, τ} convey different pieces of information.

The TPF (i.e. sensitivity) can be plotted against FPF (i.e. 1 - specificity) for any risk score threshold h. This is the receiver operating characteristic (ROC) curve (Figure 1). The ROC curve is always anchored at coordinates (0, 0) and (1,1) and, for large samples and at least some continuously measured predictors, it is smooth with a monotonically declining but always positive slope. Krzanowski and Hand (2009) give a detailed discussion of how to estimate ROC curves. The diagonal line joining the point [0, 0] (sensitivity = 0, specificity = 1: everyone is predicted to be a respondent so the threshold on the probability scale is one) to [1, 1] (sensitivity = 1, specificity = 0: everyone is predicted to be a non-respondent so the threshold is zero) is the ROC that would be obtained if the

variables used to construct the risk score do not explain any of the variation in the outcome. Consequently, it is the AUC – the area enclosed by the ROC curve and the x-axis in Fig. 1 – that is of interest and this can vary from 1 (perfect discrimination) down to 0.5, the area below the diagonal (no discrimination). 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) - sometimes referred to as a Gini coefficient and equivalent to Somer's D rank correlation index (Harrell, 1996) - is commonly used as a more natural measure than AUC because it varies from 0 to 1.

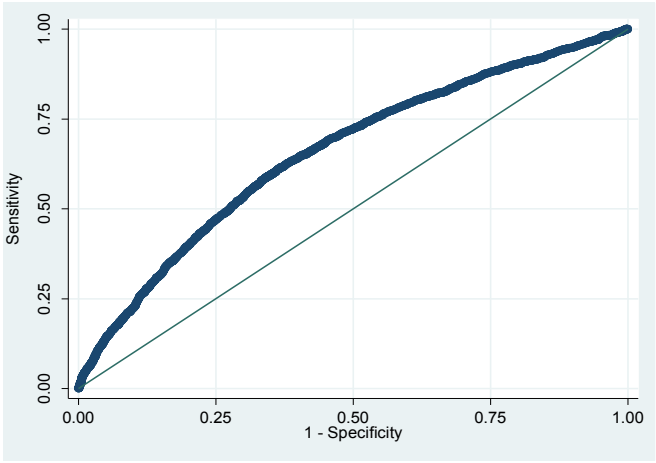

Figure 1: ROC curve

Copas (1999) proposes using the logit rank plot as an alternative to the ROC as a means of assessing the predictiveness of a risk or propensity score. If the propensity score is derived from a logistic regression then a logit rank plot is just a plot of the linear predictor from the logistic regression model against the logistic transformation of the proportional rank of the propensity scores as shown in Figure 2. More generally, it is a plot of logit(p_i) where p_i is the estimated probability from any form of (1) i.e. $\hat{R}(D \mid \mathbf{x}, \mathbf{x}^*, \mathbf{z})$, against the logits of the proportional ranks (r/n) where r is the rank position of case i (i = 1..n) on the propensity score. This relation is usually close to being linear and its slope – which can vary from zero to one - is a measure of the predictive strength of the propensity score. Copas argues that the slope is more sensitive to changes in the specification of the model underpinning the propensity score and to changes in the prevalence of the outcome than the Gini coefficient is. The slope is scale-independent and can therefore be used to compare the predictive strength of different propensity scores for the outcome of interest. A good estimate of the slope can be obtained by calculating quantiles of the variables on the y and x axes and then fitting a simple regression model.

One issue to bear in mind when assessing Gini coefficients and logit rank plot slopes is that they are both subject to shrinkage: they will be lower when applied to a new set of cases. The degree of shrinkage is directly proportional to the number of explanatory variables in the model but inversely proportional to the sample size (Copas, 1999; Copas and Corbett, 2002). As the sample size of the dataset we use here to illustrate the methods is very large, shrinkage will be minimal.

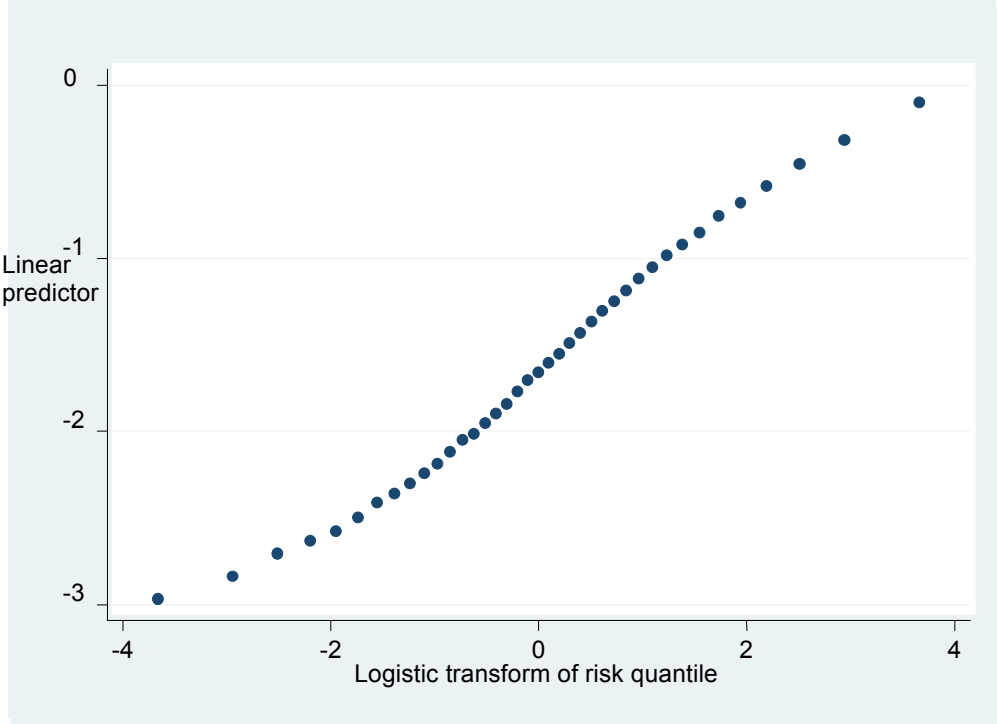

Figure 2: Logit rank plot: illustration

It is also possible to estimate, from (1), d_1 at wave t^*: the difference between the means of the estimated linear predictors, and d_2 - the difference between the means of the predicted probabilities (p_i) of not responding - for non-respondents and respondents:

$$d_1 = [1/NR \sum_{i=1}^{NR} X_i b + X_i^* c + Z_i d)]_{r_i=1} - [1/R \sum_{j=1}^{R} X_j b + X_j^* c + Z_j d)]_{r_i=0} \quad (2a)$$

and

$$d_2 = [1/NR \sum_{i=1}^{NR} p_i]_{r_i=1} - [1/R \sum_{j=1}^{R} p_j]_{r_i=0} \quad (2b)$$

where *b*, *c* and *d* are vectors of estimated coefficients (i.e. *β, γ* and *δ* from (1)),

$p_i = f^{-1}(X_i b + X_i^* c + Z_i d)$ and $NR = \sum_{i=1}^{n} r_i \; ; R = n - NR$

We would expect d_k (k = 1,2) to be positive but it is not clear how much greater than zero d_k should be for a model to be useful, nor how values of d_k should be compared across studies. A problem with d_1 is that it is not bounded; also, it will vary according to the chosen function *f*. The same arguments would hold were d_k to be defined in terms of ratios rather than differences.

The extent to which propensity scores discriminate between respondents and non-respondents can be used as an indication of how influential, and possibly how effective our statistical adjustments are going to be. A lack of discrimination suggests either that there are important predictors missing from the propensity score or that a substantial part of the process that drives the missingness is essentially random. The extent to which propensity scores predict whether a case will be a non-respondent in subsequent waves – and what kind of non-respondent they will be - is an indication of whether any intervention to reduce non-response, how ever well designed and targeted, will be successful.

3. 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 as described in Plewis (2007b). 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.

3.1 Sample loss from the Millennium Cohort Study

Table 1 shows how the MCS sample has diminished over time after wave one. The sample loss consists of wave non-respondents – cases that are missing at wave t but not at one or more subsequent waves – and attrition cases that, once missing, have not reappeared. It is not possible definitively to allocate all cases to one of these two non-response categories until the end of the study but Table 1 does indicate that sample loss from the MCS, in common with most longitudinal studies in the social sciences, consists of a mixture of wave non-response and attrition and is therefore not monotone. Table 1 also shows that non-respondents are equally divided between refusals and other non-productives (not located, not

contacted etc.) at wave two but that refusal becomes more dominant thereafter. There is an association between type of, and reasons for non-response: 62% of the attrition cases but only about one third of the wave non-respondents are refusals. The proportion of not located cases within the heterogeneous 'other non-productive' group rose from about 40% for waves two and three to 61% in wave four. Note that the eligible sample size – which excludes child deaths and emigrants - increased between waves two and three because some cases omitted at wave one in England were recruited for the first time at wave two.

Table 1: Sample loss from MCS after wave one by non-response type

	Wave 2, age 3 yrs	Wave 3, age 5 yrs	Wave 4, age 7 yrs
(i) Wave non-response	9.0%	3.4%	n.a.
(ii) Attrition	10%	16%	n.a.
Total (= i + ii = iii+ iv)	19%	20%	26%
(iii) Refusal	9.5%	12%	19%
(iv) Other non-productive	9.5%	7.3%	7.4%
Eligible sample size	18,385	18,944	18,756

n.a.: not applicable as wave non-response is undefined at the most recent wave.

4. Analyses of non-response

In this section, we use MCS data to answer the three broad questions about accuracy posed in the Introduction.

4.1 Accuracy of classification and prediction

Research reported in Plewis (2007b) and Plewis et al. (2008) on the predictors of different types of non-response in the MCS is summarised in Table 2. We see that variables measured at wave one that are associated with attrition are not necessarily associated with wave non-response (and vice-versa). The same is true for correlates of refusal and other non-productives. The estimate of 0.39 for the Gini coefficient for overall non-response is relatively low: it corresponds to an AUC of 0.69 which is the probability of correctly assigning (based on their predicted probabilities) a pair of cases (one respondent, one non-respondent), indicating that classification of (or discrimination between) non-respondents and respondents from the propensity score is not especially good. Classification is slightly better for wave non-respondents than it is for attrition and notably better for other non-productive than it is for refusal (although this is partly because the moving residence variable and not being located are so closely related). These estimates were obtained from pair-wise comparisons of each non-response category with being a respondent. A similar picture emerges when we look at the slopes of the logit rank plots although these bring out more clearly the differences in predictiveness for the different types of non-response. The pattern of estimates

for d_1 is in line with those from the ROC and logit rank plots but this not true for d_2 (for example, the estimate for attrition is higher than it is for wave non-response), suggesting that comparing mean estimated probabilities from models like (1) might be misleading. We do not consider d_1 and d_2 further.

The correct specification of models for explaining non-response can be difficult to achieve. New candidates for inclusion in a model can appear after the model and the corresponding inverse probability weights have been estimated and disseminated, others remain unknown. How much effect on measures of accuracy might these new variables have? Here, we examine the effects of adding three new variables to the MCS models: (i) whether or not respondents gave consent to having their survey records linked to health records at wave one; (ii) a neighbourhood conditions score varying from zero to 20 and derived from interviewer observations at wave two (see Appendix N of Edwards et al. (2006) for more details); and (iii) whether, at wave one, the main respondent reported voting at the last UK general election. The first two of these variables were not available for the analyses summarised in Table 2 but there are grounds for supposing that they might be important: refusing consent at wave t might be followed by overall refusal at wave t + 1, and non-response might be greater in poorer neighbourhoods. The voting variable was suggested as an indicator of social engagement that might also be related to the probability of responding. Although the neighbourhood conditions score was obtained for productive cases, refusals and non-contacts it could not, inevitably, be obtained for cases that were not located. Hence, we use this variable just for the model that compares refusals with productives.

Table 2: Explanatory variables for non-response, MCS wave two

Wave one explanatory variable [1]	Overall non-response [3]	Wave non-response [4]	Attrition [4]	Refusal [5]	Other non-productive [5]
Moved residence after wave one	✓	✓	×	×	✓
UK country	✓	✓	✓	✓	✓
Family income	✓	×	✓	✓	×
Refused to answer income qn.	✓	×	×	✓	×
Ethnic group	✓	✓	✓	×	✓
Tenure	✓	✓	✓	×	✓
Accommodation type	✓	✓	✓	✓	✓
Mother's age	✓	✓	✓	✓	✓
Education	✓	✓	✓	✓	✓
Provided stable address	✓	×	✓	✓	✓
Cohort member breast fed	✓	✓	✓	✓	✓
Longstanding illness	✓	✓	✓	✓	✓
Partner present	✓	✓	✓	✓	✓
Partner but no interview	✓	✓	✓	✓	✓
Sample size	18,230	16,210	16,821	16,543	16,513
AUC [2]	0.69	0.71	0.69	0.68	0.76
Gini [2]	0.39	0.42	0.38	0.37	0.52
Logit rank plot: slope [2]	0.45	0.51	0.44	0.40	0.63
d_1	0.53	0.65	0.51	0.46	1.0
d_2	0.078	0.054	0.057	0.049	0.097

Notes

[1] See Plewis (2007) and Plewis et al. (2008) for more details: ✓, related to non-response category; ×, not related to non-response category.

[2] 95% confidence limits for AUC, Gini coefficient and logit rank plot slope generally ± 0.02.

[3] Based on a logistic regression, allowing for the survey design using the *svy* commands in STATA.

[4] Based on a multinomial regression with three categories, allowing for the survey design and with the sample size based on the sum of the productive and relevant non-response category.

[5] Based on a multinomial regression with three categories, allowing for the survey design and with the sample size based on the sum of the productive and relevant non-response category.

Table 3 presents the results. We see that each of the three variables is associated with at least one kind of non-response. The accuracy of the propensity scores is, however, little changed by the inclusion of extra variables except for refusal where the inclusion of the three new variables does make a

difference; the estimate of the Gini coefficient increases from 0.37 to 0.42 and the slope of the logit rank plot increases from 0.40 to 0.45 (although missing data for the neighbourhood conditions score does reduce the sample size).

Table 3: Accuracy estimates for enhanced models, MCS wave two

Accuracy measure	Overall non-response [1]	Wave non-response [2]	Attrition [3]	Refusal [4]	Other non-productive [5]
AUC	0.70	0.71	0.71	0.71	0.76
Gini	0.40	0.43	0.41	0.42	0.53
Logit rank plot: slope	0.47	0.52	0.47	0.45	0.65
Sample size	18,148	16,177	16,745	15,656	16,443

Notes
[1] Includes consent (odds ratio (OR) = 2.1, s.e. = 0.20) and vote (OR = 1.4, s.e. = 0.08) variables.
[2] Includes vote variable only (OR = 1.4, s.e. = 0.11), consent not important ($t = 1.33$; $p > 0.18$).
[3] Includes consent (OR = 2.7, s.e. = 0.26) and vote (OR = 1.4, s.e. = 0.09).
[4] Includes consent (odds ratio (OR) = 2.6, s.e. = 0.32), vote (OR = 1.3, s.e. = 0.10) and neighbourhood score (OR = 1.03, s.e. = 0.014) variables.
[5] Includes consent (odds ratio (OR) = 1.6, s.e. = 0.20) and vote (OR = 1.5, s.e. = 0.11) variables.

4.2 Classifying wave non-respondents using data from subsequent waves

It is possible to use variables measured at waves $t + k$, and changes between waves $t – k$ and $t + k$ ($k \geq 1$) in models for discriminating between wave non-response and being productive at wave t. If such variables do discriminate – and are ignored – then it is more likely that missingness for wave non-response will not be at random (i.e. MNAR) in the sense that changes after wave $t – 1$ predict missingness at wave t. It is not, of course, possible to establish this for the attrition cases. If, however, changes from say $t – 1$ to $t + 1$ are incorporated into an imputation model then this would strengthen the MAR assumption usually required for multiple imputation.

We find that change in accommodation type and in partnership status between waves one and three are both associated with wave non-response at wave two with wave non-response greater when a partner is lost (OR = 1.7, s.e. = 0.18), and less when a partner is acquired (OR = 0.72, s.e. = 0.11), and also less with any change in type of accommodation (OR = 0.57, s.e. = 0.09). Classification is slightly improved with the Gini coefficient rising from 0.43 to 0.46. The slope of the logit rank plot goes up from 0.52 to 0.55.

4.3 Alternative strategies for classifying and predicting non-response

One of the difficulties faced by analysts wishing to use inverse probability weights to adjust for non-response is that, ideally, the weights need to be re-estimated at each wave. The search for new predictors of non-response at each wave, re-estimating the model and then recalculating the inverse probability weights before archiving them for secondary analysts can, however, be time-consuming. Consequently, it is worth investigating whether the changes in the non-response weights after wave two are likely to be sufficiently large to justify recalculation at each subsequent wave. Here we focus on wave four of MCS and consider to what extent discrimination between response categories changes as model specification varies by comparing estimates of Gini coefficients and slopes of logit rank plots for four models for overall non-response based on:

1. The wave one variables used in the response propensity model at wave two, the wave one values of these variables and the wave one coefficients, i.e. applying exactly the same wave two model to wave four outcomes, without any re-estimation.
2. The wave one variables, the wave one values, but the wave three coefficients, i.e. re-estimating the wave two model using the same data as at wave two but with the wave four response categories as the outcome.
3. The wave one variables, but the wave three values of those variables and the wave three coefficients, i.e. using the same explanatory variables for non-response as were used at wave two.
4. Variables measured up to and including at wave three, i.e. starting afresh.

Clearly the amount of analytic work required increases from model 1 to model 4. We find that only four of the 15 wave one variables that were associated with non-response at wave two are not associated with wave four non-response. The estimated Gini coefficients and logit rank plot slopes at wave four for the first three models are (0.36; 0.47: n = 17819), (0.37; 0.44: n = 17819) and (0.37; 0.41: n = 14257), not substantially smaller than the estimates (0.40: 0.47) at wave two for the model that includes the voting and consent variables (Table 3). Model four includes one new variable – the cohort child's score on a cognitive test with lower scores associated with more non-response – along with 10 of the variables used at wave two. The relevant estimates are (0.37; 0.42: n = 13790) indicating that essentially no discrimination is gained by using this wave three variable. Note also that item non-response leads to a substantial decrease in sample size for the third and fourth strategies and this is a general difficulty when constructing non-response weights.

It is not entirely surprising that discrimination is so little affected by changes in model specification because about half of the non-respondents at wave four had also been non-respondents at wave two. In addition, 8.8% of the respondents at wave four had been non-respondents at wave two. For prediction (and therefore possibly intervention) we are more interested in the cases that become non-

responders after wave two, particularly those who appear to be permanently lost from the study. We find that 6% of the eligible productive sample at wave two are non-respondents at both waves three and four. The predictors of these 'attrition' cases are generally similar to those given in Table 2 together with the consent and vote variables described in Section 4.1. In addition, child cognitive test scores at wave two are associated with this attrition which declines as test scores increase. The slope of the logit rank plot (0.48; 95% CI: 0.47 to 0.49) is very close to the value presented in Table 3 (i.e. 0.47) for the cases missing at waves two, three and four.

5. Discussion

Survey methodologists, particularly those working with longitudinal data, have long been exercised by the problem of non-response. Nearly all longitudinal studies suffer from accumulating non-response over time and it is common even for well-conducted mature studies to obtain data for less than half the wave one target sample. On the other hand, non-response researchers have learned a lot about the correlates of different types of non-response by exploiting available data – sometimes known as auxiliary variables - from earlier waves. The main purpose of this paper has been to introduce a different way of thinking about the utility of the approaches that generally rely on binary and multinomial logistic regressions to construct inverse probability weights and to inform imputations. We suggest treating the linear predictors from the logistic regressions as response propensity scores and, by so doing, enabling the methods for summarising the information in these scores to be used to assess the possibilities both for classifying and for predicting different kinds of non-response.

ROCs and their associated summary statistics offer a valuable way of assessing the ability of a model to discriminate between respondents and non-respondents and hence to compare models both within and across studies. The application of this approach to data from the Millennium Cohort Study has shown that, despite using a wide range of explanatory variables, discrimination is on the low side. There are two, not necessarily exclusive, implications of this finding. One is that some non-response is generated by a myriad of circumstantial factors, none of them important on their own, which we can reasonably regard as chance. There is some support for this hypothesis in that the accuracy of the models for overall non-response, wave non-response and other non-productive (the latter two being related) were little changed by the introduction of the voting and consent variables. On the other hand, these variables (and the neighbourhood conditions score) did improve the discrimination between productives, and attrition cases and refusals (which are also related). Nevertheless, discrimination for these two categories remained lower than for the other types of non-response.

A second possible implication is that the models do not discriminate well because data are missing not at random (MNAR) in Little and Rubin's (2002) sense. In other words, it is changes in circumstances after the previous wave that

influences non-response at the current wave. These changes are generally unobserved – except for the wave non-respondents. The finding of an increase in discrimination when changes in partnership status and accommodation type were added to the model for wave non-response at wave two provides some support for the hypothesis that missingness might be informative and that the assumption of MAR might not hold.

Another conclusion to emerge from the application of the methodology, albeit just from one study, is that the functions used to generate inverse probability weights at wave two might reasonably be used at subsequent waves without sacrificing much in the way of discrimination. This would also have the advantage of avoiding the problem of missing weights that arises when variables from later waves (when more data are missing) are included in the non-response models. It would be useful to determine whether this finding is replicated on other studies, particularly panel studies where annual revisions of non-response weights can be expensive.

Turning to prediction, we argue that measures based on simple comparisons across categories of the estimated mean linear predictor and the estimated probabilities are not especially helpful, mainly because they are not scale-invariant. This problem does not affect the slope of the logit rank plot which also has the advantage of being bounded by zero and one. The implications of our findings for prediction are that it might not be generally possible to predict which cases will become non-respondents with a high degree of accuracy. In turn, this suggests that effective interventions might be difficult to target efficiently.

If interventions to prevent non-response in longitudinal studies are to be effective then they need to be based on sound theoretical and empirical foundations, the latter ideally coming from randomised experiments, and they ought to be costed. But interventions also need to be targeted at those cases most likely not to respond. This is where the ROC approach can be especially useful because, as Swets et al. (2000) show, it is possible to determine the optimum threshold for the response propensity score based on the costs and benefits of intervening according to the true and false positive rates implied by the threshold. A more detailed assessment of these issues is beyond the scope of this paper but would include considering interventions to prevent different kinds of non-response and the benefits of potential reductions in both bias and variability arising from a larger sample. It could also draw on the work of Pepe et al. (2008) who argue that a plot of the probability (or risk) of being a non-respondent against the quantiles of the risk score – the predictiveness curve – is a useful way of describing risk and of comparing propensity models.

6. Conclusions

Three main points emerge from this paper. The first is that using a framework that is constructed around different kinds of conditional probabilities and response

propensity scores generates summary measures of accuracy like Gini coefficients and slopes of logit rank plots that enable us to make comparisons across models that predict non-response. These comparisons provide a means of assessing the usefulness of introducing extra predictors into the models and of comparing predictiveness and discrimination for different kinds of non-response.

The second point is that models developed to generate non-response weights at wave two of a study might be satisfactory to use at later waves. If this point were supported by further investigations then this would suggest that efforts to estimate fresh non-response models at each wave might be misplaced.

The final point relates to the implications of our results for statistical adjustment other than by using inverse probability weights. We find that our models of missingness are, for wave non-respondents, improved by including variables from a later wave and therefore these variables could be included in a selected imputation process for a particular model of interest. Moreover, the fact that some of the important explanatory variables for non-response are variables that are unlikely ever to feature in a model of interest – providing a stable address, for example - means that they might be used as identifying instruments in a joint Heckman-type model that considers the models of interest and missingness simultaneously and thus allows for non-ignorable missingness. Carpenter and Plewis (2011) provide an example.

Acknowledgements

This research was funded by the U.K. Economic and Social Research Council under its Survey Design and Measurement Initiative (ref. RES-175-25-0010).

References

Carpenter, J. and Plewis, I. (2011). Analysing longitudinal studies with non-response: issues and statistical methods. In *Handbook of methodological innovations* (Eds., Williams and Vogt). Newbury Park, Ca.: Sage.

Copas, J. (1999). The effectiveness of risk scores: The logit rank plot. *Applied Statistics*, 48, 165-183.

Copas, J. B. and Corbett, P. (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika,* 89, 315-331.

Edwards, A., Barnes, M., Plewis, I. and Morris, K. et al. (2006). *Working to Prevent the Social Exclusion of Children and Young People*. London: Department for Education and Skills Research Report No. 734.

Harrell, F. E. Jr., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361-387.

Hawkes, D., and Plewis, I. (2006). Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society A*, 169, 479–491.

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

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

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

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: OUP.

Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M. and Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*, 167, 362-368.

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

Plewis, I. (2007b). Non-response in a birth cohort study: The case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, 10, 325-334.

Plewis, I., Ketende, S. C., Joshi, H. and 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, 365-385.

Swets, J. A., Dawes, R. M. and Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Sciences in the Public Interest*, 1, 1-26.