Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Assessing the accuracy of response propensity models in longitudinal studies

by Ian Plewis, Sosthenes Ketende and Lisa Calderwood

December 2012

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

This product is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$10.00	CAN$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2012
Vol. 38, No. 2, pp. 167-171
Statistics Canada, Catalogue No. 12-001-X

Assessing the accuracy of response propensity models in longitudinal studies

Ian Plewis, Sosthenes Ketende and Lisa Calderwood [1]

Abstract

Non-response in longitudinal studies is addressed by assessing the accuracy of response propensity models constructed to discriminate between and predict different types of non-response. Particular attention is paid to summary measures derived from receiver operating characteristic (ROC) curves and logit rank plots. The ideas are applied to data from the UK Millennium Cohort Study. The results suggest that the ability to discriminate between and predict non-respondents is not high. Weights generated from the response propensity models lead to only small adjustments in employment transitions. Conclusions are drawn in terms of the potential of interventions to prevent non-response.

Key Words: Longitudinal studies; Missing data; Weighting; Propensity scores; ROC curves; Millennium Cohort Study.

1. Introduction

Examples of studies that have modelled the predictors of different kinds of, and different reasons for the non-response that affect longitudinal studies are plentiful, stimulated by being able to draw on auxiliary variables obtained from sample members before (and after) the occasions at which they are non-respondents. See, for example, Lepkowski and Couper (2002) for an analysis that separates refusals from not being located or contacted; Hawkes and Plewis (2006) who separate wave non-respondents from attrition cases in the UK National Child Development Study; and Plewis (2007a) and Plewis, Ketende, Joshi and Hughes (2008) who consider non-response in the first two waves of the UK Millennium Cohort Study. The focus of this paper is on how we can assess the accuracy of these response propensity models (Little and Rubin 2002). The paper is built around a framework that is widely used in epidemiology (Pepe 2003) and criminology (Copas 1999) to evaluate risk scores but has not, to our knowledge, been used in survey research before. Response propensity models can be used to construct weights intended to remove biases from estimates, to inform imputations, and to predict potential non-respondents at future waves thereby directing fieldwork resources to those respondents who might otherwise be lost. The accuracy of response propensity models 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. Good estimates of accuracy can be used to compare the efficacy of different weighting methods, and to help to determine the allocation of scarce fieldwork resources in order to reduce 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 the methods are illustrated using data from this study in Section 4. Section 5 concludes.

2. Models for predicting non-response

A typical response propensity model for a binary outcome (*e.g.*, Hawkes and Plewis 2006) is:

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

where

- $\pi_{it} = E(r_{it})$ is the probability of not responding for subject i at wave t; $r_{it} = 0$ for a response and 1 for non-response; 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.

1. 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.

Model (1) 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 there is no wave non-response so that non-response has a monotonic rather than an arbitrary pattern, to model time to attrition as a survival process.

The estimated response probabilities p_i, for $t = t^*$, are derived from the estimated non-response probabilities in (1) and they can be used to generate inverse probability weights $g_i (= 1/p_i)$. These are widely applied (see Section 4.2 for an example) 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

A widely used method of assessing the accuracy of models like (1) is to estimate their goodness-of-fit by using one of several possible pseudo-R^2 statistics. Estimates of pseudo-R^2 are not especially useful in this context, partly because they are difficult to compare across datasets but also because they assess the overall fit of the model and do not, therefore, distinguish between the accuracy of the model for the respondents and non-respondents separately.

As Pepe (2003) emphasises, there are two related components of accuracy: discrimination (or classification) and prediction. Discrimination refers to the conditional probabilities of having a propensity score (s: the linear predictor from (1)) above a chosen threshold (c) 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 $+ (s > c)$ and $- (s \leq c)$ as positive and negative tests derived from the propensity score and its threshold. Then, for discrimination, 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 $(P(+) = \tau)$ is the same as the prevalence of the poor outcome $(P(D) = \rho)$ then inferences about discrimination and prediction are essentially the same: sensitivity equals PPV and specificity equals NPV. Generally, however, {TPF, FPF, ρ} and {PPV, NPV, τ} convey different pieces of information. TPF can be plotted against FPF for any risk score threshold c. This is the receiver operating characteristic (ROC) curve (Figure 1). Krzanowski and Hand (2009) give a detailed discussion of how to estimate ROC curves. The AUC – the area enclosed by the ROC curve and the x-axis in Figure 1 – is of particular interest and can vary from 1 (perfect discrimination) down to 0.5, the area below the diagonal (implying 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, Lee and Mark 1996) – is commonly used as a more natural measure than AUC because it varies from 0 to 1.

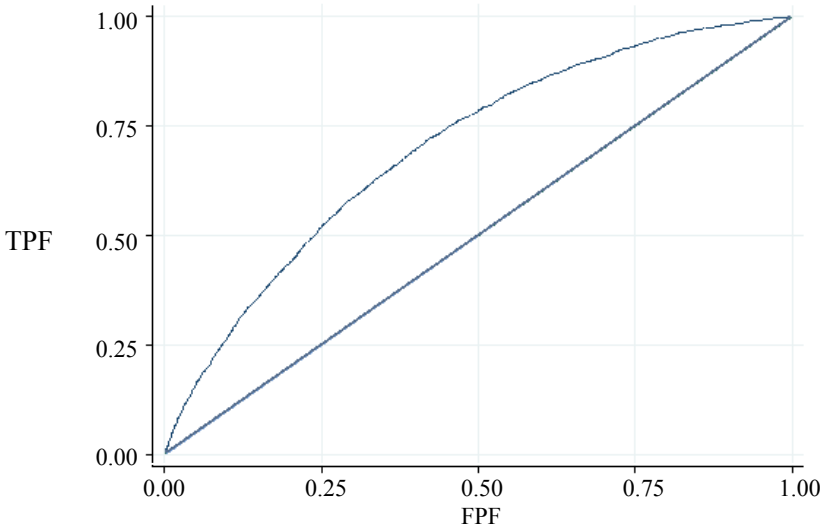

Figure 1 ROC curve

Copas (1999) proposes the logit rank plot as an alternative to the ROC as a means of assessing the predictiveness of a 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 model against the logistic transformation of the proportional rank of the propensity scores. More generally, it is a plot of $\text{logit}(p_i)$ where p_i is the estimated probability from any form of (1) *i.e.*, $p(D|\mathbf{x}, \mathbf{x}^*, \mathbf{z})$, 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 propensity model, and to changes in the prevalence of the outcome, than the Gini coefficient is. 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.

The extent to which propensity scores discriminate between respondents and non-respondents is one indicator of the effectiveness of any statistical adjustments for missingness. A lack of discrimination suggests either that there are important predictors absent 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 will be successful.

3. The Millennium Cohort Study

The wave one sample of the UK Millennium Cohort Study (MCS) includes 18,552 families born over a 12-month period during the years 2000 and 2001, and living in selected UK electoral wards at age nine months. 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. At wave two, 19% of the target sample – which excludes child deaths and emigrants – were unproductive. The unproductive cases were equally divided between wave non-response and attrition, and between refusals and other non-productives (not located, not contacted *etc*.).

4. Analyses of non-response

4.1 Accuracy of discrimination and prediction

Plewis (2007a) and Plewis *et al.* (2008) show that variables measured at wave one of the MCS that are associated with attrition at wave two are not necessarily associated with wave non-response then (and vice-versa). The same is true for correlates of refusal and other non-productives. Table 1 gives the accuracy estimates from the response propensity models. The estimate of the Gini coefficient for overall non-response (0.38) 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 discrimination between non-respondents and respondents from the propensity score is not especially good. Discrimination is slightly better for wave non-respondents than it is for attrition and notably better for other non-productive than it is for refusal. These estimates were obtained from pairwise 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, and reasons for non-response.

Table 1
Accuracy estimates from response propensity models, MCS wave two

Accuracy measure	Overall non-response [2]	Non-response type [2]		Non-response reason [2]	
		Wave non-response	Attrition	Refusal	Other non-productive
AUC [1]	0.69	0.71	0.69	0.68	0.77
Gini [1]	0.38	0.42	0.39	0.37	0.53
Logit rank plot: slope [1]	0.45	0.51	0.44	0.40	0.63
Sample size	18,230	16,210	16,821	16,543	16,513

[1] AUC estimated under the binormal assumption (Krzanowski and Hand 2009); 95% confidence limits for (a) AUC not more than ± 0.015, (b) Gini coefficient and logit rank plot slope not more than ± 0.03.
[2] Based on a logistic regression, allowing for the survey design using the *svy* commands in STATA with the sample size based on the sum of the productive and relevant non-response category.

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, others remain unknown. How much effect on measures of accuracy might the inclusion of 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 derived from interviewer observations at wave two; 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 1: 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 is an indicator of social engagement that might be related to the probability of responding. As the neighbourhood conditions score could not be obtained for cases that were not located, we use this variable just in the model that compares refusals with productives.

Table 2 presents the results using the same methods of estimation as for Table 1 with corresponding levels of precision. We see (from the notes) that each of the three variables is associated with at least one kind of non-response. The increase in accuracy of the AUC is more than would be expected by chance ($p < 0.001$ apart from wave non-response: $p > 0.06$) but is small 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 0.41 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).

4.2 Using weights to adjust for non-response

Although non-response at wave two of MCS is systematically related to a number of variables measured at or after wave one, we have seen that the models' ability to discriminate between and predict categories of non-response is not high. We now consider what effect the weights generated from the response propensity models have on a longitudinal estimate of interest. We focus on transitions between not working and working across the two waves. As Groves (2006) argues, 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 the variable of interest. We find that all the variables that predict overall non-response are also related to whether or not the main respondent works at wave two, conditional on whether she was working at wave one so we might expect the application of non-response weights to reduce bias. The results are presented in Table 3 and show that, compared with just using the survey weights, the introduction of the non-response weights based on the model underpinning Table 1 leads to small adjustments in the estimated transition probabilities. The consent and vote variables have no additional effect, however, and this is consistent with the marginal increases in accuracy reported in Table 2.

Table 2
Accuracy estimates for enhanced response propensity models, MCS wave two

Accuracy measure	Overall non-response [1]	Non-response type		Non-response reason	
		Wave non-response [2]	Attrition [3]	Refusal [4]	Other non-productive [5]
AUC	0.70	0.72	0.71	0.70	0.77
Gini	0.41	0.44	0.41	0.41	0.54
Logit rank plot: slope	0.47	0.52	0.46	0.45	0.65
Sample size	18,148	16,177	16,745	15,656	16,443

[1] Includes consent (odds ratio (OR) = 2.1, s.e. = 0.20) and vote (OR = 1.4, s.e. = 0.08).
[2] Includes vote 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 (OR = 2.6, s.e. = 0.32), vote (OR = 1.3, s.e. = 0.10) and neighbourhood score (OR = 1.02, s.e. = 0.014).
[5] Includes consent (OR = 1.6, s.e. = 0.20) and vote (OR = 1.5, s.e. = 0.11).

Table 3
Weighted employment transitions (standard errors), MCS wave two

Variable	Survey weights only	Overall weight [1]	Overall weight [2]
No change	0.30 (0.0053)	0.30 (0.0056)	0.31 (0.0056)
Working → not working	0.34 (0.0059)	0.35 (0.0059)	0.35 (0.0060)
Not working → working	0.37 (0.0073)	0.35 (0.0073)	0.35 (0.0073)
Weight range[3]	0.23 – 2.0	0.19 – 4.1	0.19 – 6.3
Sample size	14,891	14,796	14,733

[1] Based on the product of the survey weights and the non-response weights using the model underpinning Table 1.
[2] Non-response weights based on a model that includes consent and vote.
[3] All weights standardised to have mean of one.

5. Discussion

Survey methodologists 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 target sample. On the other hand, a lot can be learnt about the correlates of different types of non-response by drawing on 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 rely on general linear models both to construct inverse probability weights and to inform imputations. Treating the linear predictors from the regression models as response propensity scores and then generating ROCs enables methods for summarising the information in these scores to be used to assess the accuracy of discrimination and prediction for different kinds of non-response.

The application of this approach to the Millennium Cohort Study has shown that, despite using a wide range of explanatory variables, discrimination is rather low. One implication of this finding is that some non-response is generated by circumstantial factors, none of them important on their own, which can reasonably be regarded 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 not missing at random (NMAR) in Little and Rubin's (2002) sense. In other words, it might be changes in circumstances after the previous wave that influences non-response at the current wave.

The implications of our findings for prediction are that it might be difficult to predict which cases will become non-respondents with a high degree of accuracy. If interventions to prevent non-response in longitudinal studies are to be effective then they need to be targeted at those cases least likely to respond because these cases are probably the most different from the respondents and therefore the major source of bias. This is where the ROC approach can be especially useful because, as Swets, Dawes and Monahan (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 bias and variability arising from a sample that is both larger and closer in its characteristics to the target sample.

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

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

Groves, R.M. (2006). Nonresponse rates and non-response bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.

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., R.M. Groves *et al*.). New York: John Wiley & Sons, Inc.

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

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

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

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

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.