Sample loss from cohort studies: patterns, characteristics and adjustments

Ian Plewis,Lisa Calderwood and Sosthenes Ketende

Centre for Longitudinal Studies
Institute of Education
University of London

s.ketende@ioe.ac.uk

Research Methods Festival 2010, Oxford, UK

July 5, 2010

Acknowledgments

This work was part of " **Predicting and preventing non-response in cohort studies**" project, funded by ESRC - Survey Design and measurement initiative.

The research team:-Principal Investigator:

Ian Plewis, Social Statistics, University of Manchester

Co-Investigators

Lisa Calderwood, CLS, Institute of Education, London
Rebecca Taylor, NatCen, London

Research Officer

Sosthenes Ketende, CLS, Institute of Education, London

Motivation

Main objective

What can we learn from modelling the predictors of different kinds of non-response in cohort studies?

For weighting purposes

Is it necessary to update non-response predictors at wave $\mathbf{t}$ with values from wave $\mathbf{t}$-$\mathbf{1}$,where $\mathbf{t}$ $\geq \mathbf{3}$?

Millennium Cohort Study

The study

The Millennium Cohort Study (MCS) is the fourth in the series of internationally renowned cohort studies in the UK.

The sample

At wave one, it 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.

Millennium Cohort Study

Over-sampling

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.

Number of waves

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.

Outline

- Patterns of non-response in MCS, waves 1 to 4
- Predicting non-response at wave 2: summary measures of accuracy
- Alternative models for predicting non-response
- Implications for statistical adjustment

Sample loss from MCS

Wave 1 response rate was 72%

	Wave 2, Age 3 years	Wave 3, age 5 yrs	4, age 7 yrs
Wave NR	8.3%	3.3%	n.a
Attrition	9.9%	16.1%	n.a
Total	**18%**	**20%**	**26%**
Refusal	9.1%	12.2%	18.7%
Other NP	9.2%	7.3%	7.4%
Eligible N	18,385	18,944	18,756

Predictors of overall response at wave 2 (Plewis, 2007)

Variable	Wave NR	Attrition	Refusal	Other NP
Moved residence	√	×	×	√
UK country	√	√	√	√
Family income	×	√	√	×
Refused income qn.	×	×	√	×
Ethnic group	√	√	×	√
Tenure	√	√	×	√
Accom. type	√	√	√	√
Mother's age	√	√	√	√
Education	√	√	√	√
Stable address	√	√	√	√
Cohort member breastfed	√	√	√	√
Long Standing illness	√	√	√	√
Partner present	√	√	√	√
Partner but no IV	√	√	√	√

How might we summarise the accuracy of our predictions?

We can think of the functions estimated from the logistic regressions as statistical prediction rules or risk scores.

How accurate are these risk scores?

We can think of accuracy in two, not necessarily equivalent ways:

I Discrimination sensitivity (true positives) and specificity (1-false positives)

II Prediction

How might we summarise the accuracy of our predictions?

The extent to which risk scores discriminate between respondents and non-respondents is an indication of how effective our statistical adjustments are going to be.

The extent to which risk scores predict whether a case will be a non-respondent in the next wave is an indication of whether any intervention to reduce non-response will be successful.

How might we summarise the accuracy of our predictions?

Discrimination

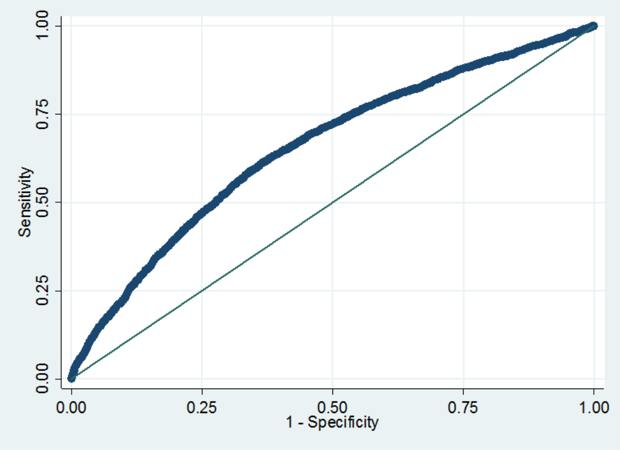
We can plot the true positive fraction (i.e. sensitivity) against the false positive fraction (i.e. 1 - specificity). This is known as a Receiver Operating Characteristic (ROC) curve.

The area under the ROC is a measure of discrimination (AUC varies from 0.5 to 1).
The Gini coefficient;

$$G = 2 \times (AUC - 1)$$

is perhaps a more natural measure, as it varies from 0 to 1.

ROC curve



Sensitivity
0.00
0.25
0.50
0.75
1.00
1 - Specificity
0.00
0.25
0.50
0.75
1.00

How might we summarise the accuracy of our predictions?

Prediction

we can plot the logit of the quantiles of the risk score distribution against the logit of the quantiles of the proportional ranks and estimate the slope.

This is a logit rank plot (Copas, 1999) and the slope will be close to one if the prediction is good.

Accuracy measures, wave 2

	AUC	GINI	Slope - logit rank plot	Prevalence
Overall NR	0.69	0.39	0.45	0.19
Wave NR	0.71	0.43	0.52	0.078
Attrition	0.69	0.39	0.41	0.11
Refusal	0.69	0.37	0.37	0.091
Other NP	0.76	0.52	0.58	0.092

95% confidence limits generally $\pm$ **0.02**

Adding an explanatory variable

Consent to linkage of birth records to administrative health records at wave 1 is highly predictive of non-response at wave two.

	Without Consent		With consent	
	Gini	Slope, logit rank plot	Gini	Slope, logit rank plot
Overll NR	0.39	0.45	0.40	0.47
Wave NR	0.43	0.52	0.43	0.53
Attrition	0.39	0.41	0.41	0.46
Refusal	0.37	0.37	0.39	0.42
Other NP	0.52	0.58	0.52	0.64

Adding an explanatory variable

Prediction is improved by introducing consent but the effects on discrimination are small.

However, even with consent, our ability to predict different kinds of non-response is not great and therefore targeted interventions might not be worthwhile.

Do variables measured at wave $\mathbf{t+1}$ predict wave non-response at wave $\mathbf{t}$?

Change in accomodation type	$\surd$
Change in tenure	$\times$
Change in partnership status	$\surd$
Family income at wave $\mathbf{t+1}$	$\surd$

Gini coefficient for wave 2 rises from **0.43** to **0.46**.

Alternative strategies for predicting non-response at wave **t**

Option 1

Use wave 1 variables, wave 1 values, wave 1 coefficients

Option 2

Use wave 1 variables, wave 1 values, wave (**t-1**) coefficients

Option 3

Use wave 1 variables, wave (**t-1**) values, wave (**t-1**) coefficients

Option 4

Use wave (**t-1**) variables, values, coefficients

Alternative strategies for predicting non-response at wave **t**

Results for MCS, wave 4:

Gini = 0.36; n = 17862

Gini = 0.37, n = 17862

Gini = 0.36, n = 12729

i.e. discrimination essentially the same for approaches (a) to (c).

Predictors at waves 2 and 4

Variable	Wave 2	Wave 3
Moved residence	$\surd$	$\times$
Country	$\surd$	$\surd$
Family income	$\surd$	$\surd$
Refused income qn.	$\surd$	$\surd$
Ethnic group	$\surd$	$\surd$
Tenure	$\surd$	$\times$
Accommodation type	$\surd$	$\surd$
Mothers age	$\surd$	$\surd$
Education	$\surd$	$\surd$
Stable address	$\surd$	$\surd$
Cohort member breast fed	$\surd$	$\surd$
Longstanding illness	$\surd$	$\times$
Partner present	$\surd$	$\surd$
Partner but no IV	$\surd$	$\surd$
Consent for linkage	$\surd$	$\times$

Implications for:

Statistical adjustment via Inverse Probability Weighting

Models developed to generate weights at wave 2 might be satisfactory for later waves, i.e. efforts to generate models for weights at each wave that are based on different sets of variables at each wave might be misplaced.

Statistical adjustment via Multiple Imputation

Imputation models can be improved by using wave **t+k** measures for imputation at wave t.

Statistical adjustment via Selection Modelling

Auxiliary variables or para data can be used as instruments in joint models of selection and outcome (Heckman models, Bayesian models etc.).

Reference

Further details of this are available from

Plewis, I; Calderwood, L and Ketende, S.(2009) Sample loss from cohort studies: patterns, characteristics and adjustments. *Statistics Canada International Symposium Series - Proceedings, Symposium 2009: Longitudinal Surveys: from Design to Analysis*