Assessing Optimal Strategies to Reduce Non-response in Longitudinal Studies

Natalie Shlomo and Ian Plewis
Social Statistics Discipline
School of Social Sciences
University of Manchester
Natalie.Shlomo@manchester.ac.uk

Topics Covered:

- Introduction
- UK Millennium Cohort Study
- Response propensity model
- R-indicators under complex survey designs
- Results
- Future work

Introduction

- Longitudinal studies need to retain sample members over time to remain representative of target population
  - How effective are strategies to retain sample members in a longitudinal study?
  - Which sample members should be the targets of intervention to improve the quality of response?
- Use R-indicators to partially address these questions

UK Millennium Cohort Study

- Wave one sample: 18,818 babies in 18,552 families born in the UK over a 12 month period during the years 2000-2001 and living in selected UK electoral wards at age 9 months
- Sample frame: child birth register
- Boost samples in areas with high proportions of Black and Asian Families, disadvantaged areas and three smaller UK countries over represented
- Design weights with respect to the sample size range from 2.0 (England advantaged stratum) to 0.23 (Wales disadvantaged stratum)
- First 4 waves: 9 months, 3, 5 and 7 years old

UK Millennium Cohort Study

- Face to face interviewing, partners interviewed where possible, data collected from cohort members and siblings
- Standard practice – reissue all eligible cases at wave t conditional on their being in the observed sample at wave 1
- Cases ineligible: emigration, institutional care or child death
- Exceptions: hard refusals not reissued and majority of eligible cases that were unproductive at waves 2 and 3 not reissued at wave 4
- Overall wave 1 response rate was 72%:

England – Advantaged 76%, Disadvantaged 72%, Ethnic 66%
Wales – Advantaged 79%, Disadvantaged 72%
Scotland – Advantaged 77%, Disadvantaged 74%
N.Ireland– Advantaged 68%, Disadvantaged 64%

Response Propensity Model

- Explanatory variables predictors of non-response taken at wave 1 (so representativeness is with respect to wave 1 and not the target population):

Family income (8), Ethnic group of cohort child (6), Accommodation type (2), Tenure (3), Mother's age (2), Mother's education (7), Child breast fed (2), Mother long term illness (2), Family status (2), Mother voted in last election (2), Mother gave consent to record linkage (2), Provided a stable address at wave 1 (2), Change of address between waves 1 and 2 (2), Interactions with Tenure and Accommodation type

Response Propensity Model

$$\text{logit}\ (\rho_i) = \sum_{k=0}^{K} \beta_k x_{ki}$$

where $\rho_i = E(r_i)$ is the probability of not responding for unit i ($i=1,..,n$), $r_i=0$ for a response and $r_i=1$ for non-response, and x_k are explanatory variables ($x_o=1$)

- ML estimates of β_k leading to predicted probabilities or propensities of responding where

$$\hat{\rho}_i = e^{\sum \hat{\beta}_k x_{ki}} \Big/ (1 + e^{\sum \hat{\beta}_k x_{ki}})$$

- Take into account sample design (disproportionate stratification and clustering) using SAS Proc SurveyLogistic
- Use predicted probabilities to estimate R-indicators

R-Indicators

- $\hat{R}_\rho = 1 - 2\hat{S}_\rho$ where $\hat{S}_\rho^2 = (N-1)^{-1}\sum_s d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)^2$, $d_i = \pi_i^{-1}$ is the design weight , $\hat{\rho}_i$ is the estimated propensity, $\hat{\bar{\rho}}_U = (\sum_s d_i \hat{\rho}_i)/N$

- Unconditional variable partial R-indicator measures the distance to representative response for single variable Z based on 'between' variance of the response propensities on categories *k*, *k=1,…,K*

- Unconditional categorical partial R-indicator for Z=k:

$$\hat{R}_{\rho(u),k} = \sqrt{\frac{\hat{N}_k}{N}}\left(\hat{\bar{\rho}}_k - \hat{\bar{\rho}}_U\right) \qquad \hat{\bar{\rho}}_k = \frac{1}{\hat{N}_k}\sum_{s_k} d_i \hat{\rho}_i \qquad \hat{N}_k = \sum_{s_k} d_i$$

s_k is the set of sample units in category k

R-Indicators

- Conditional partial R- indicator measures lack of representative response due to variable Z conditional on all other remaining variables (X^{-}) based on the 'within' variance of the response propensities on categories j, $j=1,\ldots,J$ of X^{-}
- For Millennium Cohort Study, replace population size N with the sample size n, $\hat{N}_k$ is replaced by $\hat{n}_k = \sum_{s_k} d_i$, $\hat{\bar{\rho}}_U$ is replaced by $\hat{\bar{\rho}}_s = \left(\sum_s d_i \hat{\rho}_i\right)/n$, and $\hat{\bar{\rho}}_k = \frac{1}{\hat{n}_k}\sum_{s_k} d_i \hat{\rho}_i$
- Estimates of variances for R-indicator and partial R-indicators adapted to handle complex survey designs

Re-issuing Strategies

Strategy	Explanation
Standard Practice	
S2	reissue all eligible cases at wave 2 conditional on being observed at wave 1
S3	reissue all eligible cases at wave 3 conditional on being observed at wave 1
S4	reissue all eligible cases at wave 4 conditional on being observed at wave 1
Hypothetical strategy - only reissue productive cases from previous waves	
P3.2	Only reissue at wave 3 cases that were productive at wave 2
P4.23	Only reissue at wave 4 cases that were productive at waves 2 and 3
P4.3	Only reissue at wave 4 cases that were productive at wave 3
Hypothetical strategy - of not reissuing refusals from previous waves	
C3.2	Only reissue at wave 3 cases that were not refusals from wave 2
C4.32	Only reissue at wave 4 cases that were not refusals in waves 2 and 3
C4.3	Only reissue at wave 4 cases that were not refusals from wave 3
Hypothetical strategy of not reissuing at wave 4 cases that were not productive at wave 2 but were productive at wave 3	
W4	

Re-issuing Strategies

Row label	Cases lost	Percentage of actual productive sample
S2	n.a.	-
S3	n.a.	-.
P3.2	1,444	9.5%
C3.2	473	3.1%
S4	n.a.	-
P4.23	1,668	12.0%
P4.3	639	4.6%
C4.23	536	3.9%
C4.3	216	1.6%
W4	1,029	7.4%

- Sample size at wave 1 is 18,552 but some cases omitted due to item nonresponse at wave 1
- Sample sizes after omitting ineligible cases: 18,148, 17,990, 17,819 for waves 2 to 4

Results of R-Indicators

Strategies	R-indicator (CI)	Difference from Standard Practice (CI*)
S2	0.761 (0.745 – 0.777)	-
S3	0.773 (0.758 – 0.788)	-
P3.2	0.690 (0.673 – 0.707)	-0.083* (-0.095 - -0.071)
C3.2	0.749 (0.733 – 0.765)	-0.024* (-0.031 - -0.017)
S4	0.720 (0.703 - 0.738)	-
P4.23	0.643 (0.626 - 0.661)	-0.077* (-0.089 - -0.065)
P4.3	0.695 (0.677 - 0.713)	-0.026* (-0.033 - -0.019)
C4.23	0.700 (0.682 - 0.718)	-0.021* (-0.028 - -0.013)
C4.3	0.716 (0.699 - 0.733)	-0.005 (-0.008 - -0.001)
W4	0.670 (0.654 - 0.687)	-0.050* (-0.060 - -0.040)

* CI of difference calculated by bootstrapping

Results of R-Indicators

- S2 and S3 show little difference in representativeness between waves two and three given standard reissuing practice
- S4 shows lower representativeness at wave 4 compared to the first two waves using standard reissuing practice
- P show that representativeness falls if only productive cases from previous waves are reissued and R-indicators are all significantly lower than for the standard practice in each wave
- The R-indicator for P4.3 is significantly higher to the R-indicator P4.23 ($p<0.001$) and is closer to the R-indicator of the standard practice S4

Results of R-Indicators

- Comparing C3 to S3 suggests that representativeness is less compromised if refusals from previous waves are not reissued (R-indicator more similar to standard practice)
- Comparing strategies of cases that refused just at wave three not being reissued at wave four (C4.3) to those that refused at either wave two or wave three (C4.23), there is a significant increase in representativeness for C4.3 ($p<0.001$)
- C4.3 similar representativeness compared to standard practice in S4 with a non-significant difference in R-indicator
- Strategy C4.3 had the lowest number of dropped cases
- W4 shows that representativeness is significantly reduced if wave non-respondents are not reissued

Results of Partial R-Indicators – Education Qualifications

Strategies	Unconditional Partial R-indicator (CI)	Conditional Partial R-indicator (CI)
S2	0.059 (0.053 – 0.066)	0.015 (0.007 – 0.023)
S3	0.059 (0.053 – 0.066)	0.016 (0.009 – 0.023)
P3.2	0.083 (0.075 – 0.091)	0.021 (0.013 – 0.029)
C3.2	0.066 (0.059 – 0.074)	0.018 (0.010 – 0.026)
S4	0.076 (0.069 – 0.084)	0.020 (0.011 – 0.028)
P4.23	0.099 (0.091 – 0.108)	0.025 (0.016 – 0.033)
P4.3	0.085 (0.078 – 0.093)	0.022 (0.013 – 0.031)
C4.23	0.083 (0.075 – 0.091)	0.021 (0.013 – 0.030)
C4.3	0.078 (0.071 – 0.085)	0.020 (0.011 – 0.028)
W4	0.090 (0.082 – 0.098)	0.023 (0.014 – 0.031)

* All differences in strategies significant to standard practice

Results of Partial R-Indicators

- Unconditional partial R-indicators significantly different from zero so highest educational qualification contributes to lack of representativity and highest for P4.32
- Other variables with high unconditional partial R-indicator (not shown): Family income group (8), Family status (4), Tenure (3), Ethnic group (6),
- Conditional partial R-indicators significantly different from zero so educational qualification continues to contribute to lack of representativity conditional on other variables
- Unconditional partial R-indicators larger than conditional partial R-indicators suggesting that the impact of each variable is reduced when controlling for other variables (multicollinearity of auxiliary variables)

Targeting Data Collection

- Categorical partial R-indicators indicates which groups of individuals (based on wave 1 characteristics) should be targeted for data collection
- Use the information to reassess the reissuing strategy wave 4:

Strategy S4: R=0.720 (0.703 - 0.738)

P4.23 R=0.643 (0.626 - 0.661)

P4.3 R=0.695 (0.677 - 0.713)

Not reissue cases that were unproductive at earlier waves ONLY if they belonged to majority ethnic group, had some educational qualifications, own accommodation, etc.

P4.23 R=0.694 (0.677 – 0.712) (dropped cases: 1275)

P4.3 R=0.733 (0.716 – 0.751) (dropped cases: 323)

Targeting Data Collection

- Add in paradata to response model: interviewer questionnaire on neighborhood – score from missing, 0 to 20 (in 5 groups)
- Use the information to reassess the reissuing strategy wave 4:

Strategy P4.23 original: R=0.643 (0.626 – 0.661)

With paradata in model: R=0.561 (0.544 – 0.578)

Not reissue cases that were unproductive at earlier waves ONLY (if they belonged to majority ethnic group, had some educational qualifications, own accommodation, etc.)

R=0.608 (0.589 – 0.627) (dropped cases 323)

ONLY (lives in a good neighborhood)

R=0.659 (0.643 – 0.676) (dropped cases 2990)

BOTH of the above

R=0.683 (0.665 – 0.701) (dropped cases: 3124)

Conclusions and Future Work

- Not re-issuing refusals (C) has less of an impact on representativity with respect to wave 1 than re-issuing productive cases (P) since many continue to be refusals at subsequent waves
- Indication that only re-issuing productive cases (P) impacts on representativity with respect to wave 1 but can be mitigated by targeting refusals based on profiles from R-indicators
- Including paradata such as neighborhood observations can be used to target data collection
- Further work: use ROC curve and estimate a cost-model with an optima response propensity below which interventions should be carried out (cost model should take into account clustered designs)

Thank you for your attention!