Handling non-response in the COVID-19 surveys

17 June 2021

Richard Silverwood
Associate Professor of Statistics, Chief Statistician

Centre for Longitudinal Studies, UCL Social Research Institute

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

Outline

CENTRE FOR LONGITUDINAL STUDIES

1. Introduction
2. Target population and response
3. Derivation of non-response weights
4. Effectiveness of non-response weights
5. Implementation of non-response weights

Introduction

Introduction

CENTRE FOR LONGITUDINAL STUDIES

- Non-response is common in longitudinal surveys.
- Missing values mean less efficient estimates because of reduced size of analysis sample.
- Also introduces potential for bias since respondents are often systematically different from non-respondents.
- Well known methods for dealing with missing data include multiple imputation (MI), inverse probability weighting (IPW), and full information maximum likelihood (FIML).

Introduction

CENTRE FOR LONGITUDINAL STUDIES

- To correct for non-response in the COVID-19 surveys, non-response weights are provided, so that IPW analysis can be undertaken.
- Non-response weights capitalise on the rich data cohort members have provided over many years.

Target population and response

Target population

CENTRE FOR LONGITUDINAL STUDIES

- Target population of each cohort defined as individuals born in the specified birth period who are alive and still residing in the UK.
- Non-response weights designed to make weighted results from COVID-19 survey respondents representative of the target population.
- COVID-19 surveys also issued to a relatively small number of cohort members who had already emigrated from the UK – we do not derive non-response weights for such individuals.

Target population

CENTRE FOR LONGITUDINAL STUDIES

In MCS:

- Only derived non-response weights for cohort members (i.e. not parents).
- Only derived non-response weights for singletons and one twin or triplet from each twin pair/triplet set.
- (Triplet families subsequently excluded from COVID-19 survey dataset.)

Response within target population

CENTRE FOR LONGITUDINAL STUDIES

Cohort	Target pop.	Response within target population		
		Wave 1	Wave 2	Wave 3
NSHD	3,758	1,170 (31.1%)	1,488 (39.6%)	1,325 (35.3%)
NCDS	15,291	5,119 (33.5%)	6,228 (40.7%)	6,757 (44.2%)
BCS70	17,486	4,132 (23.6%)	5,236 (29.9%)	5,684 (32.5%)
Next Steps	15,770	1,876 (11.9%)	3,609 (22.9%)	4,167 (26.4%)
MCS cohort members	19,243	2,609 (13.6%)	3,233 (16.8%)	4,422 (23.0%)
Total	71,548	14,906 (20.8%)	19,794 (27.7%)	22,355 (31.2%)

Derivation of non-response weights

Derivation of non-response weights
Overview

CENTRE FOR LONGITUDINAL STUDIES

At each wave and within each cohort separately:

1. Within sample corresponding to target population, model COVID-19 survey response conditional on a common set of covariates using logistic regression.
2. For COVID-19 survey respondents, predict probability of response from model.
3. Calculate non-response weight as inverse of probability of response.
4. Examine distribution of weights across cohorts to decide whether truncation may be desirable; apply truncation if so.
5. Calibrate weights so they sum to number of respondents in each cohort.

Derivation of non-response weights
Stage 1: Response model

CENTRE FOR LONGITUDINAL STUDIES

- Selection of covariates in response model informed by results of the CLS Missing Data Strategy and assumed associations with the probability of response and/or with key COVID-19 survey variables.
- Aimed to use broadly same set of variables in each cohort to ensure consistency.
- Not possible to include identical sets of variables due to data being collected at different ages and using different questions.
- Further technical details in User Guide.

Derivation of non-response weights
Stage 1: Response model

CENTRE FOR LONGITUDINAL STUDIES

Sex
Ethnicity
Parental social class
Number of rooms at home/persons per room
Cognitive ability
Early life mental health
Voting
Membership in organisations

Internet access prior to web survey
Consent for biomarkers
Consent for linkages
Educational qualifications
Economic activity
Partnership status
Psychological distress
BMI

Self-rated health
Smoking status
Maternal mental health
Social capital/social support
Income
Number of non-responses across all previous sweeps
Response at COVID-19 Wave 1 and 2 surveys*

Derivation of non-response weights
Stages 2-5

CENTRE FOR LONGITUDINAL STUDIES

1. Within sample corresponding to target population, model COVID-19 Survey response conditional on a common set of covariates using logistic regression.
2. For COVID-19 survey respondents, predict probability of response from model.
3. Calculate non-response weight as inverse of probability of response.
4. Examine distribution of weights across cohorts to decide whether truncation may be desirable; apply truncation if so.
5. Calibrate weights so they sum to number of respondents in each cohort.

Effectiveness of non-response weights

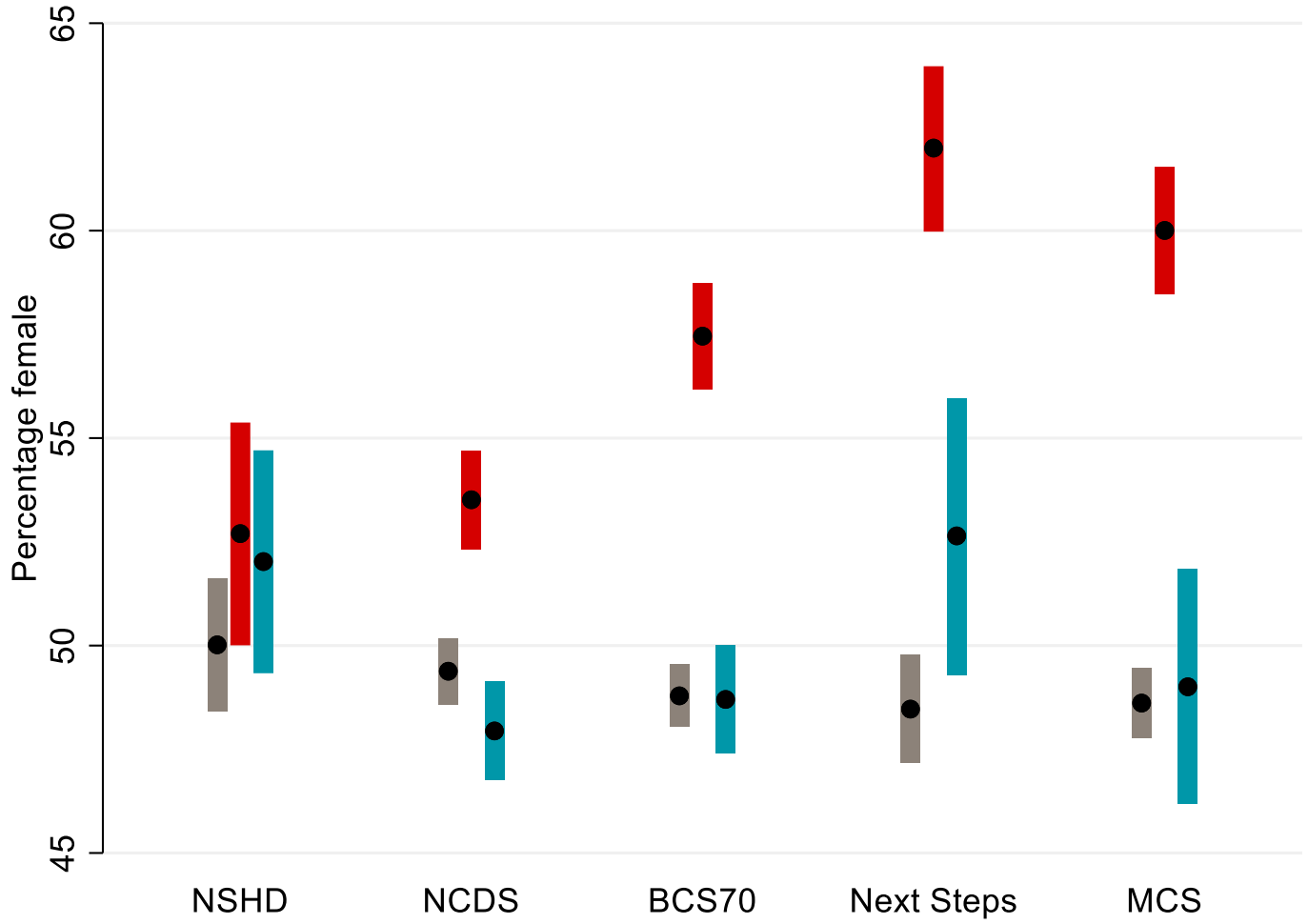
Effectiveness of non-response weights

CENTRE FOR LONGITUDINAL STUDIES

- To examine effectiveness of non-response weights in restoring sample representativeness we conducted several analyses.
- We considered the distribution of sex in each cohort, which is observed at baseline in virtually all cohort members.
- We compared the distribution of sex:
  - In all cohort members;
  - In COVID-19 survey respondents only;
  - In COVID-19 survey respondents after application of non-response weights.

Effectiveness of non-response weights

Sex



Grey: all cohort members; red: COVID-19 Wave 3 survey respondents only; blue: COVID-19 Wave 3 survey respondents after application of non-response weights.

Implementation of non-response weights

Implementation of non-response weights

CENTRE FOR LONGITUDINAL STUDIES

- Non-response weights provided as part of COVID-19 survey dataset.
- Non-response weights already combined with design weights where necessary (NSHD, Next Steps and MCS) to produce a combined weight (CW3_COMBWT).
- In other cohorts (NCDS and BCS70), same variable name used for consistency but is simply the non-response weight.

Implementation of non-response weights

CENTRE FOR LONGITUDINAL STUDIES

- NCDS & BCS70
  No study design to take into account
  → Just use combined weight
- NSHD
  Design weight to take into account
  → Just use combined weight
- Next Steps & MCS
  Design weight plus primary sampling unit, strata and finite population correction (MCS) to specify
  → svyset the data then use svy prefix in Stata

Implementation of non-response weights

CENTRE FOR LONGITUDINAL STUDIES

- Illustrated by estimating proportion of individuals reporting having Coronavirus in each cohort.
- CW3_COVID19 initially coded 1 “Yes, confirmed by a positive test”, 2 “Yes, based on strong personal suspicion”, 3 “Unsure” and 4 “No”.
- We first collapse the categories to form a binary yes/no variable.
- Illustrative analyses in Stata, but similar in other software.

Implementation of non-response weights
NSHD (note: Wave 2 data)

CENTRE FOR LONGITUDINAL STUDIES

```
. proportion CW2_COVID19 [pweight=CW2_COMBWT] if CW2_COHORT==6,
  citype(agresti)

Proportion estimation                 Number of obs    =       1,485

------------------------------------------------------------------------
             |                                        Agresti-Coull
             | Proportion   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------------
CW2_COVID19 |
          No |   .9778868   .0050314       .9689836    .9843239
         Yes |   .0221132   .0050314       .0156761    .0310164
------------------------------------------------------------------------
```

Implementation of non-response weights
NCDS

CENTRE FOR LONGITUDINAL STUDIES

```
. proportion CW3_COVID19 [pweight=CW3_COMBWT] if CW3_COHORT==1,
  citype(agresti)

Proportion estimation                 Number of obs    =       6,722

------------------------------------------------------------------------
             |                                        Agresti-Coull
             | Proportion   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------------
 CW3_COVID19 |
          No |   .8871005   .0078628       .8793098    .8944491
         Yes |   .1128995   .0078628       .1055509    .1206902
------------------------------------------------------------------------
```

Implementation of non-response weights
BCS70

CENTRE FOR LONGITUDINAL STUDIES

```
. proportion CW3_COVID19 [pweight=CW3_COMBWT] if CW3_COHORT==2,
  citype(agresti)

Proportion estimation                 Number of obs    =       5,633

------------------------------------------------------------------------
             |                                        Agresti-Coull
             | Proportion   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------------
CW3_COVID19 |
          No |    .850274   .0080814       .8407147    .8593559
         Yes |    .149726   .0080814       .1406441    .1592853
------------------------------------------------------------------------
```

Implementation of non-response weights
Next Steps

CENTRE FOR LONGITUDINAL STUDIES

```
. svyset CW3_SAMPPSU [pweight=CW3_COMBWT], strata(CW3_SAMPSTRATUM)

. svy: proportion CW3_COVID19 if CW3_COHORT==3, citype(agresti)
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      37          Number of obs   =      4,095
Number of PSUs   =     645          Population size = 4,067.1908
                                    Design df       =        608

------------------------------------------------------------------
             |             Linearized         Agresti-Coull
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
 CW3_COVID19 |
          No |   .7813959   .0133312       .7540945     .806448
         Yes |   .2186041   .0133312        .193552    .2459055
------------------------------------------------------------------
```

Implementation of non-response weights
MCS

CENTRE FOR LONGITUDINAL STUDIES

```
. svyset CW3_SPTN00 [pweight=CW3_COMBWT], strata(CW3_PTTYPE2) fpc(CW3_NH2)

. svy: proportion CW3_COVID19 if CW3_COHORT==4, citype(agresti)
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =       9          Number of obs   =      4,348
Number of PSUs   =     398          Population size = 4,471.2108
                                    Design df       =        389

------------------------------------------------------------------
             |             Linearized          Agresti-Coull
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
 CW3_COVID19 |
          No |   .7775335   .0098784       .7575102    .7963491
         Yes |   .2224665   .0098784       .2036509    .2424898
------------------------------------------------------------------
```

Analysing data across multiple timepoints

CENTRE FOR LONGITUDINAL STUDIES

- COVID-19 survey non-response weights designed to make analyses of respondents at that survey representative of target population.
- If analytical sample largely driven by non-response to specific COVID-19 survey response, non-response weights at that wave likely to perform well.
- If analytical sample doesn't (approximately) correspond to respondents at specific COVID-19 survey then a bit more complicated…
- Alternative approaches (custom weights, MI, …) may be preferred.

Thank you. Any questions?

CENTRE FOR
LONGITUDINAL
STUDIES

Derivation of non-response weights
Stage 1: Response model

CENTRE FOR LONGITUDINAL STUDIES

- Missing covariate values handled using MI, conducted in each cohort separately.
- Imputation model included above variables, COVID-19 survey response and, for relevant cohorts (NSHD, Next Steps and MCS), the design weight.
- Five imputed datasets were created using chained equations.
- Models for COVID-19 survey response fitted in each imputed dataset and combined using standard rules.