UCL Institute of Education

Handling missing data in the CLS cohort studies

User guide (Version 2)

March 2026

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

Contact

Data queries: help@ukdataservice.ac.uk

Questions and feedback about this user guide: clsdata@ucl.ac.uk.

Authors

Richard Silverwood, Martina Narayanan, Aase Villadsen, Brian Dodgeon, Michalis Katsoulis, George Ploubidis.

How to cite this guide

Silverwood, R., Narayanan, M., Villadsen, A., Dodgeon, B., Katsoulis, M., Ploubidis, G. (2026) *Handling missing data in the CLS cohort studies: User guide (Version 2)*. London: UCL Centre for Longitudinal Studies.

Data citation and acknowledgement

You should cite the data and acknowledge CLS following the guidance from cls.ucl.ac.uk/data-access-training/citing-our-data/.

Centre for Longitudinal Studies

Centre for Longitudinal Studies (CLS)

UCL Social Research Institute

University College London

20 Bedford Way, London WC1H 0AL

www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre. It is home to a unique series of UK national cohort studies. It is part of the UCL Social Research Institute, based at the UCL Institute of Education.

This document is available in alternative formats. Please email the Centre for Longitudinal Studies at clsdata@ucl.ac.uk

Contents

Missing data 1
CLS Missing Data Strategy 1
Multiple imputation 3
Running example 3
Deciding on the set of auxiliary variables 11
Developing the imputation model 14
Recoding variables 16
Preparing the data for imputation 17
Conducting the imputation 18
Troubleshooting the imputation 21
Checking the imputed values 22
Fitting the analysis model 28
Checking the analysis model 31
Missing not at random sensitivity analyses 34
Inverse probability weighting for missing data handling 38
Further reading 48
References 50
Appendices 53

Missing data

Non-response is common in longitudinal surveys. Missing values due to non-response mean less efficient estimates because of the reduced size of the of the analysis sample but also introduce the potential for bias since respondents are often systematically different from non-respondents.

Missing data are typically characterised by their corresponding missing data mechanism: (i) missing completely at random (MCAR), meaning that missingness does not depend on either observed or unobserved values (i.e. is completely at random); (ii) missing at random (MAR), meaning that, given the observed values, missingness does not depend on unobserved values; or (iii) missing not at random (MNAR), meaning that missingness depends on unobserved (and possibly observed) values (1, 2).

A complete case analysis (one restricted to study participants with complete data) is valid if data are MCAR, but also under MNAR if missingness is independent of the outcome variable given the covariates in the model (3). If data are MAR then popular analysis approaches include multiple imputation (MI) (2, 4, 5), inverse probability weighting (IPW) (6, 7), and full information maximum likelihood (FIML) (8, 9).

CLS Missing Data Strategy

Identifying predictors of non-response can help make the MAR assumption more plausible and has implications for missing data analysis with principled methods such as MI, IPW and FIML.

In the 1958 National Child Development Survey (NCDS) (10), 1970 British Cohort Study (BCS70) (11), Next Steps (12), and Millennium Cohort Study (MCS) (13) we have implemented a systematic data-driven approach to identify predictors of non-response at each sweep of data collection (wave 8 only for Next Steps). We found a number of variables to be consistently associated with non-response, across sweeps and across cohorts. These include disadvantaged socio-economic background in childhood, worse mental health and lower cognitive ability in early life. A full list of the

identified predictors of non-response at each sweep in each study is available in the Appendices. For details of the approach used in identifying these variables we refer you to the published papers (10-13).

These variables can be straightforwardly used as “auxiliary variables”, i.e. variables not in the substantive model of interest, in analyses with principled methods. Their appropriate use can help maximise the plausibility of the MAR assumption in order to reduce bias due to missing data and have the potential to help restore sample representativeness despite selective non-response.

Non-dynamic longitudinal models (regression based analyses, including interactions and/or formal mediation, using data from at least two stages of the life course) are the most common analyses using CLS cohort data. In such analyses, MI is plausible and arguably more flexible than FIML since auxiliary variables are more easily included in the imputation phase.

Dynamic longitudinal models (explicitly quantifying change over time, for example growth models, mixture modules, latent transitions models, fixed/random effects, multilevel models, generalised estimating equations, generalised methods of moments) are less common in analyses of CLS cohort data. In such analyses, it is more difficult to incorporate the longitudinal structure into the imputation model. FIML is more flexible, so may be preferred, but approaches and related software for incorporating hierarchal/multilevel structures in the imputation phase are available (5, 14, 15).

The approach to missing data handling outlined in this User Guide should not be viewed as a roadmap to undertaking the single recommended analysis in the presence of missing data. Precisely how a given analysis should be undertaken will inevitably depend on a variety of factors, such as the research question of interest, the availability of data, and the analysis model being used, which are largely beyond the scope of this User Guide. The aim of the User Guide is simply to describe and illustrate a straightforward approach to missing data handling, while detailing some more general considerations around missing data along the way.

Multiple imputation

In MI, the analyst specifies an appropriate imputation model, from which a series of imputed datasets are created. Each imputed data set is analysed using the substantive model of interest and the results are combined using standard rules (2), resulting in standard errors that incorporate the variability in results between the imputed data sets. In this way, uncertainty about the missing data is appropriately accounted for in the inference. Over recent years, MI has been widely adopted because it is practical for applied researchers in a wide range of settings and can be undertaken using standard statistical software (5).

Due to its applicability in the majority of typical analyses of data from CLS cohort studies and its ease of use with standard software, this User Guide will focus on MI. It will take you through the recommended steps of conducting an MI analysis in Stata as applicable in typical analyses. We show Stata code and output, but the basic analytic steps are broadly applicable, and similar commands are available in R and other statistical software. There exist many general guides to conducting MI analyses in Stata, alongside the very detailed Stata help files (starting from `help mi`). This User Guide will focus on the most relevant issues and commands for undertaking the most common types of analyses of data from CLS cohort studies.

Running example

The application of the CLS Missing Data Strategy will be illustrated through the use of a running example using data from NCDS, though all the issues discussed are relevant to analyses of data from any of the CLS cohort studies. The focus of the analysis is the association between partnership status at age 42 (ParStat42R) and income at age 55 (total_income55) (16). Partnership status at age 42 is a 3-level categorical variable (married/cohabiting vs. separated/divorced/widowed vs. single and never married) and income at age 55 is a continuous variable measured in British Pounds per week, log-transformed prior to analysis (log_total_income55) to deal with the characteristic positive skew of income data. Data at age 55 were collected using a sequential mixed mode design, but this feature will not be considered further here.

While this example has been chosen to closely resemble the sort of analysis that may be undertaken in practice, the results obtained should not be over-interpreted from a substantive perspective as some features of the analysis would likely be undertaken differently were this a “real” analysis.

The analysis of interest is a linear regression of log_total_income55 on ParStat42R adjusted for a number of potential confounders relating to birth (sex, birthweight, maternal smoking during pregnancy, maternal age, breastfeeding), the cohort member’s parents (mother working up to 5, parents read to child, parental interest in school, divorce, separation from child, mother not married), socioeconomic position (paternal social class at birth, financial difficulties, housing tenure, housing difficulties), early life (cognitive ability, enuresis, summary of objectively assessed health conditions, body mass index, mental health, behaviour), and midlife (age 42) (income, economic activity, mental health, education, chronic illness, disability). The variables included in the analysis model are listed in the below table.

Outcome/ exposure/ covariate	Covariate domain	Variable description	Variable name	Variable type
Outcome		Log-transformed total income at age 55	log_total_income55	Continuous
Exposure		Partnership status at age 42	ParStat42R	Categorical
Covariate	Birth	Sex	n622	Binary
Covariate	Birth	Birthweight	LBW	Binary
Covariate	Birth	Maternal smoking during pregnancy	smpreg	Binary
Covariate	Birth	Maternal age	n553	Continuous
Covariate	Birth	Breastfeeding	bfever	Binary
Covariate	Cohort member's parents	Mother working up to 5	maw5	Binary
Covariate	Cohort member's parents	Parents read to child	MotherNeverReads7	Binary
Covariate	Cohort member's parents	Parental interest in school	NoIntEdu	Binary
Covariate	Cohort member's parents	Parental divorce	DivBy7	Binary
Covariate	Cohort member's parents	Separation from child	SepMore1Month	Binary
Covariate	Cohort member's parents	Mother not married	MumNotMarried	Binary
Covariate	Socioeconomic position	Paternal social class at birth	SocialClassHusband	Categorical
Covariate	Socioeconomic position	Financial difficulties	DifficultiesFinancial	Binary

Outcome/ exposure/ covariate	Covariate domain	Variable description	Variable name	Variable type
Covariate	Socioeconomic position	Housing tenure	HousingTenure_7	Binary
Covariate	Socioeconomic position	Housing difficulties	DifficultiesHousing	Binary
Covariate	Early life	Cognitive ability	CogAbil7	Continuous
Covariate	Early life	Enuresis	enuresis7	Binary
Covariate	Early life	Summary of objectively assessed health conditions	MedExSum7	Continuous
Covariate	Early life	Body mass index	BMI7	Continuous
Covariate	Early life	Mental health	PsychoMed	Binary
Covariate	Early life	Behaviour	a16_totalscore	Continuous
Covariate	Midlife	Log-transformed total income	log_total_income42	Continuous
Covariate	Midlife	Economic activity	EconAct42R	Categorical
Covariate	Midlife	Mental health	Mal24Age42	Continuous
Covariate	Midlife	Education	NVQ42R	Categorical
Covariate	Midlife	Chronic illness	lsiany2	Binary
Covariate	Midlife	Disability	dmdisab	Binary

A complete case analysis (see below) uses data from only 1896 cohort members and estimates coefficients of -0.18 (95% confidence interval [CI] -0.36, 0.01) comparing separated/divorced/widowed to married/cohabiting and -0.37 (95% CI -0.55, -0.18) comparing single and never married to married/cohabiting. These correspond to 16% (exp(-0.18) = 0.84) and 31% (exp(-0.37) = 0.69) lower income respectively.

```
. regress log_total_income55 i.ParStat42R n622 LBW smpreg n553 bfever maw5 MotherNeverReads7
  NoIntEdu DivBy7 SepMore1Month MumNotMarried i.SocialClassHusband DifficultiesFinancial
  HousingTenure_7 DifficultiesHousing CogAbil7 enuresis7 MedExSum7 BMI7 PsychoMed a16_totalscore
  log_total_income42 i.EconAct42R Mal24Age42 i.NVQ42R lsiany2 dmdisab

      Source |       SS           df       MS      Number of obs   =     1,896
-------------+----------------------------------   F(37, 1858)     =      5.41
       Model |  244.643966        37  6.61199909   Prob > F        =    0.0000
    Residual |  2272.29158     1,858  1.22297717   R-squared       =    0.0972
-------------+----------------------------------   Adj R-squared   =    0.0792
       Total |  2516.93554     1,895  1.32819818   Root MSE        =    1.1059

------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
         log_total_income55 |      Coef.   Std. Err.      T    P>|t|     [95% Conf. Interval]
----------------------------+-------------------------------------------------------------
                 ParStat42R |
Seperated/Divorced/Widowed  |  -.1760115   .0936201    -1.88   0.060     -.359623    .0076001
       Single/NeverMarried  |  -.3652557   .0947965    -3.85   0.000    -.5511745   -.1793368
                                          [Output omitted]
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
```

However, the 1896 cohort members included in the complete case analysis are only a small proportion of the available NCDS sample. Of the original NCDS sample, 15,613 were alive and had not emigrated by age 55, of whom 9137 provided responses for at least some of the items at age 55, though only 6306 of these had data on income at age 55. The other variables in the analysis, being collected at earlier sweeps and of a less sensitive nature than income, have even lower levels of missingness.

Inferences from the complete case analysis only necessarily relate to the population of individuals with complete data on these specific variables. In general, this will not

be the population to which we would wish inferences to relate. It is important to consider the "target population" of a given analysis as this helps assess the extent to which specific analytical approaches (for example, complete case analysis or MI) will be able to address the research question of interest.

An exploration of the extent of missingness in the analysis variables can be conducted using the `misstable summarize` command (see `help misstable`). As well as exploring the extent of missingness in the analysis variables, it can also be helpful to examine the patterns of missing data (i.e. precisely which variables are missing for each NCDS participant). This can be undertaken using `misstable patterns` (see `help misstable`) (and graphical visualisations may also be also be helpful (17)), but these can become more difficult to interpret when the number of included variables is high (for example, with 29 variables there are 2^{29} > 0.5 billion possible missing data patterns, though clearly not all of these could be observed in the dataset).

```
. misstable summarize log_total_income55 ParStat42R n622 LBW smpreg n553 bfever maw5
  MotherNeverReads7 NoIntEdu DivBy7 SepMore1Month MumNotMarried SocialClassHusband
  DifficultiesFinancial HousingTenure_7 DifficultiesHousing CogAbil7 enuresis7 MedExSum7 BMI7
  PsychoMed a16_totalscore Log_total_income42 EconAct42R Mal24Age42 NVQ42R lsiany2 dmdisab

                                                                  Obs<.
                                                +------------------------------
                |                               | Unique
     Variable   |      Obs=.     Obs>.     Obs<. | values         Min          Max
  -------------+--------------------------------+------------------------------
  log_total~55  |      9,307               6,306 |   >500           0     8.146998
    ParStat42R  |      4,651              10,962 |      3           1            3
          n622  |          1              15,612 |      2           0            1
           LBW  |      1,442              14,171 |      2           0            1
        smpreg  |      1,148              14,465 |      2           0            1
          n553  |        972              14,641 |     35           8           47
        bfever  |      2,596              13,017 |      2           0            1
          maw5  |      3,069              12,544 |      2           0            1
  MotherNeve~7  |      2,639              12,974 |      2           0            1
      NoIntEdu  |      3,124              12,489 |      2           0            1
        DivBy7  |      3,098              12,515 |      2           0            1
  SepMore1Mo~h  |      2,832              12,781 |      2           0            1
  MumNotMarr~d  |        970              14,643 |      2           0            1
  SocialClas~d  |      1,723              13,890 |      4           2            5
```

```
Diffuculti~l |     3,901              11,712  |      2          0           1
HousingTen~7 |     2,532              13,081  |      2          0           1
Diffuculti~g |     3,196              12,417  |      2          0           1
    CogAbil7 |     2,675              12,938  |   >500  -3.411307    2.172811
   enuresis7 |     2,552              13,061  |      2          0           1
   MedExSum7 |     2,350              13,263  |     24          0          23
        BMI7 |     3,663              11,950  |   >500   9.280806    28.97114
   PsychoMed |     3,919              11,694  |      2          0           1
a16_totals~e |     5,116              10,497  |     31          0          31
log_total~42 |     6,326               9,287  |   >500          0    8.318254
  EconAct42R |     4,729              10,884  |      5          1           5
  Mal24Age42 |     4,595              11,018  |     25          0          24
      NVQ42R |     4,615              10,998  |      5          0           4
     lsiany2 |     4,637              10,976  |      2          0           1
     dmdisab |     4,596              11,017  |      2          0           1
--------------------------------------------------------------------------------
```

It is also sensible to explore the extent to which the distributions of the analysis variables differ between the complete cases and the non-complete cases. In the presence of a MCAR mechanism, the distributions of all variables should be the same (allowing for sampling variability) in the two groups. If it appears that there are systematic differences between the two groups, then this is suggestive of a MAR or MNAR mechanism. This is an important distinction as a complete case analysis would be valid under MCAR, but not (generally) under MAR or MNAR.

Below, we derive a variable which indicates whether a cohort member is included in the complete case analysis and then see whether the distributions of analysis variables differ between complete cases and non-complete cases, using `tabstat` for continuous variable and `tabulate` for binary/categorical variables.

```
. quietly regress log_total_income55 i.ParStat42R n622 n553 BMI7 CogAbil7 MedExSum7
  a16_totalscore log_total_income42 dmdisab lsiany2 SepMore1Month MotherNeverReads7 NoIntEdu
  smpreg maw5 HousingTenure_7 DifficultiesHousing DifficultiesFinancial DivBy7 LBW bfever
  enuresis7 MumNotMarried PsychoMed i.NVQ42R i.SocialClassHusband Mal24Age42 i.EconAct42R

. gen cc = e(sample)

. foreach var in log_total_income55 n553 CogAbil7 MedExSum7 BMI7 a16_totalscore
  log_total_income42 Mal24Age42 {
    tabstat `var', by(cc) stat(n mean sd)
```

```
}

. foreach var in n622 dmdisab lsiany2 SepMore1Month MotherNeverReads7 NoIntEdu smpreg maw5
  HousingTenure_7 DifficultiesHousing DifficultiesFinancial DivBy7 LBW bfever enuresis7
  MumNotMarried PsychoMed NVQ42R SocialClassHusband ParStat42R EconAct42R {
    tab `var' cc, col
  }
```

Some variables, such as BMI at age 7, are well balanced (mean 15.9 kg/m^2 in both groups):

```
Summary for variables: BMI7
     by categories of: cc

      cc |         N      mean        sd
---------+--------------------------------
       0 |     10054  15.91879   1.79659
       1 |      1896  15.89221  1.728181
---------+--------------------------------
   Total |     11950  15.91457  1.785867
----------------------------------------
```

Other variables, such as parental divorce, display clear imbalances (1.6% in complete cases, 4.5% in non-complete cases):

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

Divorce/se |
  peration |          cc
      by 7 |         0          1 |     Total
-----------+----------------------+----------
        No |    10,142      1,866 |    12,008
           |     95.51      98.42 |     95.95
-----------+----------------------+----------
       Yes |       477         30 |       507
           |      4.49       1.58 |      4.05
-----------+----------------------+----------
     Total |    10,619      1,896 |    12,515
           |    100.00     100.00 |    100.00
```

```
Pearson chi2(1) =  35.0398
```

More formal statistical testing of differences between the groups could also be conducted, for example through use of t-tests or chi-squared tests (subject to the usual underlying assumptions).

Analysing such a small proportion of the available data may raise concerns over the potential for bias and will certainly lead to imprecision in the results obtained. Application of a principled method for missing data handling would therefore be a sensible next step in the analysis. Over the next few sections of this User Guide we will explore how MI could be applied, in generality and then in this analysis specifically.

Deciding on the set of auxiliary variables

In MI, the analyst first specifies an appropriate imputation model, from which a series of imputed datasets are created. The imputation model should include all the variables in the substantive model (exposure(s), outcome(s), confounder(s), etc.) in the form in which they will enter the analysis model (i.e. subsequent to any recoding or transformations). Any interactions between two or more variables within the substantive model must be explicitly included in the imputation model. This ensures that the imputation model is “congenial” or “consistent” with the analysis model. Omission of, say, an interaction term from the imputation model would imply that values are being imputed assuming zero interaction. If this interaction were of substantive interest in the analysis then it would be, unsurprisingly, less likely to be apparent in the MI analysis. Such considerations are therefore very important.

The imputation model should also include the following sets of auxiliary (i.e. not included in the substantive model) variables:

- Variables that are predictive of both the probability of missingness and the underlying missing values themselves.
- Variables that are predictive of the underlying missing values only.

As missing data in NCDS is largely driven by non-response at a given sweep (as opposed to item non-response within sweeps), the first of these auxiliary variables can be selected from the pre-determined sets of variables predictive of non-response at each sweep (see Appendix A). The second of these auxiliary variables should be selected using a combination of substantive/theoretical knowledge and exploration of the data.

A further consideration is the completeness of potential auxiliary variables. All other things being equal, more complete variables should be preferred. Auxiliary variables with very little missing data themselves will add information to the imputation model, improving the quality of the imputed values; auxiliary variables with extensive missingness add limited information to the imputation model and will themselves largely require imputing, adding further uncertainty. As missingness in NCDS (as in almost all cohort studies) increases as time progresses, this suggests favouring auxiliary variables from earlier sweeps, though should not rule out variables from later sweeps with high levels of completeness, particularly if they are thought likely to be strongly associated with variables subject to missingness (even after allowing for all existing variables in the model).

Example

In our running example, the variables in the substantive model are those in the corresponding complete case analysis (exposure, outcome and potential confounders).

Missingness is largely driven by non-response at the NSCD age 55 sweep (since non-respondents at previous sweeps are usually, though not always, also non-respondents at this sweep). Variables that are predictive of non-response at age 55 are precisely those identified in Table A10 of Appendix A. There are 31 such variables in total, though 5 of these already appear in the substantive model and therefore do not require further consideration here. The remaining 26 variables could all be included in the imputation model as auxiliary variables, but there is evidence to suggest that variables that are predictive of the chance of missing values but are not predictive of the underlying missing values themselves will not add information, so should not be included in the imputation model (5). We therefore examine which of

these 26 variables that are predictive of non-response at age 55 are also predictive of income at age 55 (as the variable with the highest proportion of missing data), using the observed data.

```
. foreach var in aconnn512 aconnn504 i.acatnn660 bbinnn194 bconnage7dv10 genability11
  Ext11Dec18 dbinnn2250 dconnn1721 i.dcatnn2888 dconnn2930 Ext16MTDec18 i.ecatnn5113
  ebinnn5960 fbinn502977 fconnn504361_cont fbinnn504636_bin fbinntenure91_bin
  fconndvsoccapital_cont gbindmpart Org42 mconngenhlth i.hcatnnd7ms_cat iconnnd8nchtt_cont
  ibinn8j2101 NR08priorNR {
  regress log_total_income55 `var'
  testparm `var'
  }
```

We find that 21 of the 26 variables are associated with income at age 55 with $p < 0.001$ so are included as auxiliary variables in the imputation model. The $p < 0.001$ level is essentially arbitrary, but in this instance ensures that we have a reasonable number of this type of auxiliary variable in the imputation model. If you had a smaller pool of potential variables to choose from, then you may wish to be more liberal with your p-value cut-off to ensure that sufficient variables enter the imputation model; if you have a larger pool of potential variables to choose from, then you may wish to be even more stringent with your p-value cut-off to prevent so many auxiliary variables entering the model so as to make it unstable. The key here is to avoid including variables that are completely unassociated with the underlying values of the variable(s) subject to missingness (income in this instance); whether or not variables that are only weakly associated with the underlying values enter the imputation model should not have a substantial impact on the MI analysis. One may consider eschewing a p-value-based approach altogether and selecting variables from the pool of predictors of non-response based on the magnitude of their association with the underlying values (e.g. linear regression coefficient, odds ratio, risk ratio). Whilst this is straightforward for binary predictor variables, for continuous predictor variables the magnitude of the association will be scale-dependent, and for categorical variables there will be multiple estimated associations with magnitudes dependent on the choice of baseline category, so this type of approach requires careful consideration. Alternatively, a machine learning approach to variable selection, such as the lasso (18), could be considered.

Variables that are predictive of the underlying missing values but not predictive of the chance of missing values should also be included. Here we include two such variables: income at age 46 (total_income46) and income at age 50 (total_income50), both of which are again log-transformed prior to imputation (log_total_income46 and log_total_income50). It is perhaps worth noting that these variables, being observed subsequently to the exposure (age 42) and prior to the outcome (age 55), are potentially on the causal pathway (i.e. mediators) between the two. In an analysis where the objective was to estimate the total effect of exposure on outcome, such potential mediators should not be included in the analysis model. However, this is not a concern when constructing an imputation model; here the interest is in leveraging the statistical association with the outcome variable in order to improve the quality of the imputed values.

Developing the imputation model

Different imputation approaches are available in Stata (see `help mi impute`): monotone, chained or mvn. “Monotone” is a sequential approach using a monotone missing pattern; “chained” is a sequential approach using chained equations; “mvn” uses multivariate normal regression.

In the general setting (i.e. if missingness is known not to be monotone), we recommend imputation using chained equations, implemented using `mi impute chained` (see `help mi impute chained`). Imputation using chained equations fills in missing values in multiple variables iteratively by using chained equations, a sequence of univariate imputation models with fully conditional specification of prediction equations, accommodating arbitrary missing value patterns.

Consider the appropriate type of imputation model for each variable requiring imputation. Stata’s `mi impute chained` command can accommodate a variety of models including:

- Linear regression (`regress`) for continuous/linear variables
- Predictive mean matching (`pmm`) for continuous/linear variables
- Logistic regression (`logit`) for binary variables

- Ordinal logistic regression (`ologit`) for ordered categorical variables where the proportional odds assumption can be assumed to hold
- Multinomial logistic regression (`mlogit`) for unordered categorical variables (or ordered categorical variables where the proportional odds assumption cannot be assumed to hold)
- Poisson regression (`poisson`) for count variables.

Example

It is important to gain some familiarity with the data prior to undertaking any analysis, and MI analyses are no exception. Numerical summaries (e.g. `summarize`; see `help summarize`), tabulations (e.g. `tabulate`; see `help tabulate`) and graphical approaches (e.g. `histogram`; see `help histogram`) can all be used to examine the distributions of the variables to be included in the imputation model. Transformations may be considered for skewed continuous variables (as we have already applied to the heavily skewed income variables). However, there is evidence that MI is relatively robust to the assumption of normality if the amount of missing information is low (17).

The imputation model in our example contains 52 variables of different types, as follows:

	Continuous/linear (`regress`)	**Binary (`logistic`)**	**Unordered categorical (`mlogit`)**	**Total**
Variables in the substantive model	8	17	4	29
Variables predictive of both non-response and underlying missing values (auxiliary)	11	7	3	21
Variables predictive of underlying missing values only (auxiliary)	2	0	0	2

Note that our substantive model does not include any interactions. Even so, we could choose to include any interactions we thought relevant in the imputation model, though in this example analysis we have not done so.

Recoding variables

Binary variables to be modelled using logistic regression (`logit`) need to be coded so that they take values 0 and a non-zero integer (usually 1).

Multinomial logistic regression (`mlogit`) models used to model unordered categorical variables (or ordered categorical variables where the proportional odds assumptions cannot be assumed to hold) with many levels are often unstable and can prevent the imputation model from converging. It may therefore be advisable to collapse together some categories of such unordered categorical variables to form a variable with fewer levels. This can be done in an iterative manner – i) conduct the imputation (see later section) using the unordered categorical variable in its current form; ii) if the multinomial logistic regression imputation model relating to this variable fails to converge, then recode the variable to have fewer levels – repeated until convergence.

Example

In our example, the variable representing a cohort member's sex (n622) was originally coded 1 = "Male" and 2 = "Female", so was recoded to be 0 = "Male" and 1 = "Female".

There are seven unordered categorical variables (or ordered categorical variables where we did not want to make the proportional odds assumption), with between 3 and 5 categories per variable. We decided to retain these variables in their original form, conduct the imputation, then consider recoding them only if the multinomial logistic regression imputation model relating to one or more of these variables fails to converge.

Preparing the data for imputation

Prior to undertaking the imputation, the data need to be specified as an MI dataset using the `mi set` command (see `help mi set`). The data are given a specific style: wide, mlong, flong, or flongsep. We recommend using the wide style in most instances.

The variables should then be registered as being either imputed, passive or regular using the `mi register` command (see `help mi register`). "Imputed" variables are variables that have missing values and for which you will have imputations; "passive" variables are variables that are a function of imputed variables or of other passive variables; "regular" variables are variables that are neither imputed nor passive and that have the same values, whether missing or not, in all imputed datasets.

Once the data are registered, the variable _mi_miss indicates observations which are incomplete (as opposed to fully observed) across the registered variables. It is required to run analyses on the imputed data.

Example

We first `mi set` and then `mi register` the data:

```
. mi set wide

. mi register imputed ///
  log_total_income55 ParStat42R n622 LBW smpreg n553 bfever maw5 MotherNeverReads7 NoIntEdu
  DivBy7 SepMore1Month MumNotMarried SocialClassHusband DifficultiesFinancial HousingTenure_7
  DifficultiesHousing CogAbil7 enuresis7 MedExSum7 BMI7 PsychoMed a16_totalscore
  log_total_income42 EconAct42R Mal24Age42 NVQ42R lsiany2 dmdisab ///
  ///
  aconnn512 aconnn504 acatnn660 bbinnn194 bconnage7dv10 genability11 dbinnn2250
  dconnn1721 dcatnn2888 dconnn2930 Ext16MTDec18 ebinnn5960 fbinn502977 fconnn504361_cont
  fbinntenure91_bin fconndvsoccapital_cont Org42 mconngenhlth hcatnnd7ms_cat
  iconnnd8nchtt_cont ibinn8j2101 ///
  ///
  log_total_income46 log_total_income50
```

(In the above code a `///` has been used to insert a gap between the different types of variables in the imputation model, but this is not necessary.)

A new variable _mi_miss has now been created.

```
. tab _mi_miss

   _mi_miss |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        538        3.45        3.45
          1 |     15,075       96.55      100.00
------------+-----------------------------------
      Total |     15,613      100.00
```

This shows that only 538 cohort members are non-missing for all 51 variables in the imputation model.

Conducting the imputation

If using imputation by chained equations (`mi impute chained`; see `help mi impute chained`), the imputation model is specified by stating each model type (`regress`, `logit`, etc.) followed by the list of variables to be modelled using that model type. By default, all variables are included in each univariate imputation model (i.e. one as the outcome, all the others as the explanatory variables). When a variable is to be modelled by `ologit` or `mlogit`, Stata recognises this as a categorical variable and handles it as such when it appears as an explanatory variable in other univariate imputation models (i.e. as if the i. prefix had been specified in a standard regression model).

When modelling binary or categorical variables (i.e. when using `logit`, `ologit` or `mlogit`), difficulties can be encountered when certain combinations of explanatory variables lead to predicted outcome probabilities very close to 0 or 1 (so called "perfect prediction"). Augmented versions of these regressions, in which a few observations with small weights are added to the data during estimation to avoid perfect prediction, can be utilised when perfect prediction is detected through use of

the `augment` option. In many applications of MI such use of augmented regression is necessary to obtain successful model convergence.

As noted previously, the imputation model should include all the variables in the substantive model in the form in which they will enter the analysis model, including any interactions between two or more variables, to ensure that it is “congenial” or “consistent” with the analysis model. Deciding on the optimal approach for including interactions within the imputation model has generated much debate within the MI literature. The prevailing opinion at present is to treat interaction terms as “just another variable” – that is, generate a new variable representing your interaction term using the observed data (e.g. for a interaction between variables X1 and X2, X1X2 = X1 x X2), then include this variable in the imputation model in exactly the same way you would if it were any other variable (19). Whilst the resulting imputed interaction values will then not necessarily equal the product of the imputed individual variables within a given imputed dataset, this approach has been found to perform well in simulation studies. An alternative approach when at least one of the variables in the interaction is binary or categorical is to perform the imputation separately within strata of the data defined by the observed values of that variable. This can be achieved by using the `by(varlist)` option within `mi impute` (see `help mi impute`). For example, if variable X1 was binary (coded 0/1), then using the `by(X1)` option would cause the imputation to be performed separately in those with X1 = 0 and those with X1 = 1. The association between X2 and all other variables in the imputation model would then be allowed to differ depending on the value of X1, as required. However, since only a subset of the data are used, this approach can be susceptible to convergence issues, particularly with small overall sample sizes, complex imputation models, and/or large numbers of categories in the interaction variable.

The number of imputed datasets to be added is controlled by the `add(#)` option. How many imputations should you use? While a small number of imputations (say 5-20) may be sufficient for reliable estimation of point estimates in most situations, estimating 95% confidence intervals and p-values with little error requires a greater number of imputations (perhaps 100 or more) (5, 20). Such large numbers of imputations can be computationally time-consuming with large samples and/or large numbers of variables in the imputation model. We suggest that 50 imputations would

be sufficient in most situations. However, in order to ensure that the imputation procedure is working as intended, it may be sensible to initially run a “test” imputation of, say, 5 imputations.

It can be helpful for troubleshooting to review the output from all the fitted imputation models. The output can be produced by using the `noisily` option and saved in a log file.

Similarly, it can be helpful to examine the means and standard deviations of imputed values from each iteration of the imputation (the “trace data”). These data can be saved to a separate data file using the `savetrace(filename)` option.

The `rseed(#)` option allows you to specify the random-number seed of the imputation procedure, making the results exactly reproducible if run again on a different occasion.

The `burnin(#)` option specifies number of iterations for the burn-in period (how many times all the univariate imputation models are fitted before each set of imputed values is drawn). The default is 10, which is often sufficient, but may need to be increased in some cases (see later section on “Checking the imputed values”).

After conducting the MI, Stata creates a number of new variables with the prefixes _1_, _2_, … , _M_ (where M is the total number of imputed datasets), which are the variables containing the imputed values (in addition to the observed values) for each variable within each imputed dataset.

Example

We perform the imputation using `mi impute chained`.

```
. mi impute chained ///
  ///
  (logit, augment) n622 dmdisab lsiany2 SepMore1Month MotherNeverReads7 NoIntEdu smpreg maw5
  HousingTenure_7 DifficultiesHousing DifficultiesFinancial DivBy7 LBW bfever enuresis7
  MumNotMarried PsychoMed bbinnn194 dbinnn2250 ebinnn5960 fbinn502977 fbinntenure91_bin Org42
  ibinn8j2101 ///
  ///
  (mlogit, augment) NVQ42R SocialClassHusband ParStat42R EconAct42R acatnn660 dcatnn2888
  hcatnnd7ms_cat ///
```

```
///
(regress) log_total_income55 Mal24Age42 n553 BMI7 CogAbil7 MedExSum7 a16_totalscore
log_total_income42 aconnn512 aconnn504 bconnage7dv10 genability11 dconnn1721 dconnn2930
Ext16MTDec18 fconnn504361_cont fconndvsoccapital_cont mconngenhlth iconnnd8nchtt_cont
log_total_income46 log_total_income50 ///
///
, rseed(12345) dots noisily force add(5) savetrace(MI_test_trace, replace)
```

As noted previously, there are no interactions in this MI model.

Initially we produce just 5 imputed datasets (`add(5)`) to test that the imputation procedure is working correctly. Subsequently we would want to run the code again, increasing the number of imputations.

Note that the `burnin(#)` option has not been specified, so the default of 10 will be used.

In addition, you would probably want to open a log file prior to running the above code (`log using`; see `help log`) and close it afterwards (`log close`), as well as saving the imputed dataset at the end (`save`; see `help save`).

Troubleshooting the imputation

With only a small number of variables in an imputation model, the procedure will usually converge successfully with no problems. With a greater number of variables, convergence issues and other problems will often be encountered. Here, we describe some of the more common problems.

If the model fitting procedure continues to iterate without the log-likelihood being maximised (i.e. non-convergence), there is an underlying issue with the specification of the current univariate imputation model which needs to be rectified. As highlighted above, this is of particular relevance for multinomial logistic regression (`mlogit`) models used to model unordered categorical variables (or ordered categorical variables where the proportional odds assumptions cannot be assumed to hold) with many levels. In such cases, it may be advisable to collapse together some categories of the unordered categorical variable to form a variable with fewer levels.

The imputation procedure may cease prematurely with the error message “perfect predictor(s) detected”. As noted above, this “perfect prediction” can occur when modelling binary or categorical variables (i.e. when using logit, ologt or mlogit), and relates to certain combinations of explanatory variables leading to predicted outcome probabilities very close to 0 or 1. Augmented versions of these regressions, in which a few observations with small weights are added to the data during estimation to avoid perfect prediction, can be utilised when perfect prediction is detected through use of the `augment` option in `mi impute`. The above error message suggests that the `augment` option was not specified but should be.

A related warning message is that a certain expression "predicts success perfectly" and that a number of observations have been dropped. This warning message may be present even if the imputation procedure has completed successfully (which would require the specification of the `augment` option in `mi impute`). It suggests that there is an issue involving two categorical variables in the imputation model: for a certain level of a categorical variable (identified in the warning message), the outcome variable of the univariate imputation model always takes the same value. This can be straightforwardly diagnosed by cross-tabulating the two variables concerned and identifying the perfect prediction. It can be rectified by collapsing together categories of one of the variables until perfect prediction no longer occurs. If both variables concerned are binary then no such recoding is possible and one of the variables will need to be excluded from the imputation model.

Example

In our example analysis the MI procedure converged successfully and none of the above issues were apparent.

Checking the imputed values

Having conducted the imputation and saved the imputed datasets, it is important to check that the imputed values themselves appear sensible.

One approach is to plot the means and standard deviations of imputed values from each iteration of the imputation (“trace data”), saved as part of the imputation

procedure (see “Conducting the imputation” section). The dataset containing the means and standard deviations of imputed values should be opened and these values plotted against the iteration number for each imputed variable separately. A useful command for doing so is `xtline` (see `help xtline`) as it allows the values for each imputed dataset to be plotted in a different colour. Using these plots, one can examine whether the means and standard deviations of imputed values have sufficiently stabilised over the course of the iterations in the burn-in period. An underlying trend in the trace data is not in itself problematic, but if such a trend remains at the end of the burn-in period (and looks like it would continue in subsequent iterations) it suggests that the values had not sufficiently stabilised at the end of burn-in period when the values for each imputed dataset were drawn. In such situations the imputation model should be re-fitted with a greater number of iterations in the burn-in period using the `burnin(#)` option of `mi impute chained`.

It is also good practice to compare the distributions of variables between the observed data and the imputed datasets. There are many approaches one could use to achieve this, for example plotting histograms for continuous variables and comparing prevalences for binary/categorical variables. Substantial differences between the distributions of observed and imputed variables should be investigated further. However, differences themselves are not necessarily indicative of a problem – it may be that the underlying values of a variable *should* differ between individuals with observed data and individuals with missing data, and therefore the imputation procedure is providing appropriate imputations. Moreover, the whole rationale for MI is that we are in essence imputing a *distribution* of plausible values rather than a single value, and therefore a given imputed value (i.e. a given draw from this distribution) should not be over-interpreted in isolation.

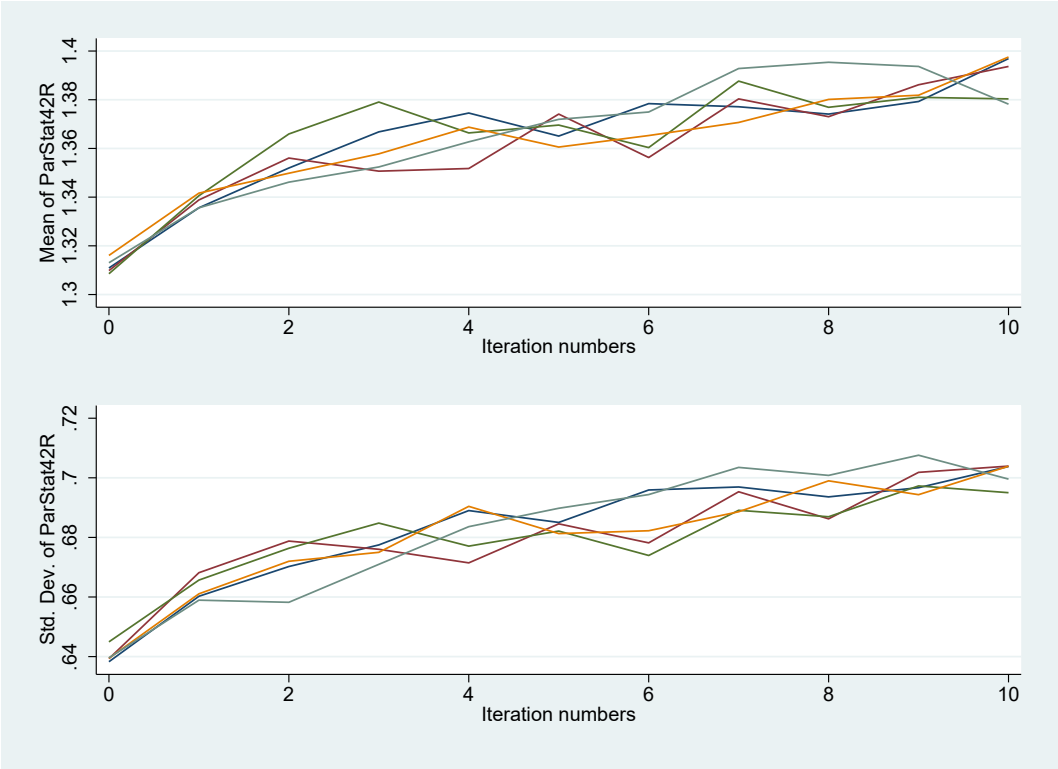
Example

The below code will loop through each variable in the imputation model in turn, producing plots of the means and standard deviations of imputed values from each iteration of the imputation against the imputation number (“trace plots”), with a different coloured line for each imputation. The `pause` command (see `help pause`) means that once each plot has been produced Stata will wait before producing the next plot (otherwise it would be produced before you had chance to examine the

previous one). To move to the next plot type “end” or “q”; to exit the plots completely type “BREAK”.

```
. pause on
. foreach var in log_total_income55 ParStat42R n622 n553 BMI7 CogAbil7 MedExSum7 ///
  a16_totalscore log_total_income42 dmdisab lsiany2 SepMore1Month MotherNeverReads7 NoIntEdu ///
  smpreg maw5 HousingTenure_7 DifficultiesHousing DifficultiesFinancial DivBy7 LBW ///
  bfever enuresis7 MumNotMarried PsychoMed NVQ42R SocialClassHusband Mal24Age42 EconAct42R ///
  ///
  aconnn512 aconnn504 acatnn660 bbinnn194 bconnage7dv10 genability11 dbinnn2250 dconnn1721 ///
  dcatnn2888 dconnn2930 Ext16MTDec18 ebinnn5960 fbinn502977 fconnn504361_cont ///
  fbinntenure91_bin fconndvsoccapital_cont Org42 mconngenhlth hcatnnd7ms_cat ///
  iconnnd8nchtt_cont ibinn8j2101 ///
  ///
  log_total_income46 log_total_income50 {
    xtline `var'_mean, t(iter) i(m) overlay legend(off) name(graph1, replace)
    xtline `var'_sd, t(iter) i(m) overlay legend(off) name(graph2, replace)
    graph combine graph1 graph2, xcommon cols(1)
    pause
  }
. pause off
```

For most variables, the means and standard deviations of imputed values appear to have sufficiently stabilised over the course of the iterations in the burn-in period. For a few variables there remains an underlying trend in the means and/or standard deviations at the end of the burn-in period, suggesting that the imputation model should be re-fitted with a greater number of iterations in the burn-in period. For example, ParStat42R:

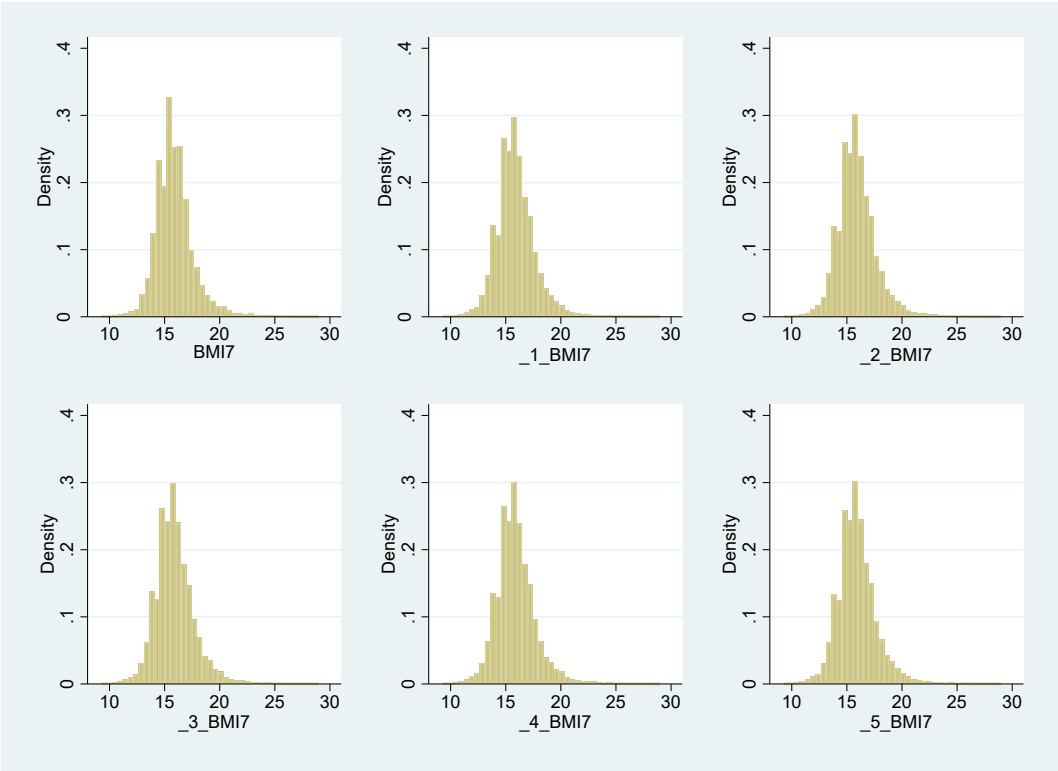


The below code will loop through each continuous variable in the imputation model in turn, producing histograms of the values in the observed data and then in each imputed dataset separately, before combining them at the end. The `pause` command is used as above.

```
. pause on
. foreach var of varlist log_total_income55 Mal24Age42 n553 BMI7 CogAbil7 MedExSum7 ///
  a16_totalscore log_total_income42 ///
  aconnn512 aconnn504 bconnage7dv10 genability11 dconnn1721 dconnn2930 Ext16MTDec18 ///
  fconnn504361_cont fconndvsoccapital_cont mconngenhlth iconnnd8nchtt_cont ///
  log_total_income46 log_total_income50 {
    hist `var', name(graph0, replace)
    hist _1_`var', name(graph1, replace)
    hist _2_`var', name(graph2, replace)
    hist _3_`var', name(graph3, replace)
    hist _4_`var', name(graph4, replace)
    hist _5_`var', name(graph5, replace)
    graph combine graph0 graph1 graph2 graph3 graph4 graph5, xcommon ycommon
    pause
  }
. pause off
```

The plots for BMI at age 7 (BMI7) show that the distributions of values in the imputed

datasets (note that this includes both imputed and observed values) are very similar to the distribution of values in the observed sample. However, as noted above, even if the distributions did differ somewhat, this is not necessarily a cause for concern, though substantial differences should be investigated further.



The below code will loop through each binary/categorical variable in the imputation model in turn, displaying a tabulation of the values in the observed data and then in each imputed dataset separately.

```
. foreach var of varlist n622 dmdisab lsiany2 SepMore1Month MotherNeverReads7 NoIntEdu ///
  smpreg maw5 HousingTenure_7 DifficultiesHousing DifficultiesFinancial DivBy7 LBW bfever ///
  enuresis7 MumNotMarried PsychoMed NVQ42R SocialClassHusband ParStat42R EconAct42R ///
  bbinnn194 dbinnn2250 ebinnn5960 fbinn502977 fbinntenure91_bin Org42 ibinn8j2101 acatnn660 ///
  dcatnn2888 hcatnnd7ms_cat {
  tab1 `var' _1_`var' _2_`var' _3_`var' _4_`var' _5_`var'
}
```

The tabulations for chronic illness (lsiany2) show a prevalence of 28.8% among the 10,976 cohort members with observed data on this variable, but a prevalence of 30.0% among the 15,613 cohort members alive and still living in Britain at Sweep 9 in the first imputed dataset (noting again that this includes both imputed and observed values). Prevalences in the other four imputed datasets (not shown in the interest of

space) ranged between 29.6% and 30.2%. Such small differences should not be unexpected – if we think that cohort members with chronic illness may be more likely to drop out of the study then the prevalence of chronic illness post-imputation should be somewhat greater than in the observed data. However, as noted previously, more substantial differences should be investigated further.

```
-> tabulation of lsiany2


   1st long |
   standing |
    illness |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |      7,820       71.25       71.25
        Yes |      3,156       28.75      100.00
------------+-----------------------------------
      Total |     10,976      100.00


-> tabulation of _1_lsiany2

 _1_lsiany2 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     10,926       69.98       69.98
          1 |      4,687       30.02      100.00
------------+-----------------------------------
      Total |     15,613      100.00
```

Once the above checks of the imputed values have been satisfactorily concluded using the test dataset of 5 imputations, the imputation model should be re-fitted using a greater number of imputations (here we use 50) and with a greater number of iterations in the burn-in period if this was deemed necessary (here we use 20). The checks of the imputed values should then be repeated on the new imputed dataset to ensure that everything now/still looks okay.

Fitting the analysis model

The analysis model can be fitted using `mi estimate` (see `help mi estimate`), which is followed by whatever command would usually (i.e. in the non-MI setting) be used to fit the analysis model. A useful option here is `dots`, which displays dots in the Stata results window as the estimations are performed within each imputed dataset (indicating that progress is being made, which otherwise would not be apparent). The output produced in the Stata results window in relation to the fitted model will be very similar to that produced by the same command in the non-MI setting, with some additional MI-specific information (number of imputations, etc.) at the top.

While there is consensus in the MI literature that imputation should be conducted using all individuals in the sample, there remains some debate around who should be included in the MI analysis model in the case where there is imputation of the outcome variable. If no auxiliary variables are included in the imputation model, then the only information being used in the imputation of the outcome variable is that contained within the analysis model and there is therefore no advantage (in terms of bias reduction) in including imputed values of the outcome in the MI analysis – indeed this may just add noise. This suggests an “impute and delete” approach whereby missing values of the outcome variable are imputed during the imputation phase but then deleted prior to (or at least excluded from) the analysis phase (21). On the other hand, if strong predictors of the underlying values of the outcome variable are included in the imputation model as auxiliary variables, then there is additional information in the imputation model beyond that in the analysis model, and the imputed values of the outcome variable should be retained in the MI analysis model. How strong do predictors of the underlying values of the outcome variable have to be for this to be the case? This will depend on the specifics of the analysis. A further consideration is that the greater the number of individuals included in an MI analysis, the more precise the estimates (in general). All other things being equal, one would therefore try to include as many individuals as possible in the MI analysis, but not (in general) at the expense of introducing (or limiting the reduction of) bias. Whilst these may be more statistical considerations, the choice of the MI analysis

sample has implications on the interpretation of any findings, as it contributes to the definition of the target population to which sample inferences are being made.

The issue of who to include in the MI analysis is therefore a little complex, and inevitably this will differ by context. One suggestion would be to perform the analysis using different analysis samples in order to explore the sensitivity of the findings to this issue.

Example

In our example we have imputed missing data using the full dataset of 15,613 cohort members who were alive and had not emigrated by age 55. Of these, 6306 cohort members had data on income at age 55, so we would certainly feel comfortable analysing these individuals. However, because we included auxiliary variables which we believe are strongly predictive of the outcome variable (income at earlier ages) we are happy to extend our MI analysis sample to include all 9137 Sweep 9 (age 55) respondents. All subsequent analyses will relate to this sample. (In fact, we did repeat the MI analysis using both the 6306 cohort members with observed income data at age 55 and all 15,613 cohort members who were alive and had not emigrated by age 55, and the results were very similar to those presented below.)

The MI analysis model is fitted using the same `regress` command as in the complete case analysis, but now preceded by the `mi estimate` command. The variable NR09 is an indicator variable for non-response at Sweep 9, so using the subsample with NR09 = 0 is just restricting the model fitting to the 9137 Sweep 9 (age 55) respondents. Note that we are now using the full imputed dataset with 50 imputations.

```
. mi estimate, dots: regress log_total_income55 i.ParStat42R n622 n553 BMI7 CogAbil7 MedExSum7
  a16_totalscore Log_total_income42 dmdisab lsiany2 SepMore1Month MotherNeverReads7 NoIntEdu
  smpreg maw5 HousingTenure_7 DifficultiesHousing DifficultiesFinancial DivBy7 LBW
  bfever enuresis7 MumNotMarried PsychoMed i.NVQ42R i.SocialClassHusband Mal24Age42
  i.EconAct42R if NR09==0

Imputations (50):
  .........10.........20.........30.........40.........50 done

Multiple-imputation estimates                   Imputations       =         50
```

```
Linear regression                               Number of obs     =      9,137
                                                Average RVI       =     0.6292
                                                Largest FMI       =     0.5454
                                                Complete DF       =       9099
DF adjustment:   Small sample                   DF:     min       =     161.58
                                                        avg       =     321.56
                                                        max       =     476.02
Model F test:       Equal FMI                   F(  37, 5173.5)   =      20.67
Within VCE type:          OLS                   Prob > F          =     0.0000

------------------------------------------------------------------------------
         log_total_income55 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------------------+-------------------------------------------------------------
                 ParStat42R |
Seperated/Divorced/Widowed  |  -.1442798    .049738    -2.90   0.004    -.2420863   -.0464733
       Single/NeverMarried  |  -.3962706   .0546227    -7.25   0.000    -.5036618   -.2888793

                                        [Output omitted]
------------------------------------------------------------------------------
```

The MI analysis used data from 9137 cohort members and estimates coefficients of -0.14 (95% CI -0.24, -0.05) comparing separated/divorced/widowed to married/cohabiting and -0.40 (95% CI -0.50, -0.29) comparing single and never married to married/cohabiting. These correspond to 13% (exp(-0.14) = 0.87) and 33% (exp(-0.40) = 0.67) lower income respectively.

Comparing the results from the complete case and MI analyses (see below table), we note a number of differences. In the MI analysis the coefficients differ, though only slightly, from those in the complete case analysis. If we believe that data are MAR then the MI analysis will (assuming we have correctly specified our imputation model) give us unbiased results, whereas the complete case analysis will not, and we should interpret the difference between the two sets of results as being suggestive of bias in the complete case analysis. The MI analysis includes a greater number of cohort members. These 9137 cohort members represent all Sweep 9 respondents, including the 1896 complete cases included in the complete case analysis, 4410 additional cohort members who had observed income at age 55 data but who were missing data on one or more other analysis variables, and a further 2831 cohort members who were missing income data at age 55 (and may or may not have also been

missing data on one or more other analysis variables). As well as the potential to reduce bias in the estimated coefficients, a consequence of utilising all this additional information is increased precision – the standard errors of the coefficients in the MI analysis are reduced by almost a half and the 95% CIs are therefore much narrower.

Analysis	**Category**	**Coefficient**	**Standard error**	**95% CI**
Complete case	Married/cohabiting	0.00		(ref)
(n = 1896)	Separated/divorced/widowed	-0.18	0.09	-0.36, 0.01
	Single and never married	-0.37	0.09	-0.55, -0.18
MI	Married/cohabiting	0.00		(ref)
(n = 9137)	Separated/divorced/widowed	-0.14	0.05	-0.24, -0.05
	Single and never married	-0.40	0.05	-0.51, -0.29

Checking the analysis model

It is good practice to perform some additional checks once the analysis model has been fitted. The `mi estimate` command can be reissued with the `vartable` and `dftable` options (but without specifying the analysis model) (see `help mi estimate`).

`vartable` displays a table reporting variance information about MI estimates. The table contains estimates of within-imputation variances, between-imputation variances, total variances, relative increases in variance due to nonresponse (RVI), fractions of information about parameter estimates missing due to nonresponse (FMI), and relative efficiencies for using the chosen number of imputations rather than a hypothetically infinite number of imputations. The RVI is the proportional increase in total sampling variance that is due to missing information. Variables with large amounts of missing data and/or that are weakly correlated with other variables in the imputation model will tend to have high RVIs. The closer this number is to zero, the less effect missing data have on the variance of the estimate. The FMI is the proportion of the total sampling variance that is due to missing data. The higher the

FMI is, the greater the number of imputations required for reliable results. One rule of thumb is to have the number imputations (at least) equal the highest FMI percentage. The relative efficiency is an estimate of the efficiency (how well the true population parameters are estimated) relative to performing an infinite number of imputations. If the relative efficiency is not close to 1, then it indicates that the analysis should be repeated with a greater number of imputations.

`dftable` displays a table containing parameter-specific degrees of freedom (DF) and percentages of increase in standard errors due to nonresponse. The parameter-specific degrees of freedom depend not only on the number of imputations but also (inversely) on the RVI due to nonresponse. The closer the RVI is to zero, the larger the degrees of freedom regardless of the number of imputations.

Example

The tables corresponding to the above fitted model are shown below (with some variables omitted for brevity). In the table produced by `vartable`, we see that for many variables the between-imputation variability is quite large relative to the within-imputation variability, leading to high RVI and FMI values. For example, a16_totalscore has a RVI of 1.17, indicating that much of the total variance is due to between-imputation variability, and a FMI of 0.55, indicating that much of the total sampling variance is due to missing data. This FMI value (which was also the highest of the omitted values) suggests that at least ~50 imputations are required for reliable results, and we also note that the relative efficiency is close to 1 for all variables, so we can be reasonably confident in the results using 50 imputations.

```
. mi estimate, vartable

Multiple-imputation estimates                   Imputations       =         50
Linear regression

Variance information
--------------------------------------------------------------------------------
              |        Imputation variance                             Relative
              |    Within   Between     Total       RVI       FMI    efficiency
--------------+-----------------------------------------------------------------
  ParStat42R  |
Seperated~d   |     .0016   .000857   .002474   .546406   .356685       .992917
Single/Ne~d   |   .001962   .001002   .002984   .520641    .34558       .993136
              |
        n622  |   .000733   .000344   .001084   .478842    .32674       .993508
        n553  |   4.3e-06   2.8e-06   7.1e-06   .658577      .401       .992044
        BMI7  |   .000042   .000022   .000064   .525621   .347755       .993093
    CogAbil7  |   .000299    .00024   .000544   .816857   .454179       .990998
   MedExSum7  |   .000035   .000019   .000054   .556507   .360938       .992833
a16_totals~e  |   .000012   .000014   .000026    1.1737   .545446       .989209

                                  [Output omitted]

--------------------------------------------------------------------------------
```

In the output produced by `dftable`, we see from the header information (which was also reported when `mi estimate` was initially run) that the parameter-specific degrees of freedom vary between 162 and 476, with a mean of 322. We see from the table output that the smallest degrees of freedom correspond to a16_totalscore, which is to be expected given that this variable has the largest RVI. The largest degrees of freedom correspond to MumNotMarried (omitted from the below output), suggesting that the loss of information due to non-response is the smallest for the estimation of this coefficient. Consequently, the percentage increase in standard error is largest for a16_totalscore (47.4%) and smallest for MumNotMarried (20.3%).

```
. mi estimate, dftable

Multiple-imputation estimates                   Imputations       =         50
Linear regression                               Number of obs     =      9,137
                                                Average RVI       =     0.6292
                                                Largest FMI       =     0.5454
                                                Complete DF       =       9099
```

```
DF adjustment:   Small sample                    DF:     min       =     161.58
                                                         avg       =     321.56
                                                         max       =     476.02
Model F test:       Equal FMI                    F(  37, 5173.5)    =      20.67
Within VCE type:          OLS                    Prob > F           =     0.0000

------------------------------------------------------------------------------------------
                            |                                                   % Increase
         log_total_income55 |      Coef.   Std. Err.      t    P>|t|           DF   Std. Err.
----------------------------+-------------------------------------------------------------
                 ParStat42R |
Seperated/Divorced/Widowed  |  -.1442798    .049738    -2.90   0.004        367.9       24.35
       Single/NeverMarried  |  -.3962706   .0546227    -7.25   0.000        390.7       23.31
                            |
                       n622 |   .0147763   .0329194     0.45   0.654        434.4       21.61
                       n553 |   .0018927   .0026708     0.71   0.479        294.1       28.79
                       BMI7 |  -.0076632   .0079808    -0.96   0.338        386.1       23.52
                   CogAbil7 |   .0638111   .0233175     2.74   0.007        231.2       34.79
                  MedExSum7 |   .0063789   .0073428     0.87   0.386        359.7       24.76
             a16_totalscore |  -.0027983   .0050665    -0.55   0.581        161.6       47.43

                                          [Output omitted]

------------------------------------------------------------------------------------------
```

Missing not at random sensitivity analyses

MI will provide unbiased results on the assumption that data are MAR. In practice, it is impossible to know that data truly are MAR (as opposed to MNAR) and therefore we might wish to explore how robust our results are to the MAR assumption. A variety of such “MNAR sensitivity analyses” have been proposed, which typically involve imputing data under a MNAR mechanism – or at least approximating the results of doing so (5). One simple approach to this is to take the existing multiply imputed (under MAR) datasets, modify the imputed values for one or more subsets of the sample according to a MNAR scenario of interest, and re-fit the analysis model. Formally, this is a “pattern-mixture model” approach to MNAR sensitivity analysis (5).

For example, you may imagine that individuals with lower income might be less likely to complete questions on income (even after taking all other available information into account). The resultant income data would then be MNAR, as the probability of

missingness would be affected by values of the partly-unobserved variable itself. A MI analysis of these data would only be unbiased under MAR, so a MNAR sensitivity analysis should be undertaken. Using the simple approach to MNAR sensitivity analysis mentioned above, hypothesised scenarios such as "study members with missing income data have *X* units lower income than otherwise similar study members with observed income data" can be explored by subtracting *X* units from the imputed income values in each imputed dataset and re-fitting the analysis model. Assuming that there was previously evidence of the association of interest, does this remain the case in the sensitivity analysis? Different values of *X* could be considered in a more thorough sensitivity analysis. At what value of *X* is there no longer evidence of the association of interest? How plausible is such a value of *X* in practice? The MNAR scenarios being considered could be more complex, for example the value of *X* may be hypothesised to differ between subgroups defined by another variable (e.g. males and females), with the imputed value modification approach extended in the obvious way.

Example

As an example, let's assume we believe that cohort members with lower income at age 55 are less likely to report their income than those with higher income, despite the richness of the information include in the imputation phase. First, let's explore the hypothesised scenario that "cohort members with missing income at age 55 have 10% lower income than otherwise similar cohort members with observed income data". Since we are analysing income at age 55 on the log scale, it is more straightforward to consider multiplicative (as opposed to additive) differences on the original scale – a 10% reduction on the original scale (i.e. multiplying by 0.9) is equivalent to an additive difference of log(0.9) = -0.105 on the log scale.

Having opened our multiply imputed dataset, let's first generate an indicator variable for whether log-income at age 55 is imputed (i.e. whether it is missing in the original sample).

```
. gen log_total_income55_imputed = (log_total_income55 >= .)
```

We now generate a new adjusted log-income variable (log_total_income55_adj) within each imputed dataset, which takes the observed value of log_total_income55

for cohort members in whom income was observed and the imputed value plus log(0.9) (i.e. the imputed value minus 0.105) for cohort members in whom income was not observed (i.e. in whom income was imputed). Note that we use the `mi passive` command as a prefix to the `generate` and `replace` commands. This ensures that the commands are applied within each imputed dataset and that the resultant variables are registered as passive variables (recall that "passive" variables are variables that are a function of imputed variables or of other passive variables; see `help mi passive`).

```
. mi passive: generate log_total_income55_adj = log_total_income55
. mi passive: replace log_total_income55_adj = log_total_income55_adj + log(0.9) if
  log_total_income55_imputed==1
```

We are now in a position to re-run the analysis using the new adjusted log-income variable in place of the original log-income variable. Again, we fit this model to all sweep 8 respondents (n = 9,137).

```
. mi estimate, dots: regress log_total_income55_adj i.ParStat42R n622 n553 BMI7 CogAbil7
  MedExSum7 a16_totalscore log_total_income42 dmdisab lsiany2 SepMore1Month MotherNeverReads7
  NoIntEdu smpreg maw5 HousingTenure_7 DifficultiesHousing DifficultiesFinancial DivBy7 LBW
  bfever enuresis7 MumNotMarried PsychoMed i.NVQ42R i.SocialClassHusband Mal24Age42 i.EconAct42R
  if NR09==0

Imputations (50):
  .........10.........20.........30.........40.........50 done

Multiple-imputation estimates                   Imputations       =         50
Linear regression                               Number of obs     =      9,137
                                                Average RVI       =     0.6276
                                                Largest FMI       =     0.5439
                                                Complete DF       =       9099
DF adjustment:   Small sample                   DF:     min       =     162.46
                                                        avg       =     322.63
                                                        max       =     473.89
Model F test:       Equal FMI                   F(  37, 5180.6)   =      20.86
Within VCE type:          OLS                   Prob > F          =     0.0000

------------------------------------------------------------------------------------------
     log_total_income55_adj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------------------+-------------------------------------------------------------
                  ParStat42R |
```

```
Seperated/Divorced/Widowed  |  -.1424689   .0497407    -2.86   0.004    -.2402772   -.0446606
       Single/NeverMarried  |  -.3949269   .0546427    -7.23   0.000    -.5023546   -.2874992

                                          [Output omitted]
------------------------------------------------------------------------------------------------
```

We see that there is very little change in the estimated coefficients relative to the primary analysis. This suggests that the findings are robust to the hypothesised scenario that “cohort members with missing income at age 55 have 10% lower income than otherwise similar cohort members with observed income data”. We could now continue to explore increasingly extreme hypothesised scenarios (20% lower income, 30% lower income, etc.) to see whether the findings continue to appear so robust.

Inverse probability weighting for missing data handling

This User Guide has so far focussed on MI as a principled method for missing data handling due to its applicability in the majority of typical analyses of data from CLS cohort studies and its ease of use with standard software. Other approaches are available, and here we will briefly discuss the application of inverse probability weighting (IPW) for missing data handling (6, 7).

As previously discussed, when study respondents are systematically different from non-respondents, a complete case analysis will often provide biased estimates. In IPW, the complete cases are weighted by the inverse of their probability of being a complete case. This means that study members who were unlikely to be a complete case (but were anyway) are up-weighted relative to cohort members who were likely to be complete cases (and were), so can be conceptualised as the former effectively representing both themselves and all other similar study members with missing data.

How should the probability of being a complete case be estimated? A common approach is to identify a set of variables which are predictive of being a complete case, which may or may not overlap with the set of variables in the analysis model, and fit a model for being a complete case as a function of these variables. From this fitted model, the probability of being a complete case can be predicted for each study member. This is a straightforward approach, but a major limitation is that if a study member is missing data on any one or more of the variables used to model being a complete case, then they will not be included in the sample to which this model is fitted and will not have a predicted probability of being a complete case, meaning that they cannot be included in the reweighted analysis model. It is possible to avoid this by choosing variables which are themselves completely observed as predictors of being a complete case. In the birth cohort setting this often means restricting to variables observed at or prior to birth. But what if you believe it to be important to include such partially observed variables as predictors of being a complete case? Methods exist for incorporating predictors subject to monotone missingness and more advanced methods can even handle non-monotone missingness (7).

A straightforward approach that can handle arbitrary missing data patterns would be to use MI in the estimation of the probabilities of being a complete case. Values of

the variables used to model being a complete case would first be imputed, so that the models for being a complete case are fitted using all study members, meaning all study members have a predicted probability of being a complete case, so can be included in the reweighted analysis model.

In the above described approach, IPW utilises the probability of being a complete case. But whether or not a given study member is a complete case will usually depend on the analysis model being considered, and therefore any derived probabilities are not transportable between analyses, meaning that the probability-derivation procedure must be repeated on an analysis-specific basis.

In cohort studies, where attrition is the main driver of missing data, it is often sufficient to consider response at the survey sweep where the outcome variable was collected as a proxy for being a complete case. Non-responders at this sweep will by definition not be complete cases, and responders at this sweep will only not be complete cases in the presence of non-monotone attrition (i.e. non-response at earlier sweeps causing missingness on analysis variables) or item non-response (on any of the analysis variables).

An alternative approach would therefore be to conduct a non-analysis-specific probability-derivation procedure considering response at the survey sweep where the outcome variable was collected, rather than being a complete case per se. This would still require using MI in the estimation of the probabilities of response to ensure that all study members have a predicted probability. The imputation model should include predictors of response at the survey sweep where the outcome variable was collected, which are exactly the same predictors of non-response that we have been considering previously (and which are provided for NCDS, BCS70, Next Steps, and MCS in the Appendices).

Example

There are 1896 cohort members in the complete case analysis. These are the cohort members who would be reweighted in an IPW analysis in order to regain sample representativeness.

To maintain comparability with our previous MI analyses, we will consider the 21 predictors of non-response at age 55 that are also associated with income at age 55 as the predictors of being a complete case. As noted above, in the birth cohort setting interest is often restricted to variables observed at or prior to birth to ensure that the extent of missing data is minimised, so the conclusions from applying this approach should not be assumed to apply to a more typical application of IPW.

First, we generate a new variable (cc) which is an indicator for being a complete case in the analysis model.

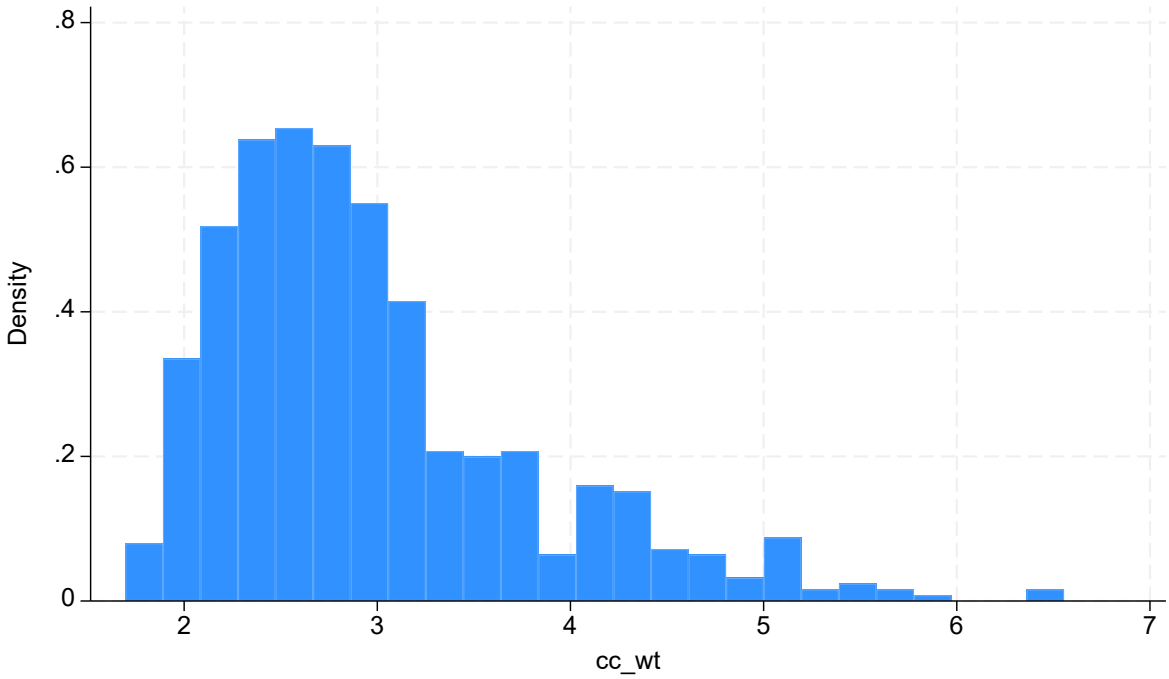
```
. quietly regress log_total_income55 i.ParStat42R n622 n553 BMI7 CogAbil7 MedExSum7
  a16_totalscore  Log_total_income42 dmdisab lsiany2 SepMore1Month MotherNeverReads7 NoIntEdu
  smpreg maw5 HousingTenure_7 DifficultiesHousing DifficultiesFinancial DivBy7 LBW bfever
  enuresis7 MumNotMarried PsychoMed i.NVQ42R i.SocialClassHusband Mal24Age42 i.EconAct42R
. gen cc = (e(sample)==1)
```

Then we fit a logistic regression model for being a complete case in the analysis model as a function of the 21 predictors of non-response at age 55. We restrict this model to the 9137 Sweep 9 (age 55) respondents, so that the resultant probabilities of being a complete case used in the IPW analysis relate to this sub-sample, similarly to the previous MI analysis.

```
. logit cc aconnn512 aconnn504 i.acatnn660 bbinnn194 bconnage7dv10 genability11 dbinnn2250 ///
  dconnn1721 i.dcatnn2888 dconnn2930 Ext16MTDec18 ebinnn5960 fbinn502977 fconnn504361_cont ///
  fbinntenure91_bin fconndvsoccapital_cont Org42 mconngenhlth i.hcatnnd7ms_cat ///
  iconnnd8nchtt_cont ibinn8j2101 if NR09==0

Logistic regression                               Number of obs     =      1,909
                                                  LR chi2(29)       =      59.83
                                                  Prob > chi2       =     0.0006
Log likelihood = -1191.1104                       Pseudo R2         =     0.0245

------------------------------------------------------------------------------------------
                            cc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------+----------------------------------------------------------
                     aconnn512 |   .0457647   .0651298     0.70   0.482    -.0818875    .1734168
                     aconnn504 |  -.0308995   .0404437    -0.76   0.445    -.1101676    .0483687

                                          Output omitted]

------------------------------------------------------------------------------------------
```

Note that this model was only fitted to 1909 cohort members (those with complete data on all variables in the model) out of the 9137 Sweep 9 (age 55) respondents. We now derive the probabilities of being a complete case according to the above model for cohort members included in the complete case analysis, then generate the weights as the reciprocal of the probabilities. It is good practice to examine the distribution of weights (here we restrict to cohort members included in the complete case analysis to whom the weights will be applied) to ensure there are no extremely large values which would dominate the IPW analysis, resulting in a large reduction in effective sample size (7).

```
. predict cc_prob
. gen cc_wt = 1/cc_prob
. sum cc_wt if cc==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       cc_wt |        645    2.967592    .8079965   1.695903   6.555767

. hist cc_wt
```



The resultant weights vary between 1.7 and 6.6, with no particularly extreme values. However, the weights are only defined for 645 out of the 1896 cohort members in the

complete case analysis. These are the cohort members with complete data on the 21 predictors of non-response at age 55 used to model being a complete case. This means that the IPW analysis model will only be fitted on these 645 cohort members.

```
. regress log_total_income55 i.ParStat42R n622 n553 BMI7 CogAbil7 MedExSum7 a16_totalscore ///
  Log_total_income42 dmdisab lsiany2 SepMore1Month MotherNeverReads7 NoIntEdu smpreg maw5 ///
  HousingTenure_7 DifficultiesHousing DifficultiesFinancial DivBy7 LBW bfever enuresis7 ///
  MumNotMarried PsychoMed i.NVQ42R i.SocialClassHusband Mal24Age42 i.EconAct42R ///
  [pweight = cc_wt]
(sum of wgt is 1,914.0969414711)

Linear regression                               Number of obs     =        645
                                                F(37, 607)        =       2.16
                                                Prob > F          =     0.0001
                                                R-squared         =     0.1357
                                                Root MSE          =     1.2082

------------------------------------------------------------------------------------------
                            |               Robust
         log_total_income55 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------------------+-------------------------------------------------------------
                 ParStat42R |
Seperated/Divorced/Widowed  |   -.173244   .1599381    -1.08   0.279    -.4873433    .1408552
       Single/NeverMarried  |  -.6940282   .2688696    -2.58   0.010    -1.222056   -.1660007

                                        [Output omitted]

------------------------------------------------------------------------------------------
```

As expected, the IPW analysis model is only fitted on 645 cohort members. The estimated coefficients of -0.17 (95% CI -0.49, 0.14) comparing separated/divorced/widowed to married/cohabiting and -0.69 (95% CI -1.22, -0.17) comparing single and never married to married/cohabiting correspond to 16% (exp(-0.17) = 0.84) and 50% (exp(-0.69) = 0.50) lower income respectively.

This analysis is inefficient because not all Sweep 9 (age 55) respondents (only 1909 out of 9137) were used in the derivation of the weights and not all cohort members included in the complete case analysis (only 645 out of 1896) were included in the IPW analysis. The reason for both these features is missing data on the variables

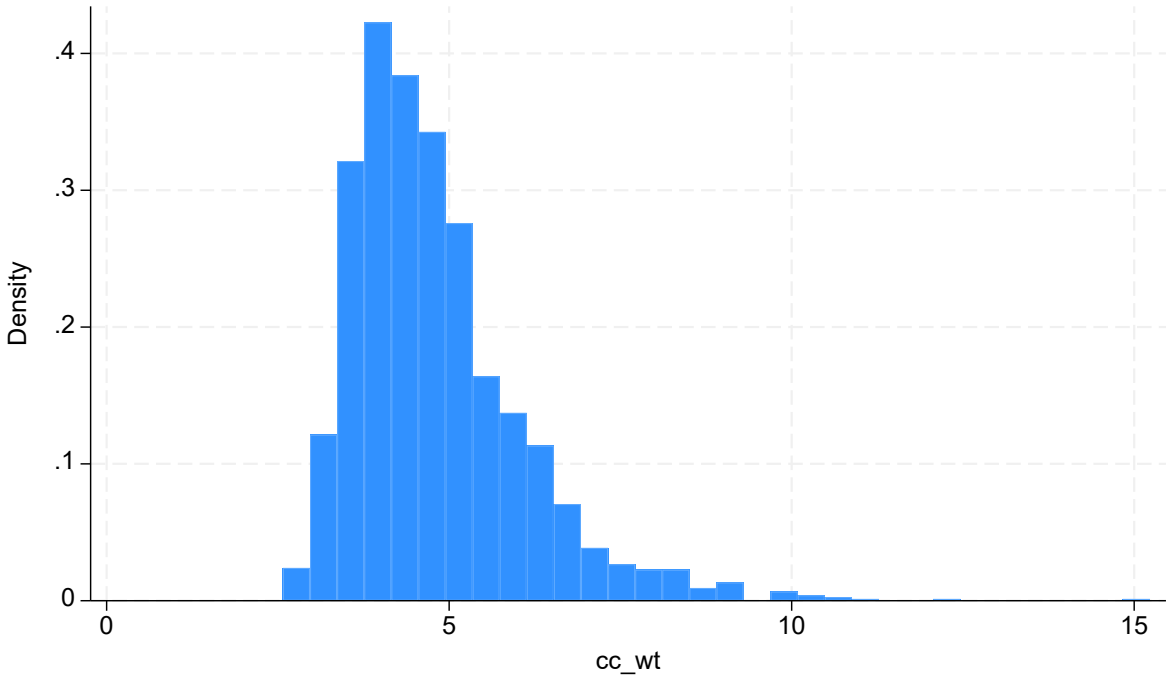
used as predictors of being a complete case (here, the 21 predictors of non-response at age 55).

Alternatively, we could conduct the IPW analysis having first multiply imputed any missing values to ensure that there will be no such problems. To illustrate this, we use the previously obtained MI dataset (recall that the imputation model contained 52 variables: 29 in the substantive model, 21 predictive of both non-response and underlying missing values and 2 predictive of underlying missing values only). We again fit a logistic regression model for being a complete case in the analysis model as a function of the 21 predictors of non-response at age 55, restricted to the 9137 Sweep 9 (age 55) respondents.

```
mi estimate, saving(mi_ipw_estfile, replace) dots: logit cc aconnn512 aconnn504 i.acatnn660 ///
  bbinnn194 bconnage7dv10 genability11 dbinnn2250 dconnn1721 i.dcatnn2888 dconnn2930 ///
  Ext16MTDec18 ebinnn5960 fbinn502977 fconnn504361_cont fbinntenure91_bin ///
  fconndvsoccapital_cont Org42 mconngenhlth i.hcatnnd7ms_cat iconnnd8nchtt_cont ibinn8j2101 ///
  if NR09==0

Imputations (50):
  .........10.........20.........30.........40.........50 done

Multiple-imputation estimates                   Imputations       =         50
Logistic regression                             Number of obs     =      9,137
                                                Average RVI       =     0.1124
                                                Largest FMI       =     0.2281
DF adjustment:   Large sample                   DF:     min       =     954.87
                                                        avg       =  13,493.41
                                                        max       =  95,278.78
Model F test:       Equal FMI                   F(  29,130353.5)  =       4.53
Within VCE type:          OIM                   Prob > F          =     0.0000

------------------------------------------------------------------------------------------
                           cc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------------+-----------------------------------------------------------
                     aconnn512 |  -.0501063   .0339963    -1.47   0.141    -.1167399    .0165274
                     aconnn504 |  -.0018362    .020436    -0.09   0.928    -.0418906    .0382182

                                          [Output omitted]

------------------------------------------------------------------------------------------
```

We see that this model was fitted to all 9137 Sweep 9 (age 55) respondents. The `saving(mi_ipw_estfile)` option is necessary here as this file is required by the subsequent `mi predict` command. We again derive the probabilities of being a complete case according to the above model, generate the weights as the reciprocal of the probabilities, and examine the distribution of weights for cohort members included in the complete case analysis.

```
. mi predict xb using mi_ipw_estfile
. mi passive: gen cc_prob = invlogit(xb)
. mi passive: gen cc_wt = 1/cc_prob
. sum cc_wt if cc==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       cc_wt |      1,896    4.799125    1.259972   2.573817   15.23007

. hist cc_wt if cc==1
```



The resultant weights vary between 2.6 and 15.2, with some slightly more extreme values (we could consider truncating these larger values at a certain threshold, e.g. 10, but have not done so here). The weights are now defined for all 1896 cohort

members in the complete case analysis, meaning that they will all be included in the IPW analysis model.

```
. regress log_total_income55 i.ParStat42R n622 n553 BMI7 CogAbil7 MedExSum7 a16_totalscore ///
  Log_total_income42 dmdisab lsiany2 SepMore1Month MotherNeverReads7 NoIntEdu smpreg maw5 ///
  HousingTenure_7 DifficultiesHousing DifficultiesFinancial DivBy7 LBW bfever enuresis7 ///
  MumNotMarried PsychoMed i.NVQ42R i.SocialClassHusband Mal24Age42 i.EconAct42R ///
  [pweight = cc_wt]
(sum of wgt is 9,099.14102745056)

Linear regression                               Number of obs     =      1,896
                                                F(37, 1858)       =       6.22
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1022
                                                Root MSE          =     1.1211

------------------------------------------------------------------------------------------
                            |               Robust
         log_total_income55 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------------------+-------------------------------------------------------------
                 ParStat42R |
Seperated/Divorced/Widowed  |   -.146184   .1005034    -1.45   0.146    -.3432955    .0509274
       Single/NeverMarried  |  -.3128876   .0916653    -3.41   0.001    -.4926654   -.1331097

                                        [Output omitted]

------------------------------------------------------------------------------------------
```

The estimated coefficients of -0.15 (95% CI -0.34, 0.05) comparing separated/divorced/widowed to married/cohabiting and -0.31 (95% CI -0.49, -0.13) comparing single and never married to married/cohabiting correspond to 14% (exp(-0.15) = 0.86) and 27% (exp(-0.31) = 0.73) lower income respectively.

These estimated coefficients differ markedly from the previous IPW analysis, particularly for single and never married (-0.31 vs. -0.69). Given that this is a weighted analysis of all 1896 cohort members included in the complete case analysis rather than just the 645 cohort members with complete data on the 21 predictors of non-response at age 55 used to model being a complete case, this should perhaps not be too surprising.

The MI IPW analysis is also more efficient than the previous IPW analysis because, in addition to all cohort members included in the complete case analysis being included in the IPW analysis, all Sweep 9 (age 55) respondents were used in the derivation of the weights. This is illustrated by the standard errors in the below table (which also includes the previous MI results): those for the first IPW analysis (IPW-CC) are much larger than those for the second IPW analysis (IPW-MI).

Analysis	**Category**	**Coefficient**	**Standard error**	**95% CI**
Complete case	Married/cohabiting	0.00		(ref)
(n = 1896)	Separated/divorced/widowed	-0.18	0.09	-0.36, 0.01
	Single and never married	-0.37	0.09	-0.55, -0.18
MI	Married/cohabiting	0.00		(ref)
(n = 9137)	Separated/divorced/widowed	-0.14	0.05	-0.24, -0.05
	Single and never married	-0.40	0.06	-0.50, -0.29
IPW-CC	Married/cohabiting	0.00		(ref)
(n = 645 -> 9137)	Separated/divorced/widowed	-0.17	0.16	-0.49, 0.14
	Single and never married	-0.69	0.27	-1.22, -0.17
IPW-MI	Married/cohabiting	0.00		(ref)
(n = 1896 -> 9137)	Separated/divorced/widowed	-0.15	0.10	-0.34, 0.05
	Single and never married	-0.31	0.09	-0.49, -0.13

Comparing the IPW results to the complete case and MI results we see that the estimated single and never married IPW-CC coefficients differ markedly from the complete case/MI coefficients, but the estimated IPW-MI coefficients are somewhat closer to the complete case analysis/MI estimates. The IPW-CC standard errors are larger than under any other approach, the MI standard errors are the smallest, and those for the complete case analysis and IPW-CC are in between and comparable to one another. The IPW-MI analysis therefore does not display the efficiency gains seen in the MI analysis relative to the complete case analysis, an observation which is true in general (7).

Finally, to emphasise a point made earlier, our use of predictors of being a complete case from across several sweeps of data collection which are consequently subject to high levels of missingness is not typical. The conclusions from applying this approach should therefore not be assumed to necessarily apply to more standard applications of IPW, which would typically rely on more completely observed variables observed at or prior to birth. However, if variables observed at later sweeps of data collection are important predictors of being a complete case then their inclusion should be considered, be that via an IPW-based approach or MI.

Further reading

There are many journal articles and books devoted to the handling of missing data generally and to specific approaches for doing so. The following are a selection that we recommend for further reading (in alphabetical order within section):

Missing data

- Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. Stat Methods Med Res. 2012; 21(3): 243-56
- Enders CK. Applied missing data analysis. New York: Guilford; 2010.
- Little RJA, Rubin DB. Statistical Analysis with Missing Data. Third Edition. Hoboken, NJ: John Wiley & Sons; 2020.

MI

- Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011; 20(1): 40-9.
- Carpenter JR, Bartlett JW, Morris TP, Wood AM, Quartagno M, Kenward MG. Multiple Imputation and its Application. Second edition. Chichester, UK: John Wiley & Sons, Ltd; 2023.
- Harel O, Mitchell EM, Perkins NJ, Cole SR, Tchetgen Tchetgen EJ, Sun B, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. Am J Epidemiol. 2018; 187(3): 576-84.
- Spratt M, Carpenter J, Sterne JAC, Carlin JB, Heron J, Henderson J, et al. Strategies for Multiple Imputation in Longitudinal Studies. American Journal of Epidemiology. 2010; 172(4): 478-87.
- Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009; 338: b2393.
- van Buuren S. Flexible Imputation of Missing Data. Second Edition. Boca Raton, FL: CRC Press; 2018.
- White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011; 30(4): 377-99.

IPW for missing data

- Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013; 22(3): 278-95.

- Sun B, Perkins NJ, Cole SR, Harel O, Mitchell EM, Schisterman EF, et al. Inverse-Probability-Weighted Estimation for Monotone and Nonmonotone Missing Data. Am J Epidemiol. 2018; 187(3): 585-91.

References

1. Rubin DB. Inference and missing data. Biometrika. 1976;63:581–92.
2. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Third Edition. Hoboken, NJ: Wiley; 2020.
3. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine. 2010;29(28):2920-31.
4. Little RJA, Rubin DB. The Analysis of Social-Science Data with Missing Values. Sociological Methods & Research. 1989;18(2-3):292-326.
5. Carpenter JR, Kenward MG. Multiple Imputation and Its Application2013.
6. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. Journal of Econometrics. 2007;141(2):1281-301.
7. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278-95.
8. Enders CK. Applied missing data analysis. New York: Guilford; 2010.
9. Enders CK. The performance of the full information maximum likelihood estimator in multiple regression models with missing data. Educational and Psychological Measurement. 2001;61(5):713-40.
10. Mostafa T, Narayanan M, Pongiglione B, Dodgeon B, Goodman A, Silverwood RJ, Ploubidis GB. Missing at random assumption made more plausible: evidence from the 1958 British birth cohort. J Clin Epidemiol. 2021;136:44-54.
11. Katsoulis M, Narayanan M, Dodgeon B, Ploubidis G, Silverwood R. A data driven approach to address missing data in the 1970 British birth cohort. BMC Medical Research Methodology. in press.
12. Silverwood RJ, Calderwood L, Henderson M, Sakshaug JW, Ploubidis GB. A data-driven approach to understanding non-response and restoring sample representativeness in the UK Next Steps cohort. Longitudinal and Life Course Studies. 2024;15(2):227-50.
13. Narayanan M, Villadsen A, Katsoulis M, Dodgeon B, Ploubidis G, Fitzsimons E, Silverwood R. A data driven approach to handling missing data in the UK Millennium Cohort Study. SocArXiv. 2026; https://doi.org/10.31235/osf.io/agqmu_v1.
14. Carpenter JR, Goldstein H, Kenward MG. REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. 2011. 2011;45(5):14.
15. Quartagno M, Grund S, Carpenter J. jomo: A Flexible Package for Two-level Joint Modelling Multiple Imputation. The R Journal. In press.
16. Cohort and Longitudinal Studies Enhancement Resources. Harmonised Earnings and Income in Four Longitudinal Cohort Studies: National Survey of Health and Development, National Child Development Study, 1970 British Cohort Study and Millennium Cohort Study [data collection]Datasets in preparation.
17. Harel O, Mitchell EM, Perkins NJ, Cole SR, Tchetgen Tchetgen EJ, Sun B, Schisterman EF. Multiple Imputation for Incomplete Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):576-84.
18. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.
19. Wagstaff DA, Kranz S, Harel O. A preliminary study of active compared with passive imputation of missing body mass index values among non-Hispanic white youths. Am J Clin Nutr. 2009;89(4):1025-30.
20. Graham JW, Olchowski AE, Gilreath TD. How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. Prevention Science. 2007;8(3):206-13.
21. Von Hippel PT. Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. Sociological Methodology. 2007;37(1):83-117.

22. University of London, Institute of Education, Centre for Longitudinal Studies. National Child Development Study: Childhood Data from Birth to Age 16, Sweeps 0-3, 1958-1974. [data collection]. 3rd Edition. National Children's Bureau, National Birthday Trust Fund, [original data producer(s)]. National Children's Bureau. SN: 5565, DOI: http://doi.org/10.5255/UKDA-SN-5565-22023.
23. University of London, Institute of Education, Centre for Longitudinal Studies. National Child Development Study: Age 23, Sweep 4, 1981, and Public Examination Results, 1978. [data collection]. 2nd Edition. National Children's Bureau, [original data producer(s)]. National Children's Bureau. SN: 5566, DOI: http://doi.org/10.5255/UKDA-SN-5566-12023.
24. University of London, Institute of Education, Centre for Longitudinal Studies. National Child Development Study: Age 33, Sweep 5, 1991. [data collection]. 2nd Edition. City University, Social Statistics Research Unit, [original data producer(s)]. City University, Social Statistics Research Unit. SN: 5567, DOI: http://doi.org/10.5255/UKDA-SN-5567-12023.
25. University of London, Institute of Education, Centre for Longitudinal Studies. National Child Development Study: Age 42, Sweep 6, 1999-2000. [data collection]. 2nd Edition. Joint Centre for Longitudinal Research, [original data producer(s)]. Joint Centre for Longitudinal Research. SN: 5578, DOI: http://doi.org/10.5255/UKDA-SN-5578-12023.
26. University of London, Institute of Education, Centre for Longitudinal Studies, NatCen Social Research. National Child Development Study: Biomedical Survey 2002-2004. [data collection]. UK Data Service. SN: 8731, DOI: http://doi.org/10.5255/UKDA-SN-8731-12023.
27. University of London, Institute of Education, Centre for Longitudinal Studies. National Child Development Study: Age 46, Sweep 7, 2004-2005. [data collection]. 3rd Edition. UK Data Service. SN: 5579, DOI: http://doi.org/10.5255/UKDA-SN-5579-12023.
28. University of London, Institute of Education, Centre for Longitudinal Studies. National Child Development Study: Age 50, Sweep 8, 2008-2009. [data collection]. 3rd Edition. UK Data Service. SN: 6137, DOI: http://doi.org/10.5255/UKDA-SN-6137-22023.
29. Chamberlain G, Chamberlain R, University of London, Institute of Education, Centre for Longitudinal Studies. 1970 British Cohort Study: Birth and 22-Month Subsample, 1970-1972. [data collection]. 3rd Edition. UK Data Service. SN: 2666, DOI: http://doi.org/10.5255/UKDA-SN-2666-22023.
30. University of London, Institute of Education, Centre for Longitudinal Studies, Osborn A, Dowling S, Butler N. 1970 British Cohort Study: Age 5, Sweep 2 1975. [data collection]. 5th Edition. UK Data Service. SN: 2699, DOI: http://doi.org/10.5255/UKDA-SN-2699-42023.
31. University of London, Institute of Education, Centre for Longitudinal Studies, Butler N, Bynner J. 1970 British Cohort Study: Age 10, Sweep 3, 1980. [data collection]. 7th Edition. UK Data Service. SN: 3723, DOI: http://doi.org/10.5255/UKDA-SN-3723-82023.
32. Bynner J, Butler N, University of London, Institute of Education, Centre for Longitudinal Studies. 1970 British Cohort Study: Age 16, Sweep 4, 1986. [data collection]. 9th Edition. UK Data Service. SN: 3535, DOI: http://doi.org/10.5255/UKDA-SN-3535-62023.
33. University of London, Institute of Education, Centre for Longitudinal Studies, Bynner J. 1970 British Cohort Study: Age 26, Sweep 5, 1996. [data collection]. 5th Edition. UK Data Service. SN: 3833, DOI: http://doi.org/10.5255/UKDA-SN-3833-32023.
34. University of London, Institute of Education, Centre for Longitudinal Studies. 1970 British Cohort Study: Age 29, Sweep 6, 1999-2000. [data collection]. 4th Edition. Joint Centre for Longitudinal Research, [original data producer(s)]. Joint Centre for Longitudinal Research. SN: 5558, DOI: http://doi.org/10.5255/UKDA-SN-5558-32023.
35. University of London, Institute of Education, Centre for Longitudinal Studies. 1970 British Cohort Study: Age 34, Sweep 7, 2004-2005. [data collection]. 5th Edition. UK Data Service. SN: 5585, DOI: http://doi.org/10.5255/UKDA-SN-5585-42023.
36. University of London, Institute of Education, Centre for Longitudinal Studies. 1970 British Cohort Study: Age 38, Sweep 8, 2008-2009. [data collection]. 5th Edition. UK Data Service. SN: 6557, DOI: http://doi.org/10.5255/UKDA-SN-6557-42023.

37. University of London, Institute of Education, Centre for Longitudinal Studies. 1970 British Cohort Study: Age 42, Sweep 9, 2012. [data collection]. 3rd Edition. UK Data Service. SN: 7473, DOI: http://doi.org/10.5255/UKDA-SN-7473-32023.
38. University College London, UCL Institute of Education, Centre for Longitudinal Studies. Next Steps: Sweeps 1-8, 2004-2016. [data collection]. 16th Edition. UK Data Service. SN: 5545, DOI: http://doi.org/10.5255/UKDA-SN-5545-82024.
39. University of London, Institute of Education, Centre for Longitudinal Studies. Millennium Cohort Study: Age 9 months, Sweep 1, 2001. [data collection]. 14th Edition. UK Data Service. SN: 4683, DOI: http://doi.org/10.5255/UKDA-SN-4683-62024.
40. University of London, Institute of Education, Centre for Longitudinal Studies. Millennium Cohort Study: Age 3, Sweep 2, 2004. [data collection]. 12th Edition. UK Data Service. SN: 5350, DOI: http://doi.org/10.5255/UKDA-SN-5350-72024.
41. University of London, Institute of Education, Centre for Longitudinal Studies. Millennium Cohort Study: Age 5, Sweep 3, 2006. [data collection]. 9th Edition. UK Data Service. SN: 5795, DOI: http://doi.org/10.5255/UKDA-SN-5795-62024.
42. University of London, Institute of Education, Centre for Longitudinal Studies. Millennium Cohort Study: Age 7, Sweep 4, 2008. [data collection]. 9th Edition. UK Data Service. SN: 6411, DOI: http://doi.org/10.5255/UKDA-SN-6411-92024.
43. University of London, Institute of Education, Centre for Longitudinal Studies. Millennium Cohort Study: Age 11, Sweep 5, 2012. [data collection]. 6th Edition. UK Data Service. SN: 7464, DOI: http://doi.org/10.5255/UKDA-SN-7464-62024.
44. University of London, Institute of Education, Centre for Longitudinal Studies. Millennium Cohort Study: Age 14 S, 2015. [data collection]. 7th Edition. UK Data Service. SN: 8156, DOI: http://doi.org/10.5255/UKDA-SN-8156-7, editor2024.
45. University of London, Institute of Education, Centre for Longitudinal Studies. Millennium Cohort Study: Age 17, Sweep 7, 2018. [data collection]. 2nd Edition. UK Data Service. SN: 8682, DOI: http://doi.org/10.5255/UKDA-SN-8682-22024.

Appendices

Note

The variables described in the Appendices are the versions that we used in our systematic data-driven approach to identify predictors of non-response in the CLS cohort studies. In some instances, the variables are identical to those available in the datasets available to researchers, but in many cases some recoding or further derivation was employed. This was, firstly, to ensure that different types of missing data were appropriately handled and, secondly, to try and avoid issues of data sparsity while maintaining sufficient information in the variables. Whilst these versions of the variables were appropriate for our analyses, exactly how variables are recoded/derived should be considered on an analysis-specific basis – it may be that alternative approaches are more appropriate in some cases.

Appendix A: Identified predictors of non-response in the 1958 National Child Development Study (NCDS).

Sweep 0 (age 0) to Sweep 8 (age 50) NCDS data are available from the UK Data Service:

National Child Development Study: Childhood Data from Birth to Age 16, Sweeps 0-3, 1958-1974 (22)

National Child Development Study: Age 23, Sweep 4, 1981, and Public Examination Results, 1978 (23)

National Child Development Study: Age 33, Sweep 5, 1991 (24)

National Child Development Study: Age 42, Sweep 6, 1999-2000 (25)

National Child Development Study: Biomedical Survey 2002-2004 (26)

National Child Development Study: Age 46, Sweep 7, 2004-2005 (27)

National Child Development Study: Age 50, Sweep 8, 2008-2009 (28)

Stata code that derives a dataset with all the variables predicting non-response in the NCDS listed below, is available on the CLS Data GitHub.

Table A1. Predictors of non-response at Sweep 1 (age 7).

Sweep	Variable description	Variable derivation details
Sweep 0 (age 0)	Region	Recode of n0region (“Region at PMS (1958) – Birth”): 1 "North" 2 "Midlands" 3 "East & South East" 4 "South & South West" 5 "Wales" 6 "Scotland" “Not in PMS” → Missing
	Number of persons per room	Recode of n512 (“0 M Number of persons per room”): “NA” → Missing
	Social class of mother’s husband	Recode of n236 (“0P Social class of mother's husband (GRO 1951)”): “NA, NMH” / “Unemployed, sick” → Missing In some non-response models it was necessary to merge categories “I” and “II” to ensure convergence. If you do not encounter such convergence issues in your own analysis then we would encourage you to use the original, richer coding of the variable. If convergence issues are encountered then this (or similar) recoding may be necessary.

Table A2. Predictors of non-response at Sweep 2 (age 11).

Sweep	Variable description	Variable derivation details
Sweep 0 (age 0)	Mother's present marital status	Recode of n545 (“0 Mother's present marital status”): “NA, incomplt info” → Missing “Sep,Div,Widowed” / “Stable union” / “Unmarried” → “Unmarried / stable union / sep, div, widowed” “Twice married” / “Married” → “Married/Twice married”
Sweep 1 (age 7)	Number of kids under 21 in the household, including living away	Recode of n99 (“1P No. kids undr 21 hhld,inc liv away”): “NA,not prvt fam” → Missing
	Common difficulties age 7 (mother)	Derived from summing all positive replies to variables n123 to n131 (headaches, temper tantrums, reluctance to go to school, etc)
	Hospital admissions	Derived from summing all positive replies to variables n246 to n254 (admissions for up to 9 specific conditions: tonsils, adenoids, tummy operation, hernia, accidents etc)
	Cognitive ability summary	Principal Component Analysis (PCA) score. PCA indicators were the Problem Arithmetic Test score, Total score on Copying Designs Test, Drawing a Man Test score and the Southgate Group Reading Test score.
	Non-response at sweep 1	Derived from OUTCME01 as follows: 0 if respondent at sweep 1 1 if non-respondent at sweep 1

Table A3. Predictors of non-response at Sweep 3 (age 16).

Sweep	Variable description	Variable derivation details
Sweep 0 (age 0)	Region	Recode of n0region (“Region at PMS (1958) – Birth”): 1 "North" 2 "Midlands" 3 "East & South East" 4 "South & South West" 5 "Wales" 6 "Scotland" “Not in PMS” → Missing
Sweep 1 (age 7)	Number of kids under 21 in the household, including living away	Recode of n99 (“1P No. kids undr 21 hhld,inc liv away”): “NA,not prvt fam” → Missing
	Mother worked birth to 5	Derived from n197 and n198
	Ever breastfed	Recode of n222 (“1P Breast fed partially or wholly”): “NA” / “Dont know” → Missing “No” → “Never breastfed “Under one month” / “Over one month” → “Ever breastfed”
Sweep 2 (age 11)	Non-response at sweeps 1-2	Derived from OUTCME01- OUTCME02 as follows: 0 if respondent at both sweeps 1 if non-respondent at one or both sweeps

Table A4. Predictors of non-response at Sweep 4 (age 23).

Sweep	Variable description	Variable derivation details
Sweep 0 (age 0)	Region	Recode of n0region (“Region at PMS (1958) – Birth”): 1 "North" 2 "Midlands" 3 "East & South East" 4 "South & South West" 5 "Wales" 6 "Scotland" “Not in PMS” → Missing
	Number of persons per room	Recode of n512 (“0 M Number of persons per room”): “NA” → Missing
	Sex of child	Recode of n622 (“0-3D Sex of child”) “Not known” → Missing
	Social class of mother’s husband	Recode of n236 (“0P Social class of mother's husband (GRO 1951)”): “NA, NMH” / “Unemployed, sick” → Missing In some non-response models it was necessary to merge categories “I” and “II” to ensure convergence. If you do not encounter such convergence issues in your own analysis then we would encourage you to use the original, richer coding of the variable. If convergence issues are encountered then this (or similar) recoding may be necessary.
Sweep 1 (age 7)	Family moves since child's birth	Recode of n95 (“1P Family moves since child's birth”): “NA” → Missing “5” to “22” → “5+”
	Cognitive ability summary	Principal Component Analysis (PCA) score. PCA indicators were the Problem Arithmetic Test score, Total score on Copying Designs Test, Drawing a Man Test score and the Southgate Group Reading Test score.

Sweep	Variable description	Variable derivation details
	Dad reads to child	Recode of n180 (“1P Father reads to child”): “NA” / “DK,inapplicable” → Missing “Yes,every week” / “Yes,occasionally” → “Every week / occasionally”
Sweep 2 (age 11)	Area of world in which mother born	Recode of n1434 (“2P Area of world in which mother born”): “NA” → Missing“Europe inc USSR” / “Cyprus,Gib,Malta” → “Europe inc USSR” “India & Pakistan” / “West Indies” / “Asia” / “Africa” / “Others” → “Outside Europe”
	Number of family moves since child’s birth	Recode of n1150 (“2P No. of family moves since chlds birth”): “NA” → Missing “4” to “9” →”4+”
	Cognitive ability summary	Derived from genability11 (i.e. summing n914 and n917: the NFER General Ability Test)
	Number of household amenities	Sum of sole use of: indoor toilet, outside toilet, cooking facilities, hot water supply, garden or yard (variables n204 to n209)
Sweep 3 (age 16)	Number of family moves since child’s birth	Recode of n2492 (“3P No. family moves since childs birth”): “NA” → Missing
	Sum of favourable learning environments/outcomes re sex educ etc)	Derived from summing positive answers to each of 39 variables n2825 to n2863
	Conduct problems	Both mother and teacher ratings; eight items for a unidimensional 2 parameter probit measurement model with the Mplus defaults. Simply summing the items is an acceptable approximation for missing data generation.
	Non-response at sweeps 1-3	Derived from OUTCME01-OUTCME03 as follows:

Sweep	Variable description	Variable derivation details
		0 if respondent at all sweeps 1 if non-respondent at any one or more sweeps

Table A5. Predictors of non-response at Sweep 5 (age 33).

Sweep	Variable description	Variable derivation details
Sweep 0 (age 0)	Number of persons per room	Recode of n512 ("0 M Number of persons per room"): "NA" → Missing
	Sex of child	Recode of n622 ("0-3D Sex of child") "Not known" → Missing
	Social class of mother's husband	Recode of n236 ("0P Social class of mother's husband (GRO 1951)"): "NA, NMH" / "Unemployed, sick" → Missing In some non-response models it was necessary to merge categories "I" and "II" to ensure convergence. If you do not encounter such convergence issues in your own analysis then we would encourage you to use the original, richer coding of the variable. If convergence issues are encountered then this (or similar) recoding may be necessary.
Sweep 1 (age 7)	Family moves since child's birth	Recode of n95 ("1P Family moves since child's birth"): "NA" → Missing "5" to "22" → "5+"
	Social problems (alcoholism etc.)	Derived from summing positive answers to 13 variables n314 to n326 (housing probs, financial, illness/disability/death of parent, unemployment, divorce etc)
	Cognitive ability summary	Principal Component Analysis (PCA) score. PCA indicators were the Problem Arithmetic Test score, Total score on Copying Designs Test, Drawing a Man Test score and the Southgate Group Reading Test score.

Sweep	Variable description	Variable derivation details
	Summary of medical conditions	Sum of: n215 n216 n217 n218 n219 n220 n221 n263 n225 n226 n266 n279 n281 n362 n363 n367 n368 n369 n370 n421 n422 n423 n371 n424 n372 n425 n350 n354 n351 n345 n352 n342 n353 n259 n364 (measles/German measles, whooping cough, chicken pox, mumps, scarlet fever, glandular fever, heart conditions, disability, hernia, spasms, limb defect, malfunction of upper limb, stuttering, skull deformity, spinal disorders, congenital lower limb defect, talipes, malfunction of the lower limb, neurological skeletal disorder, asthma, mouth or palate abnormality, major handicap/disfiguring condition, abnormal chest shape, eczema, external ear deformity)
	Ever breastfed	Recode of n222 (“1P Breast fed partially or wholly”): “NA” / “Dont know” → Missing “No” → “Never breastfed “Under one month” / “Over one month” → “Ever breastfed”
Sweep 2 (age 11)	Child’s positive activities outside school	Derived from summing positive answers (i.e. 'done often' or 'done sometime') to 22 variables n933 to n954 (meets friends outside school, goes to cinema, reads books/newspapers, listens to music etc.)
	Cognitive ability summary	Derived from genability11 (i.e. summing n914 and n917: the NFER General Ability Test)
	Number of household amenities	Sum of sole use of: indoor toilet, outside toilet, cooking facilities, hot water supply, garden or yard. (variables n204 to n209)
Sweep 3 (age 16)	Number of family moves since child’s birth	Recode of n2492 (“3P No. family moves since childs birth”): “NA” → Missing
	How long since child drank alcohol	Recode of n2888 (“3I How long since chld drank alcohol”): 1 "Less than 1 week" 2 "2-4 weeks" 3 "5 or more weeks" 4 "Do not remember" 5 "Never had one" “NA” → Missing

Sweep	Variable description	Variable derivation details
	Test 2 – mathematics comprehension	Recode of n2930 (“3T Test 2-mathematics comprehension”): “NA” → Missing
	Sum of favourable learning environments/outcomes re sex educ etc)	Derived from summing positive answers to each of 39 variables n2825 to n2863
Sweep 4 (age 23)	Type of current accommodation	Recode of n5318 (“4I Type of current accommodation”): “Not answered” → Missing “ROOMS” / “CARAVAN” / “OTHER” → “Other”
	Voted in 1979 general election	Recode of n5959 (“4I Whether voted in 1979 general election”): “Don’t know” / “Not answered” → Missing
	Economic status	Recode of econstrg (“4D Economic status grouped - 4 categories”): “NO INFORMATION” → Missing
	Number of voluntary activities (youth club, church etc.)	Derived from summing positive answers to 5 variables n5950, n5953, n5956, n5965 & n5969. Note that n5965 is routed through n5964 (ever trades union member), and n5969 is routed through n5967 (whether has a religion)
	Non-response at sweeps 1-4	Derived from OUTCME01-OUTCME04 as follows: 0 if respondent at all sweeps 1 if non-respondent at any one or more sweeps

Table A6. Predictors of non-response at Sweep 6 (age 42).

Sweep	Variable description	Variable derivation details
Sweep 0 (age 0)	Number of persons per room	Recode of n512 (“0 M Number of persons per room”): “NA” → Missing
	Sex of child	Recode of n622 (“0-3D Sex of child”) “Not known” → Missing
	Social class of mother’s husband	Recode of n236 (“0P Social class of mother's husband (GRO 1951)”): “NA, NMH” / “Unemployed, sick” → Missing In some non-response models it was necessary to merge categories “I” and “II” to ensure convergence. If you do not encounter such convergence issues in your own analysis then we would encourage you to use the original, richer coding of the variable. If convergence issues are encountered then this (or similar) recoding may be necessary.
Sweep 1 (age 7)	Cognitive ability summary	Principal Component Analysis (PCA) score. PCA indicators were the Problem Arithmetic Test score, Total score on Copying Designs Test, Drawing a Man Test score and the Southgate Group Reading Test score.
Sweep 2 (age 11)	Area of world in which father born	Derived from n1436 (“2P Area of world in which father born”): “NA” → Missing “Europe inc USSR” / “Cyprus,Gib,Malta” → “Europe inc USSR” “India & Pakistan” / “West Indies” / “Asia” / “Africa” / “Others” → “Outside Europe”
	Child’s positive activities outside school	Derived from summing positive answers (i.e. 'done often' or 'done sometime') to 22 variables n933 to n954 (meets friends outside school, goes to cinema, reads books/newspapers, listens to music etc.)
	Cognitive ability summary	Derived from genability11 (i.e. summing n914 and n917: : the NFER General Ability Test)

Sweep	Variable description	Variable derivation details
Sweep 3 (age 16)	How long since child drank alcohol	Recode of n2888 (“3I How long since chld drank alcohol”): 1 "Less than 1 week" 2 "2-4 weeks" 3 "5 or more weeks" 4 "Do not remember" 5 "Never had one" “NA” → Missing
	Sum of good activities performed outside school	Derived from summing positive answers (i.e. 'often' and 'sometimes') to 8 variables n2864 to n2871 (reads, plays sport, swims, watches TV, goes to parties, dancing, voluntary work)
	Conduct problems	Both mother and teacher ratings; eight items for a unidimensional 2 parameter probit measurement model with the Mplus defaults. Simply summing the items is an acceptable approximation for missing data generation.
Sweep 4 (age 23)	Legal marital status	Recode of n5113 (“4I Legal marital status”): “Separated” / “Divorced” / “Widowed” → “Separated / divorced / widowed”
	Voted in 1979 general election	Recode of n5959 (“4I Whether voted in 1979 general election”): “Don’t know” / “Not answered” → Missing
Sweep 5 (age 33)	Type of accommodation	Recode of n502940 (“CMI:56 D4 Type of accommodation”): “Converted flat, etc” / “Rooms” / “Caravan” / “Mobile home” / “House-boat” / “Other” → “Converted flat, rooms, caravan, miscell”
	Current member of a Trade Union/Staff Association	Recode of n504646 (“CMI:96 G11 Current member of a Trade Union/Staff Association”): “Yes-Trade Union” / “Yes-Staff Assoc” →”Yes”
	Social capital score (people turn to for advice, support)	Derived from summing positive answers to 24 questions n509531 to n509554 on ‘who you turn to for support – discuss important life changes’, etc Considered as a continuous variable; risk ratio reported as an association per 10 years for ease of interpretation.

Sweep	Variable description	Variable derivation details
	Life contentment score	Derived from summing positive answers to nine questions: n509765 (very or fairly happy), n509766 (very or fairly happy), n509767, n509768, n509769, n509770, n509771, n509772 (scores 7/10), n509774 (scores 7/10), regarding general level of happiness, trust in other people, whether drawn into arguments or not, etc.
	Non-response at sweeps 1-5	Derived from OUTCME01-OUTCME05 as follows: 0 if respondent at all sweeps 1 if non-respondent at any one or more sweeps

Table A7. Predictors of non-response at Biomedical Sweep (age 44).

Sweep	Variable description	Variable derivation details
Sweep 0 (age 0)	Number of persons per room	Recode of n512 (“0 M Number of persons per room”): “NA” → Missing
	Abnormality during pregnancy	Recode of n522 (“0 Abnormality during pregnancy”): “NA” → Missing “None” → “No” “Accidental APH” / “Placenta praevia” / “Other APH” / “APH & 8” / “3 & 8” / “4 & 8” / “Vaginal bleeding” / “Other abnormalty” → “Yes”
	Social class of mother's father when she left school	Recode of n660 (“0 Mother's,father's social class 1958 (GRO 1951)”): “NA” / “No information” → Missing In some non-response models it was necessary to merge categories “I” and “II” to ensure convergence. If you do not encounter such convergence issues in your own analysis then we would encourage you to use the original, richer coding of the variable. If convergence issues are encountered then this (or similar) recoding may be necessary.
	Sex of child	Recode of n622 (“0-3D Sex of child”) “Not known” → Missing
	Social class of mother’s husband	Recode of n236 (“0P Social class of mother's husband (GRO 1951)”): “NA, NMH” / “Unemployed, sick” → Missing

Sweep	Variable description	Variable derivation details
		In some non-response models it was necessary to merge categories “I” and “II” to ensure convergence. If you do not encounter such convergence issues in your own analysis then we would encourage you to use the original, richer coding of the variable. If convergence issues are encountered then this (or similar) recoding may be necessary.
Sweep 1 (age 7)	Dad stayed on at school after minimum age	Recode of n194 (1P Dad stayed on at school after min age”): “NA” / “Dont know” → Missing
	Attendance	Recode of n458 (“1S Attendance-MC1:4”): “NA” → Missing “Long absences” / “Malingerer etc” / “Helps at home” → “Long absences / malingerer etc / helps at home”
	Social problems (alcoholism etc.)	Derived from summing positive answers to 13 variables n314 to n326 (housing probs, financial, illness/disability/death of parent, unemployment, divorce etc)
	Cognitive ability summary	Principal Component Analysis (PCA) score. PCA indicators were the Problem Arithmetic Test score, Total score on Copying Designs Test, Drawing a Man Test score and the Southgate Group Reading Test score.
	Body mass index	Derived from weight (n337) & height (n332, n334) at age 7
Sweep 2 (age 11)	Cognitive ability summary	Derived from genability11 (i.e. summing n914 and n917: the NFER General Ability Test)
Sweep 3 (age 16)	Emotional or behavioural problem	Recode of n2021 (“3M Emotional,behavioural problem”):

Sweep	Variable description	Variable derivation details
		“NA” → Missing “Cond uncertain” / “Cond,no handcp” / “Slight handicap” / “Moderate handicap” / “Severe handicap” / “Cond,DK degree” → “Any condition or handicap”
	How long since child drank alcohol	Recode of n2888 (“3I How long since chld drank alcohol”): 1 "Less than 1 week" 2 "2-4 weeks" 3 "5 or more weeks" 4 "Do not remember" 5 "Never had one" “NA” → Missing
	Test 2 – mathematics comprehension	Recode of n2930 (“3T Test 2-mathematics comprehension”): “NA” → Missing
	Conduct problems	Both mother and teacher ratings; eight items for a unidimensional 2 parameter probit measurement model with the Mplus defaults. Simply summing the items is an acceptable approximation for missing data generation.
Sweep 4 (age 23)	Voted in 1979 general election	Recode of n5959 (“4I Whether voted in 1979 general election”)003A “Don’t know” / “Not answered” → Missing
Sweep 5 (age 33)	Any work related training course since March 1981	n501237 (“CMI:27 B10) Any work related training course since March 1981”)
	Number of hospital admissions since March 1981	Recode of n504215 (“CMI:84 F38b) No. of hospital admissions since March 1981”) incorporating information from n504214 (“CMI:84 F38a) Admitted to hospital overnight since March 1981”): Missing → 0 if n504214 = “No”

Sweep	Variable description	Variable derivation details
	Driven/ridden after drinking alcohol in last 7 days	Recode of n504427 (“CMI:92 F69 Driven/ridden after drinking alcohol in last 7 days”) incorporating information from n504366 (“CMI:91 F64 Holds licences to drive/ride motor vehicles”): Missing → “Doesn’t drive” if n504366 = “No” 1 "Yes", 2 "No", 3 "Doesn't drive"
	Social capital score (people turn to for advice, support)	Derived from summing positive answers to 24 questions n509531 to n509554 on ‘who you turn to for support – discuss important life changes’, etc Considered as a continuous variable; risk ratio reported as an association per 10 years for ease of interpretation.
Sweep 6 (age 42)	Normally has access to a car or van	Derived from caracces (“Does CM normally have access to a car or van”): “=Dont know” / “=Not answered” → Missing
	Participated in NCDS V	dmpart (“If NCDS CM participated in NCDS V”)
	Intends to move in near future	Recode of wantmove ("Does CM intend to move in near future"): “Don’t know” / “Not answered” → Missing
	Has a computer at home	Recode of pchome (“Do you have a computer at home?”): “Dont know” / “Not answered” → Missing
	Non-response at sweeps 1-6	Derived from OUTCME01-OUTCME06 as follows: 0 if respondent at all sweeps 1 if non-respondent at any one or more sweeps

Table A8. Predictors of non-response at Sweep 7 (age 46).

Sweep	Variable description	Variable derivation details
Sweep 0 (age 0)	Number of persons per room	Recode of n512 (“0 M Number of persons per room”): “NA” → Missing
	Sex of child	Recode of n622 (“0-3D Sex of child”) “Not known” → Missing
	Social class of mother’s husband	Recode of n236 (“0P Social class of mother's husband (GRO 1951)”): “NA, NMH” / “Unemployed, sick” → Missing In some non-response models it was necessary to merge categories “I” and “II” to ensure convergence. If you do not encounter such convergence issues in your own analysis then we would encourage you to use the original, richer coding of the variable. If convergence issues are encountered then this (or similar) recoding may be necessary.
Sweep 1 (age 7)	Dad stayed on at school after minimum age	Recode of n194 (1P Dad stayed on at school after min age”): “NA” / “Dont know” → Missing
	Attendance	Recode of n458 (“1S Attendance-MC1:4”): “NA” → Missing “Long absences” / “Malingerer etc” / “Helps at home” → “Long absences / malingerer etc / helps at home”

Sweep	Variable description	Variable derivation details
	Social problems (alcoholism etc.)	Derived from summing positive answers to 13 variables n314 to n326 (housing probs, financial, illness/disability/death of parent, unemployment, divorce etc)
	Cognitive ability summary [per unit]	Principal Component Analysis (PCA) score. PCA indicators were the Problem Arithmetic Test score, Total score on Copying Designs Test, Drawing a Man Test score and the Southgate Group Reading Test score.
Sweep 2 (age 11)	Source of family income last year	Recode of n1176 (“2P Source-family income last year-MC 1:5”): “NA" → Missing “Sickness benefit” / “Unemployment pay” / “Supplementary” / “Pension,retired” / “Pension,disabled” / “Other sources” → “Other sources”
	Child’s positive activities outside school	Derived from summing positive (i.e. 'done often' or 'done sometime') answers to 22 variables n933 to n954 (meets friends outside school, goes to cinema, reads books/newspapers, listens to music etc.)
	Cognitive ability summary	Derived from genability11 (i.e. summing n914 and n917: the NFER General Ability Test)
Sweep 3 (age 16)	Local Authority & voluntary schools	Recode of n2102 ("3S Local Authority & voluntary schools"): "Technical" / "All-age" / "Res.spec.non ESN" / "Day spec.non ESN" / "Other" / "Day ESN school" / "Res.ESN school" → "Other" / "Non LEA school" → Missing
	Wish could leave school at 15 – study child	Recode of n2741 (“3I Wish could leave sch at 15-stdy child”): “NA” → Missing
	How long since child drank alcohol	Recode of n2888 (“3I How long since chld drank alcohol”):

Sweep	Variable description	Variable derivation details
		1 "Less than 1 week" 2 "2-4 weeks" 3 "5 or more weeks" 4 "Do not remember" 5 "Never had one" “NA” → Missing
	Test 2 – mathematics comprehension	Recode of n2930 (“3T Test 2-mathematics comprehension”): “NA” → Missing
Sweep 4 (age 23)	Number of accidents since 16th birthday	Recode of n5819 (“4I Number of accidents since 16th birthday”) incorporating information from n5819 (“4I Any accidents since 16th birthday”): “CANT REMEMBER” → Missing Missing → 0 if n5819 = “No”
	Voted in 1979 general election	Recode of n5959 (“4I Whether voted in 1979 general election”): “Don’t know” / “Not answered” → Missing
Sweep 5 (age 33)	Voted in 1987 general election	n504635 (“CMI:96 G4 Voted in General Election in June 1987”)
	Social capital score (people turn to for advice, support)	Derived from summing positive answers to 24 questions n509531 to n509554 on ‘who you turn to for support – discuss important life changes’, etc Considered as a continuous variable; risk ratio reported as an association per 10 years for ease of interpretation.
Sweep 6 (age 42)	Participated in NCDS V	dmpart (“If NCDS CM participated in NCDS V”)
	Intends to move in near future	Recode of wantmove ("Does CM intend to move in near future"):

Sweep	Variable description	Variable derivation details
		“Don’t know” / “Not answered” → Missing
	Membership in organisations	Derived from **orgever1** asking whether ever member of an organisation (political party, environmental group, other charity, women’s groups, Townswomen’s Guild/Women’s Institute, parent school organisation, or tenants’ associations) "Don't know" (98) / "Not answered" (99) → Missing
Biomedical Sweep (age 44)	Current legal marital status	Recode of marital (“SC2 (C3): What is your curent legal marital status”): “Not answered (9)” / “Self Completion 2 not done“ → Missing “Legally separated” / “Divorced” / “Widowed“ → “Separated/divorced/widowed”
	Is current accommodation owned or rented?	Recode of ownhome (“SC1 (G1): Do you own or rent your home or is there some other arrangement?”): “Not answered (99)” / “Self Completion 1 not done“ → Missing “Own – outright” / “Own - buying with help of mortgage/loan” → “Own home” “Pay part rent and part mortgage” / “Rent from local authority or housing as” / “Rent from private landlord, relative or” / “Live here rent-free, including rent-fre” / “Other arrangement” → “Don’t own home”
	Non-response at sweeps 1-biomedical	Derived from OUTCME01-OUTCMEBM as follows: 0 if respondent at all sweeps 1 if non-respondent at any one or more sweeps

Table A9. Predictors of non-response at Sweep 8 (age 50).

Sweep	Variable description	Variable derivation details
Sweep 0 (age 0)	Number of persons per room	Recode of n512 ("0 M Number of persons per room"): "NA" → Missing
	Sex of child	Recode of n622 ("0-3D Sex of child") "Not known" → Missing
	Social class of mother's husband	Recode of n236 ("0P Social class of mother's husband (GRO 1951)"): "NA, NMH" / "Unemployed, sick" → Missing In some non-response models it was necessary to merge categories "I" and "II" to ensure convergence. If you do not encounter such convergence issues in your own analysis then we would encourage you to use the original, richer coding of the variable. If convergence issues are encountered then this (or similar) recoding may be necessary.
Sweep 1 (age 7)	Social problems (alcoholism etc.)	Derived from summing positive answers to 13 variables n314 to n326 (housing probs, financial, illness/disability/death of parent, unemployment, divorce etc)
	Cognitive ability summary	Principal Component Analysis (PCA) score. PCA indicators were the Problem Arithmetic Test score, Total score on Copying Designs Test, Drawing a Man Test score and the Southgate Group Reading Test score.
	Summary of medical conditions	Sum of: n215 n216 n217 n218 n219 n220 n221 n263 n225 n226 n266 n279 n281 n362 n363 n367 n368 n369 n370 n421 n422 n423 n371 n424 n372 n425 n350 n354 n351 n345 n352 n342 n353 n259 n364 (measles/German measles, whooping cough, chicken pox, mumps, scarlet fever, glandular fever, heart

Sweep	Variable description	Variable derivation details
		conditions, disability, hernia, spasms, limb defect, malfunction of upper limb, stuttering, skull deformity, spinal disorders, congenital lower limb defect, talipes, malfunction of the lower limb, neurological skeletal disorder, asthma, mouth or palate abnormality, major handicap/disfiguring condition, abnormal chest shape, eczema, external ear deformity)
Sweep 2 (age 11)	Cognitive ability summary	Derived from genability11 (i.e. summing n914 and n917: the NFER General Ability Test)
	Conduct problems	Maternal reported conduct problems summed from n1450 n1454 n1458 n1460
Sweep 3 (age 16)	C'ild's school attendance	Recode of n1721 ("3SD NCDS child's school attendance"): -1 → Missing
	How long since child drank alcohol	Recode of n2888 ("3I How long since chld drank alcohol"): 1 "Less than 1 week" 2 "2-4 weeks" 3 "5 or more weeks" 4 "Do not remember" 5 "Never had one" "NA" → Missing
	Test 2 – mathematics comprehension	Recode of n2930 ("3T Test 2-mathematics comprehension"): "NA" → Missing
	Conduct problems	Both mother and teacher ratings; eight items for a unidimensional 2 parameter probit measurement model with the Mplus defaults. Simply summing the items is an acceptable approximation for missing data generation.
Sweep 4 (age 23)	Legal marital status	Recode of n5113 ("4I Legal marital status"): "Separated" / "Divorced" / "Widowed" → "Separated / Divorced / Widowed"

Sweep	Variable description	Variable derivation details
	Voted in 1979 general election	Recode of n5959 ("4I Whether voted in 1979 general election"): "Don't know" / "Not answered" → Missing
	Economic status	Recode of econstrg ("4D Economic status grouped - 4 categories"): "NO INFORMATION" → Missing
Sweep 5 (age 33)	Voted in 1987 general election	n504635 ("CMI:96 G4 Voted in General Election in June 1987")
	Social capital score (people turn to for advice, support)	Derived from summing positive answers to 24 questions n509531 to n509554 on 'who you turn to for support – discuss important life changes', etc Considered as a continuous variable; risk ratio reported as an association per 10 years for ease of interpretation.
	Life contentment score	Derived from summing positive answers to nine questions: n509765 (very or fairly happy), n509766 (very or fairly happy), n509767, n509768, n509769, n509770, n509771, n509772 (scores 7/10), n509774 (scores 7/10), regarding general level of happiness, trust in other people, whether drawn into arguments or not, etc.
Sweep 6 (age 42)	Frequency of eating biscuits and cakes of all kinds	Recode of cakes ("Freq of eating biscuits and cakes of all kinds"): "=Dont know" / "=Not answered" → Missing
	Is current accommodation owned or rented?	Recode of tenure2 ("Is current accom owned or rented"): "Not answered" → Missing "Own – outright" / "Own - with mortgage" → "Own home" "Shared ownership" / "Rent it" / "Live rent-free" / "Other" →"Don't own home"

Sweep	Variable description	Variable derivation details
	Participated in NCDS V	dmpart (“If NCDS CM participated in NCDS V”)
	Ever wanted improve your maths?	Recode of mthimp ("CM ever wanted improve your maths?"): “Not answered” → Missing
	Membership in organisations	Derived from **orgever1** asking whether ever member of an organisation (political party, environmental group, other charity, women’s groups, Townswomens Guild/Womens Institute, parent school organisation, or tenants’ associations) "Don't know" (98) / "Not answered" (99) → Missing
Biomedical Sweep (age 44)	Consent to access NHS records	nhsok (“Consent to access NHS records”)
	How many children do you have living with you aged 18 or less	Recode of childnow (“(Modify var from CHILDNOW) How many children aged 18 or less live with you, whom”): “Not answered (99)” / “Self Completion 2 not done” → Missing “Not applicable” → 0 "5"/"6 or more" → "4 or more" Categories: "0", "1", "2, "3, "4 or more"
	How many natural (biological) children have you ever had	Recode of childnum (“(Modify var from CHILDNOW) How many biological children have you ever had, that”): “Not answered (99)” / “Self Completion 2 not done” → Missing "6"/"7 or more" → "5 or more"

Sweep	Variable description	Variable derivation details
		Categories: "0", "1", "2, "3, "4", "5 or more"
Sweep 7 (age 46)	Non-response at sweeps 1-7	Derived from OUTCME01-OUTCME07 as follows: 0 if respondent at all sweeps 1 if non-respondent at any one or more sweeps

Table A10. Predictors of non-response at sweep 9 (age 55).

Sweep	Variable description	Variable derivation details
Sweep 0 (age 0)	Mother's age	Recode of n553 ("0 Mother's age last birthday, in years"): "NA" / 8 → Missing Considered as a continuous variable; risk ratio reported as an association per 10 years for ease of interpretation.
	Number of persons per room	Recode of n512 ("0 M Number of persons per room"): "NA" → Missing
	Parity	Recode of n504 ("0 Parity"): "NA" → Missing
	Social class of mother's father when she left school	Recode of n660 ("0 Mother's,father's social class 1958 (GRO 1951)"): "NA" / "No information" → Missing In some non-response models it was necessary to merge categories "I" and "II" to ensure convergence. If you do not encounter such convergence issues in your own analysis then we would encourage you to use the original, richer coding of the variable. If convergence issues are encountered then this (or similar) recoding may be necessary.
	Sex of child	Recode of n622 ("0-3D Sex of child") "Not known" → Missing

Sweep	Variable description	Variable derivation details
	Social class of mother's husband	Recode of n236 ("0P Social class of mother's husband (GRO 1951)"): "NA, NMH" / "Unemployed, sick" → Missing In some non-response models it was necessary to merge categories "I" and "II" to ensure convergence. If you do not encounter such convergence issues in your own analysis then we would encourage you to use the original, richer coding of the variable. If convergence issues are encountered then this (or similar) recoding may be necessary.
Sweep 1 (age 7)	Dad stayed on at school after minimum age	Recode of n194 (1P Dad stayed on at school after min age"): "NA" / "Dont know" → Missing
	Social problems (alcoholism etc.)	Derived from summing positive answers to 13 variables n314 to n326 (housing probs, financial, illness/disability/death of parent, unemployment, divorce etc)
	Cognitive ability summary	Principal Component Analysis (PCA) score. PCA indicators were the Problem Arithmetic Test score, Total score on Copying Designs Test, Drawing a Man Test score and the Southgate Group Reading Test score.
	Ever breastfed	Recode of n222 ("1P Breast fed partially or wholly"): "NA" / "Dont know" → Missing "No" → "Never breastfed "Under one month" / "Over one month" → "Ever breastfed"
Sweep 2 (age 11)	Cognitive ability summary	Derived from genability11 (i.e. summing n914 and n917: the NFER General Ability Test)

Sweep	Variable description	Variable derivation details
	Conduct problems	Maternal reported conduct problems summed from n1450 n1454 n1458 n1460
Sweep 3 (age 16)	Child receiving help at school – backwardness	Recode of n2250 (“3S Child receiving help sch-backwardness”): “NA” → Missing “Yes “ / “Yes,not enough” / “No,but desirable” → “Yes” “No,not necessary” →”No”
	Child's school attendance	Recode of n1721 (“3SD NCDS child's school attendance”): -1 → Missing
	How long since child drank alcohol	Recode of n2888 (“3I How long since chld drank alcohol”): 1 "Less than 1 week" 2 "2-4 weeks" 3 "5 or more weeks" 4 "Do not remember" 5 "Never had one" “NA” → Missing
	Test 2 – mathematics comprehension	Recode of n2930 (“3T Test 2-mathematics comprehension”): “NA” → Missing
	Conduct problems	Both mother and teacher ratings; eight items for a unidimensional 2 parameter probit measurement model with the Mplus defaults. Simply summing the items is an acceptable approximation for missing data generation.
Sweep 4 (age 23)	Legal marital status	Recode of n5113 (“4I Legal marital status”): “Separated” / “Divorced” / “Widowed” → “Separated / Divorced / Widowed”

Sweep	Variable description	Variable derivation details
	Voted in 1979 general election	Recode of n5959 (“4I Whether voted in 1979 general election”): “Don’t know” / “Not answered” → Missing
Sweep 5 (age 33)	Telephone in home	n502977 (“CMI:58 D18 Telephone in home (sole/shared use)”)
	How much physical effort in job	Recode of n504361 (“CMI:90 F60b) How much physical effort in job”): “Can't say” / “Not answered” → Missing
	Voted in 1987 general election	n504635 (“CMI:96 G4 Voted in General Election in June 1987”)
	Housing tenure	Recode of tenure91 (“DV:Housing tenure in 1991”): “missing” → Missing "owns outright” / “mortgage” → ”Owns home” “social housing” / “other rented” / “other” →”Doesn’t own home”
	Social capital score (people turn to for advice, support)	Derived from summing positive answers to 24 questions n509531 to n509554 on ‘who you turn to for support – discuss important life changes’, etc Considered as a continuous variable; risk ratio reported as an association per 10 years for ease of interpretation.
Sweep 6 (age 42)	Participated in NCDS V	dmpart (“If NCDS CM participated in NCDS V”)

Sweep	Variable description	Variable derivation details
	Membership in organisations	Derived from **orgever1** asking whether ever member of an organisation (political party, environmental group, other charity, women's groups, Townswomen's Guild/Women's Institute, parent school organisation, or tenants' associations) "Don't know" (98) / "Not answered" (99) → Missing
Biomedical Sweep (age 44)	Self-rated general health	Recode of genhlth ("SC2 (A1): How would you describe your health generally"): "Not answered (9)" / "Self Completion 2 not done" → Missing
Sweep 7 (age 46)	Marital status - de facto	Recode of nd7ms ("(Derived) marital status - de facto"): "Other missing" / "Did not complete HH grid" → Missing "Separated" / "Divorced" / "Widowed" → "Separated / divorced / widowed"
Sweep 8 (age 50)	Total number of natural children	Recode of ND8NCHTT ("(Derived) Total number of cohort member's natural children (in HH and absent)"): "HH grid/absent child grid not complete" → Missing
	Employer provided pension scheme	Recode of n8j2101 ("[J211] Pensions: employer provided Pension sch"): "Refusal" / "Don't Know" / "Item not applicable" → Missing
	Non-response at sweeps 1-8	Derived from OUTCME01-OUTCME08 as follows: 0 if respondent at all sweeps 1 if non-respondent at any one or more sweeps

Appendix B: Identified predictors of non-response in the 1970 British Cohort Study (BCS70).

Sweep 1 (age 0) to Sweep 9 (age 42) BCS70 data are available from the UK Data Service:

1970 British Cohort Study: Birth and 22-Month Subsample, 1970-1972 (29)

1970 British Cohort Study: Age 5, Sweep 2 1975 (30)

1970 British Cohort Study: Age 10, Sweep 3, 1980 (31)

1970 British Cohort Study: Age 16, Sweep 4, 1986 (32)

1970 British Cohort Study: Age 26, Sweep 5, 1996 (33)

1970 British Cohort Study: Age 29, Sweep 6, 1999-2000 (34)

1970 British Cohort Study: Age 34, Sweep 7, 2004-2005 (35)

1970 British Cohort Study: Age 38, Sweep 8, 2008-2009 (36)

1970 British Cohort Study: Age 42, Sweep 9, 2012 (37)

Stata code that derives a dataset with all the variables predicting non-response in the BCS70 listed below, is available on the CLS Data GitHub.

Table B1. Predictors of non-response at Sweep 2 (age 5).

Sweep	Variable description	Variable derivation details
Sweep 1 (age 0)	Marital Status	a0012 Single / Married / Divorced,Separated,Widowed
	Parity	Recoded from A0166 0 / 1 / 2 / 3 / >4
	Father's social status	Derived from a0014: 1 Social class 1 / 2 Social class 2 / 3 Social class 3 (Non-Manual) / 4 Social class 3 (Manual) / 5 Social class 4 / 6 Social class 5 / Other and unsupported (as missing)
	Country of Birth	Recoded from COB England / Wales / Scotland / Other
	Father's age at completion of education	Recoded from a0010 ≤15 year old / 16-18 year old / ≥19 years old
	Number of antenatal visits	a0190 Continuous (per visit)
	Method of contraception	a0029b None / Pill alone / Pill alone and other method / Other method

Table B2. Predictors of non-response at Sweep 3 (age 10).

Sweep	Variable description	Variable derivation details
Sweep 1 (age 0)	Marital Status	a0012 Single / Married / Divorced,Separated,Widowed
	Where was mother-date of birth	Recoded from A0357 Consultant bed/ GP bed / NHS unit / Own home / Other
	Country of Birth	Recoded from COB England / Wales / Scotland / Other
	Father's age at completion of education	Recoded from a0010 ≤15 year old / 16-18 year old / ≥19 years old
Sweep 2 (age 5)	Ethnic group (mother)	Recoded from e246a Non-European vs European
	Attitude to childhood independence	d124h Per Unit increase in z-score
	Tenure of accommodation	Recoded from e220 Owned / Bought / Council rented / Other
	Household moves (ordered)	Recoded from e249 0 / 1 / 2 / 3 / 4 / >5
Non-response at previous sweep (i.e. sweep 2)		Yes vs no

Table B3. Predictors of non-response at Sweep 4 (age 16).

Sweep	Variable description	Variable derivation details
Sweep 1 (age 0)	Father's social status	Derived from a0014: 1 Social class 1 / 2 Social class 2 / 3 Social class 3 (Non-Manual) / 4 Social class 3 (Manual) / 5 Social class 4 / 6 Social class 5 / Other and unsupported (as missing)
	Country of Birth	Recoded from COB England / Wales / Scotland / Other
	Age of mother at first birth	Derived from a0005a and a0052 Per year
	Method of contraception	a0029b None / Pill alone / Pill & Other method / Other Method
	Parity	Recoded from A0166 0 / 1 / 2 / 3 / >4
	Certainty of last menstrual period	a0196 Certain vs uncertain
	Where was mother-date of birth	Recoded from A0357 Consultant bed/ GP bed / NHS unit / Own home / Other
Sweep 2 (age 5)	Type of accommodation	Recoded from E218 Detached / Semi-detached / Terrace / Flat-Maisonette / Rooms-other
	Ethnic group (father)	Recoded from E246b European UK / European EU / Other
	Neighbourhood group	e267b Poor / Average / Well to do / Rural
Sweep 3 (age 10)	Accommodation occupied by family	Recoded from D1_1l

Sweep	Variable description	Variable derivation details
		Flat / House / Other
	Estimated reading age (in years)	bd3rdage Continuous (per year)
	British Ability Scales Matrix - Total score	BASmatrx Continuous (per unit)
Non-response at a previous sweep (i.e. sweeps 2 - 3)		Yes vs no

Table B4. Predictors of non-response at Sweep 5 (age 26).

Sweep	Variable description	Variable derivation details
Sweep 1 (age 0)	Father's social status	Derived from a0014: 1 Social class 1 / 2 Social class 2 / 3 Social class 3 (Non-Manual) / 4 Social class 3 (Manual) / 5 Social class 4 / 6 Social class 5 / Other and unsupported (as missing)
	Age of mother at 1st birth	Derived from a0005a and a0052 Continuous (per year)
	Method of contraception	a0029b None / Pill alone / Pill & Other method / Other Method
	Parity (i.e. number of older siblings)	Recoded from a0166 0 / 1 / 2 / 3 / >4
	Certainty of last menstrual period	a0196 Certain vs uncertain
	Sex	SEX Female vs male
Sweep 2 (age 5)	External score	Extern_score_5 Summary of rutter conduct and hyperactivity items: d025 d026 d027 d028 d032 d034 d038 d039 d042 d043 Per unit increase
	Harris scoring method	f114 Per unit increase
Sweep 3 (age 10)	Gross family income	grfaminc Per unit increase
	Teacher Rutter assessment	B3T_Rutt

Sweep	Variable description	Variable derivation details
		Summary of teacher rutter items: j151 j163 j160 j122 j149 j159 j156 j171 j128 j146 j169 j172 j175 Per unit increase
	Number of household accessories	b3hldstf Summary of: m291 m292 m293 m294 m295 m296 m297 m298 m299 m300 m301 m302 Per unit increase
Sweep 4 (age 16)	Satisfaction with teen's school progress	Pb3_1 Very satisfied / Fairly satisfied / Not satisfied
Non-response at a previous sweep (i.e. sweeps 2 - 4)		Yes vs no

Table B5. Predictors of non-response at Sweep 6 (age 30).

Sweep	Variable description	Variable derivation details
Sweep 1 (age 0)	Father's social status	Derived from a0014: 1 Social class 1 / 2 Social class 2 / 3 Social class 3 (Non-Manual) / 4 Social class 3 (Manual) / 5 Social class 4 / 6 Social class 5 / Other and unsupported (as missing)
	Sex	SEX Female vs male
	Number of antenatal Visits	a0190 Per visit
	Parity (i.e. number of older siblings)	Recoded from A0166 0 / 1 / 2 / 3 / >4
Sweep 2 (age 5)	Ethnic group (mother)	Recoded from E246a Non-european vs european
	Household moves	Recoded from e249 0 / 1 / 2 / 3 / 4 / >5
	Neighbourhood group	e267b Poor / Average / Well to do / Rural
Sweep 3 (age 10)	Score BAS Matrices	BASmatrx Per move
Sweep 4 (age 16)		
Sweep 5 (age 26)	How easy would you quit a job, if there was no other job to go to (range: 1-5)	B960127_new Continuous (per unit)
	N of accidents elsewhere	Recoded from aelse 0 / 1 / ≥2

Sweep	Variable description	Variable derivation details
	Tenure of current address	Recoded from B960421 Own/ Parents (rent-free) / Buying on mortgage / Rent / Other
Non-response at a previous sweep (i.e. sweeps 2-5)		Yes vs no

Table B6. Predictors of non-response at Sweep 7 (age 34).

Sweep	Variable description	Variable derivation details
Sweep 1 (age 0)	Father's social status	Derived from a0014: 1 Social class 1 / 2 Social class 2 / 3 Social class 3 (Non-Manual) / 4 Social class 3 (Manual) / 5 Social class 4 / 6 Social class 5 / Other and unsupported (as missing)
	Sex	SEX Female vs male
	Ever a teenager mother	BD1_TEENM 0-19 / 20 plus
	Parity (i.e. number of older siblings)	Recoded from a0166 0 / 1 / 2 / 3 / >4
	Present marital status	a0012 Single / Married / Widowed,Divorced,Separated
	Method of contraception	a0029b None / Pill alone / Pill & Other method / Other Method
Sweep 2 (age 5)	Household moves	Recoded from e249 0 / 1 / 2 / 3 / 4 / >5
	Number of household accessories	Derived from: e229 e230 e231 e232 e233 e234 e235 Per accessory
Sweep 3 (age 10)	Maths Test Score	BD3MATHS Per mark
Sweep 4 (age 16)	Maths 0 level or equivalent	T2c1_1 No vs yes
Sweep 5 (age 26)	Satisfaction about life (1-3) ordered	Recoded from B960666

Sweep	Variable description	Variable derivation details
		1-->0-5 / 2-->6-7 / 3-->8-9
Sweep 6 (age 30)	Does participant intend to move?	Recoded from Wantmove (8,9 turned to missing) No vs yes
	Did you vote in the 97 elections?	Recoded from vote97 (8,9 turned to missing) No vs yes
	Do you like the area? (Units 1-5)	Recoded from likearea (8,9 turned to missing) 5 item scale: 1 very satisfied – 5 very dissatisfied
Non-response at a previous sweep (i.e. sweeps 2-6)		Yes vs no

Table B7. Predictors of non-response at Sweep 8 (age 38).

Sweep	Variable description	Variable derivation details
Sweep 1 (age 0)	Father's social status	Derived from a0014: 1 Social class 1 / 2 Social class 2 / 3 Social class 3 (Non-Manual) / 4 Social class 3 (Manual) / 5 Social class 4 / 6 Social class 5 / Other and unsupported (as missing)
	Sex	SEX Female vs male
	N of antenatal visits	a0190 Continuous (per visit)
	Parity (i.e. number of older siblings)	Recoded from a0166 0 / 1 / 2 / 3 / >4
	Certainty of Data of last Menstrual Period	a0196 Certain vs uncertain
	Method of contraception	a0029b None / Pill alone / Pill & Other method / Other method
Sweep 2 (age 5)	Household moves	Recoded from e249 0 / 1 / 2 / 3 / 4 / >5
	Copying design score	f119 Per unit
Sweep 3 (age 10)	Maths Test Score	BD3MATHS Per unit increase
	Score BAS Matrices	BASMATRIX Per unit increase

Sweep	Variable description	Variable derivation details
Sweep 5 (age 26)	Age left education	Recoded from B960132 12-16 / 17-18 / 19 plus
Sweep 6 (age 30)	Does participant intend to have more children	Recoded from infertlc (9-->missing) Yes / No / Don't know
	Financial situation	Recoded from finnow (8, 9 -->missing) Living comfortably / Doing alright / Just about getting by / Finding it quite difficult / Finding it very difficult
	Voted in general elections 1997?	Recoded from vote97 (8, 9 -->missing) No vs yes
Sweep 7 (age 34)		
Non-response at a previous sweep (i.e. sweeps 2-8)		Yes vs no

Table B8. Predictors of non-response at Sweep 9 (age 42).

Sweep	Variable description	Variable derivation details
Sweep 1 (age 0)	Father's social status	Derived from a0014: 1 Social class 1 / 2 Social class 2 / 3 Social class 3 (Non-Manual) / 4 Social class 3 (Manual) / 5 Social class 4 / 6 Social class 5 / Other and unsupported (as missing)
	Country of Birth	Recoded from COB England / Wales / Scotland / Other
	Sex	SEX Female vs male
	N of antenatal visits	a0190 Continuous (per visit)
	Certainty of Data of last Menstrual Period	a0196 Certain vs uncertain
Sweep 2 (age 5)	Copying design score	f119 Continuous (Per unit)
Sweep 3 (age 10)	Score BAS Matrices	BASmmatrx Per mark
Sweep 6 (age 30)	Voted in general elections 1997?	Recoded from vote97 (8, 9 -->missing) No vs yes
Sweep 7 (age 34)	Work overtime?	b7otimny No vs Yes
	Should everyone behave responsibly?	B7_resp No vs Yes

Sweep	Variable description	Variable derivation details
	How many times have you been found guilty in a criminal court?	Recoded from B7court (8, 9 -->missing) ≥1 vs 0
Sweep 8 (age 38)	Participant willing to be contacted for parents' research project?	b8parent No vs yes
Non-response at a previous sweep (i.e. sweeps 2-8)		Yes vs no

Table B9. Predictors of non-response at Sweep 10 (age 46).

Sweep	Variable description	Variable derivation details
Sweep 1 (age 0)	Father's social status	Derived from a0014: 1 Social class 1 / 2 Social class 2 / 3 Social class 3 (Non-Manual) / 4 Social class 3 (Manual) / 5 Social class 4 / 6 Social class 5 / Other and unsupported (as missing)
	Country of Birth	Recoded from COB England / Wales / Scotland / Other
	Parity (i.e. number of older siblings)	Recoded from a0166 0 / 1 / 2 / 3 / >4
	Certainty of last menstrual period	a0196 Certain vs uncertain
	Was lactation attempted	a0297 Not attempted vs attempted
	Number of antenatal visits	a0190 Per visit
Sweep 2 (age 5)	Copying designs score	f119 Continuous (Per unit)
Sweep 3 (age 10)	Score BAS Matrices	BASmatrx Continuous (Per unit)
	Accommodation	Recoded from d2 Owned / Bought / Council rented / Other rented / Tied to occupation
Sweep 6 (age 30)	Had eczema or skin problems?	Recoded from othskin (8,9-->missing) No vs yes
	Voted in general elections 1997?	Recoded from vote97 (8, 9 -->missing)

Sweep	Variable description	Variable derivation details
		No vs yes
Sweep 7 (age 34)	Is this address participant's residence?	b7nrmal No vs yes
Sweep 8 (age 38)	Willing to be contacted for Parents Research Project	b8parent No vs Yes
	Any children aged 0-6	b8chd006 No vs Yes
Sweep 9 (age 42)	Total score (vocabulary)	B9VSCORE Per mark
Non-response at a previous sweep (i.e. sweeps 2-9)		Yes vs no

Appendix C: Identified predictors of non-response in Next Steps.

Sweep 1 (age 14) to Sweep 7 (age 20) NCDS data are available from the UK Data Service:

"Next Steps: Sweeps 1-8, 2004-2016" (38)

Stata code that derives a dataset with all the variables predicting non-response in the Next Steps listed below, is available on the CLS Data GitHub.

Table C1. Predictors of non-response at Sweep 8 (age 25).

Sweep	Variable description	Variable derivation details
Sweep 1 (age 14)	Sex of the young person	W1sexYP Coded: “Female” / “Male”
	How often the young person’s parents know where they going when they go out in the evening	W1gowhereYP Coded: “Always” / “Usually” / “Sometimes-never” / “Don't go out in the evening”
	Whether the young person has been upset by name-calling, including by text or email, in the last 12 months	W1namesYP Coded: “No” / “Yes”
	Days per week the young person uses a home computer to play games	W1HcomGYP Coded: “None” / “1 - 2 days” / “3 - 4 days” / “Most days (5 or more)”
	Whether the young person has played a musical instrument in the last 4 weeks	W1alei1YP0h Coded: “No” / “Yes”
	Housing tenure	W1hous12HH Coded: “Owned outright” / “Being bought on a mortgage/ bank loan” / “Rented/other”
	Whether the young person can access the internet from home	W1condur6MP Coded: “No” / “Yes”
Sweep 2 (age 15)	Whether the young person’s school have ever contacted their parents about their behaviour	W2SchcontMP Coded: “No” / “Yes”

Sweep	Variable description	Variable derivation details
	How much constantly under strain the young person has felt recently	W2strainYP Coded: “Not at all” / “No more than usual” / “Rather more than usual” / “Much more than usual”
Sweep 3 (age 16)	Whether the young person ever smokes cigarettes	W3cignowYP Coded: “No” / “Yes”
	Age of the young person’s main parent	W3ageMP Continuous
Sweep 4 (age 17)	How often the young person goes to nightclubs	W4ClubberYP Coded: “Once a week or more” / “Less than once a week” / “Hardly ever” / “Never”
	Whether the young person gives their permission to pass on their details to the Department for Work and Pensions	W4DWPLnk2YP Coded: “No” / “Yes”
Sweep 5 (age 18)	Whether the young person still lives at the same address as the previous interview	W5SameaddHH Coded: “No” / “Yes”
	Whether there are specific groups of people that the young person feels are usually treated better by the government than people like them	W5FairTreat2YP Coded: “No” / “Yes”
	How well the young person thought their teachers in Year 11 and earlier expected them to do in their exams	W5TeaExpecYP Coded: “Better than most pupils in their year group” / “As well as most pupils in their year group” / “Less well than most pupils in their year group”

Sweep	Variable description	Variable derivation details
	Current main activity of the young person	W5actYP Coded: “Full-time Education” / “Working or part working and part college” / “Other”
Sweep 6 (age 19)	Whether the young person has spoken to a teacher for information, advice and guidance about the future	W6IagIntroYP0b Coded: “No” / “Yes”
	Whether the young person is willing to answer questions on sexual experiences	W6SexIntroYP Coded: “No” / “Yes”
Sweep 7 (age 20)	Whether the young person is willing to answer questions on sexual experiences	W7SexIntroYP Coded: “No” / “Yes”
	Previous non-response (waves 2-7)	Coded: “Complete response up to and including wave 7” / “One or more instances of non-response” Derived using survey outcome variables for each survey: W2OUTCOME W3OUTCOME W4OUTCOME W5OUTCOME W6OUTCOME W7OUTCOME.

Appendix D: Identified predictors of non-response in the Millennium Cohort Study (MCS).

Sweep 1 (age 9 months) to Sweep 7 (age 17 years) MCS data are available from the UK Data Service:

Millennium Cohort Study: Age 9 months, Sweep 1, 2001 (39)

Millennium Cohort Study: Age 3, Sweep 2, 2004 (40)

Millennium Cohort Study: Age 5, Sweep 3, 2006 (41)

Millennium Cohort Study: Age 7, Sweep 4, 2008 (42)

Millennium Cohort Study: Age 11, Sweep 5, 2012 (43)

Millennium Cohort Study: Age 14, Sweep 6, 2015 (44)

Millennium Cohort Study: Age 17, Sweep 7, 2018 (45)

Stata code that derives a dataset with all the variables predicting non-response in the MCS listed below is available on the CLS Data GitHub.

Table D1. Predictors of non-response at Sweep 2 (age 3).

Sweep	Variable	Variable derivation details
Sweep 1 (age 9 months)	Whether voted in General Election (in last year / 2002)	Derived from: APVOTE00 Coded: “Yes” /” no”
	Age of baby when last breastfed	Derived from ACBFEV00 ACBFEA00 ACBFED00 ACBFEW00 ACBFEM00 ACBAGE00 Coded: “Never breastfed” / “Up to 1 month old” / “1-3 months old” / “3-6 months old” / “6-9 months old” / “9 months or older”
	Highest social class in household	Derived from ADD05C00, using ARESP00 to identify main parent or partner. Coded: “Managerial and professional” / “Intermediate” / “Small employer and self-employed” / “Low supervisory and technical” / “Semi-routine and routine”
	Paid work status	Derived from: ADWKST00 Coded: “Currently doing paid work or on leave” / “Past job, but not current” / “Never had paid job”
	Respondent’s age at birth of cohort member	ADDAGB00

Note: Original or recoded response categories for binary and categorical variables are reported, but not for continuous variables.

Table D2. Predictors of non-response at Sweep 3 (age 5).

Sweep	Variable	Variable derivation details
Sweep 1 (age 9 months)	Whether has a Red Book	ACREBK00 Coded: “Yes” / “no”
	Types of accounts held	Derived from: APBANK0A Coded: “Yes (bank, building society, Post Office account, or other)” / “No (no account)”
	Whether voted in General Election (in last year / 2002)	Derived from: APVOTE00 Coded: “Yes” /” no”
	Age of baby when last breastfed	Derived from ACBFEV00 ACBFEA00 ACBFED00 ACBFEW00 ACBFEM00 ACBAGE00 Coded: “Never breastfed” / “Up to 1 month old” / “1-3 months old” / “3-6 months old” / “6-9 months old” / “9 months or older”
	When child first had cow's milk	Derived from: ACAGCM00 ACCMDA00 ACCMWK00 ACCMMT00 ACCMMT00 ACADMA00 ACBAGE00 Coded: “Never had” / “Less than 6 months old” / “6 months or older”
	Paid work status	Derived from: ADWKST00 Coded: “Currently doing paid work or on leave” / “Past job, but not current” / “Never had paid job”

Sweep	Variable	Variable derivation details
	Number of hospital admissions	ACADMA00
	Baby's age in months	ACBAGE00
Sweep 2 (age 3 years)	OECD equivalised income	BDOEDE00
	Respondent's religion	Derived from: BDDRLG00 Coded: No Religion" / "Christian" / "Muslim" / "Other"

Note: Original or recoded response categories for binary and categorical variables are reported, but not for continuous variables.

Table D3. Predictors of non-response at Sweep 4 (age 7).

Sweep	Variable	Variable derivation details
Sweep 1 (age 9 months)	Whether voted in General Election (in last year / 2002)	Derived from: APVOTE00 Coded: “Yes” /” no”
	Age of baby when last breastfed	Derived from ACBFEV00 ACBFEA00 ACBFED00 ACBFEW00 ACBFEM00 ACBAGE00 Coded: “Never breastfed” / “Up to 1 month old” / “1-3 months old” / “3-6 months old” / “6-9 months old” / “9 months or older”
	Paid work status	Derived from: ADWKST00 Coded: “Currently doing paid work or on leave” / “Past job, but not current” / “Never had paid job”
	Baby’s age in months	ACBAGE00
	Highest social class in household	Derived from ADD05C00, using ARESP00 to identify main parent or partner. Coded: “Managerial and professional” / “Intermediate” / “Small employer and self-employed” / “Low supervisory and technical” / “Semi-routine and routine”
	Respondent's Ethnic Group	ADD06E00 Coded: “White” / “Mixed” / “Indian” / “Pakistani/Bangladeshi” / “Black or Black British” / “Other”
Sweep 3 (age 5 years)	Whether voted in General Election (in 2005)	CPVOTE00 Coded: “Yes” / “no”

Sweep	Variable	Variable derivation details
	Cognitive test: Pattern Construction T-score	CCPCTSCORE
Sweep 2 (age 3 years) Sweep 3 (age 5 years)	Non-response to any prior survey sweeps	Derived from: BAOUTC00 CAOUTC00 Coded: "Yes, prior non-response" / "No, no prior non-response"

Note: Original or recoded response categories for binary and categorical variables are reported, but not for continuous variables.

Table D4. Predictors of non-response at Sweep 5 (age 11).

Sweep	Variable	Variable derivation details
Sweep 1 (age 9 months)	Whether voted in General Election (in last year / 2002)	Derived from: APVOTE00 Coded: “Yes” /” no”
	Support sought since birth MC1	APSUPP0A Coded: “No none of these” / “GP, health visitor, other”
	Age of baby when last breastfed	Derived from ACBFEV00 ACBFEA00 ACBFED00 ACBFEW00 ACBFEM00 ACBAGE00 Coded: “Never breastfed” / “Up to 1 month old” / “1-3 months old” / “3-6 months old” / “6-9 months old” / “9 months or older”
	Paid work status	Derived from: ADWKST00 Coded: “Currently doing paid work or on leave” / “Past job, but not current” / “Never had paid job”
	Highest social class in household	Derived from ADD05C00, using ARESP00 to identify main parent or partner. Coded: “Managerial and professional” / “Intermediate” / “Small employer and self-employed” / “Low supervisory and technical” / “Semi-routine and routine”
	Baby’s age in months	ACBAGE00
Sweep 3 (age 5 years)	Whether voted in General Election (in 2005)	CPVOTE00 Coded: “Yes” / “no”

Sweep	Variable	Variable derivation details
	Cognitive test: Pattern Construction T-score	CCPCTSCORE
Sweep 4 (age 7 years)	Academic performance – teacher reported	Derived from: DQ2160 DQ2162 DQ2164 DQ2166 DQ2167 DQ2168 DQ2169 DQ2170. Items with valid responses are summed.
Sweep 2 (age 3 years) Sweep 3 (age 5 years) Sweep 4 (age 7 years)	Non-response to any prior survey sweeps	Derived from: BAOUTC00 CAOUTC00 DAOUTC00 Coded: "Yes, prior non-response" / "No, no prior non-response"

Note: Original or recoded response categories for binary and categorical variables are reported, but not for continuous variables.

Table D5. Predictors of non-response at Sweep 6 (age 14).

Sweep	Variable	Variable derivation details
Sweep 1 (age 9 months)	Any other sources of regular income	APREPA00 Coded: “Yes” /” no”
	Whether voted in General Election (in last year / 2002)	Derived from: APVOTE00 Coded: “Yes” /” no”
	Support sought since birth MC1	APSUPP0A Coded: “No none of these” / “GP, health visitor, other”
	Age respondent left full-time education	APLFTE00 Coded: “Under 16” / “Age 16” / “Age 17” / “Age 18/19” / “Age 20 or above”
	Respondent's Ethnic Group	ADD06E00 Coded: “White” / “Mixed” / “Indian” / “Pakistani/Bangladeshi” / “Black or Black British” / “Other”
	Baby’s age in months	ACBAGE00
	Attachment to Cohort Member	Derived from: APPCRY00 APFPAT00 APSTIM00 APTALK00 APCUDL00 APANNO00 APTHNK00 APLEAV00 APCOMP00 APPATI00 APGIUP00. Reverse code: APPCRY00 APFPAT00 APSTIM00 APTALK00 APCUDL00 APTHNK00 APLEAV00. Then summing all items.

Sweep	Variable	Variable derivation details
Sweep 2 (age 3 years)	Epstein-Barr virus panel (count of antibodies representing immunity)	EBVPANEL
Sweep 3 (age 5 years)	Whether voted in General Election (in 2005)	CPVOTE00 Coded: “Yes” / “no”
	Cognitive test: Pattern Construction T-score	CCPCTSCORE
	Frequency of CM’s physical activities with family	CPFAPA00
Sweep 2 (age 3 years) Sweep 3 (age 5 years) Sweep 4 (age 7 years) Sweep 5 (age 11 years)	Non-response to any prior survey sweeps	Derived from: BAOUTC00 CAOUTC00 DAOUTC00 EAOUTC00. Coded: "Yes, prior non-response" / "No, no prior non-response"

Note: Original or recoded response categories for binary and categorical variables are reported, but not for continuous variables.

Table D6. Predictors of non-response at Sweep 7 (age 17).

Sweep	Variable	Variable derivation details
Sweep 1 (age 9 months)	Whether voted in General Election (in last year / 2002)	Derived from: APVOTE00 Coded: “Yes” /” no”
	Type of accommodation	APMOTY00 Coded: “House or bungalow” / “Flat, maisonette, studio, room, etc”
	Age left full-time education	APLFTE00 Coded: “Under 16” / “Age 16” / “Age 17” / “Age 18/19” / “Age 20 or above”
	Baby’s age in months	ACBAGE00
	Highest social class in household	Derived from ADD05C00, using ARESP00 to identify main parent or partner. Coded: “Managerial and professional” / “Intermediate” / “Small employer and self-employed” / “Low supervisory and technical” / “Semi-routine and routine”
Sweep 3 (age 5 years)	Whether voted in General Election (in 2005)	CPVOTE00 Coded: “Yes” / “no”
	Cognitive test: Pattern Construction T-score	CCPCTSCORE
	Cohort member’s age at interview (days)	CHCAGE00
	Number of rooms - grouped	CPROMA00

Sweep	Variable	Variable derivation details
		10 rooms or more is combined, but variable is treated as continuous.
Sweep 4 (age 7 years)	Household dog owned	Derived from: DPPETH0A DPPETH0B DPPETH0C DPPETH0D DPPETH0E Coded: “Yes” / “no”
	Aspiration goal 1 (Extrinsic, Intrinsic or Otherwise)	DCAEXT0A Coded: “Extrinsic” / “Extrinsic-Intermediate” / “Intrinsic-Intermediate”
Sweep 6 (age 14 years)	School year	Derived from: FPSTSC00 FPDIFY00 Coded: “Year9/(S3)/Year10 or Year8/(S2)/Year9” / “Year 10/S4/Year11”
	Consumer Attitude Grid: I like clothing with popular labels	FCPLAB00 Coded: “Strongly Agree” / “Agree” / “Disagree” / “Strongly disagree”
	Respondent’s religion - 7 category	FDRLG00 Coded: “No Religion” / “Christian” / “Muslim” / “Other”
Sweep 2 (age 3 years) Sweep 3 (age 5 years) Sweep 4 (age 7 years) Sweep 5 (age 11 years) Sweep 6 (age 14 years)	Non-response to any prior survey sweeps	Derived from: BAOUTC00 CAOUTC00 DAOUTC00 EAOUTC00 FAOUTC00 Coded: "Yes, prior non-response" / "No, no prior non-response"

Note: Original or recoded response categories for binary and categorical variables are reported, but not for continuous variables.