Institute of Education

UCL

Handling missing data in the National Child Development Study

User guide

April 2020

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contact

Questions and feedback about this user guide should be sent to clsfeedback@ucl.ac.uk.

How to cite this guide

Silverwood, R., Narayanan, M., Dodgeon, B., Ploubidis, G. (2020) *Handling missing data in the National Child Development Study: User guide*. London: UCL Centre for Longitudinal Studies.

This guide was first published in April 2020 by the UCL Centre for Longitudinal Studies.

UCL Institute of Education

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 based at the UCL Institution of Education (IOE), University College London. It manages four internationally-renowned cohort studies: the 1958 National Child Development Study, the 1970 British Cohort Study, Next Steps, and the Millennium Cohort Study. For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies.

tel: +44 (0)20 7612 6875

email: clsfeedback@uck.ac.uk

Table of Contents

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].

NCDS 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 National Child Development Survey (NCDS; 1958 British Birth Cohort) we have implemented a systematic data-driven approach to identify predictors of non-response at Sweeps 1-9 [10]. We found disadvantaged socio-economic background in childhood, worse mental health and lower cognitive ability in early life, and lack of civic and social participation in adulthood to be consistently associated with non-response. A full list of the identified predictors of non-response at each sweep is available in the Appendix. For details of the approach used in identifying these variables we refer you to the published paper [10].

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 restore sample representativeness in NCDS.

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 most common among NCDS analyses (>80% of papers). 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 NCDS. 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, 11, 12].

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 NCDS analyses 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 using NCDS data in Stata according to the NCDS Missing Data Strategy. 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 NCDS data.

Running example

The application of the NCDS Missing Data Strategy will be illustrated through the use of a running example examining the association between partnership status at age 42 (ParStat42R) and income at age 55 (total_income55)[13]. 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	DiffucultiesFinancial	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	DiffucultiesHousing	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 DiffucultiesFinancial
  HousingTenure_7 DiffucultiesHousing 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 [14]), 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
  DiffucultiesFinancial HousingTenure_7 DiffucultiesHousing 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 DiffucultiesHousing DiffucultiesFinancial 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 DiffucultiesHousing DiffucultiesFinancial 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. (In our example, variables that are associated with non-response and income, as the association with income will increase the likelihood of association with missing values of income.)
- Variables that are predictive of the underlying missing values only. (In our example, variables associated with income.)

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). 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.

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 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 10 of the Appendix. 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, 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 [15], 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
- 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 [14].

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 DiffucultiesFinancial HousingTenure_7
  DiffucultiesHousing 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 [16]. 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 p-values with little error requires a greater number of imputations (perhaps 100 or more) [5, 17]. 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 DiffucultiesHousing DiffucultiesFinancial 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.

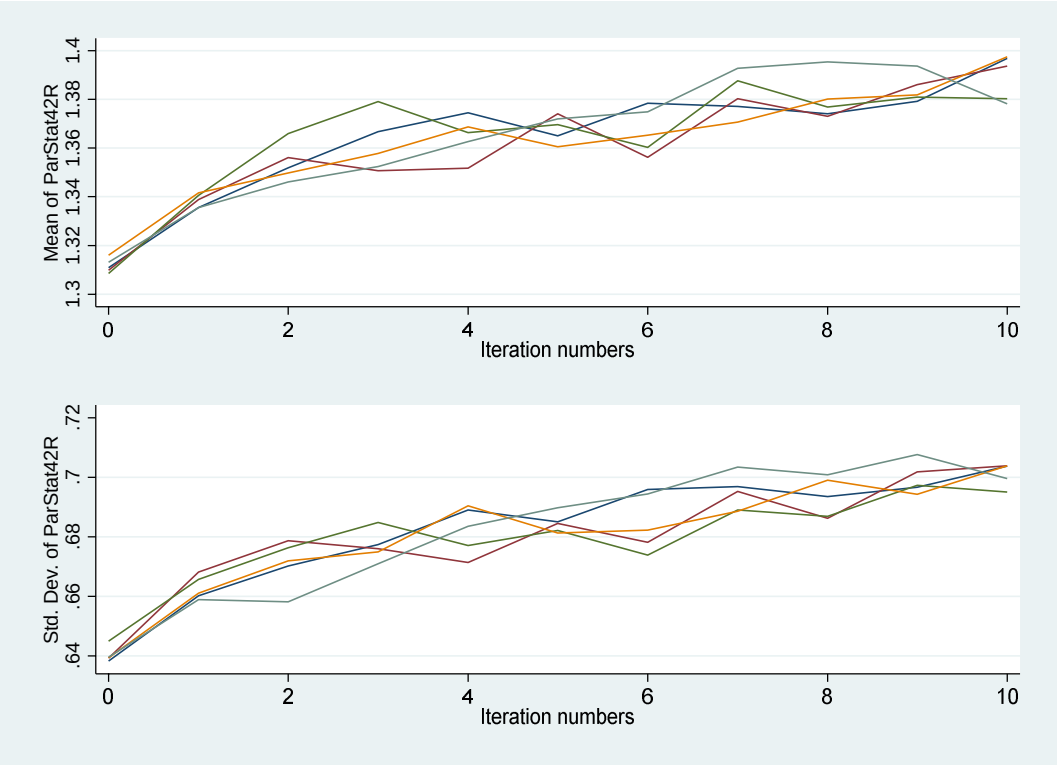
Example

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 DiffucultiesHousing DiffucultiesFinancial 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:

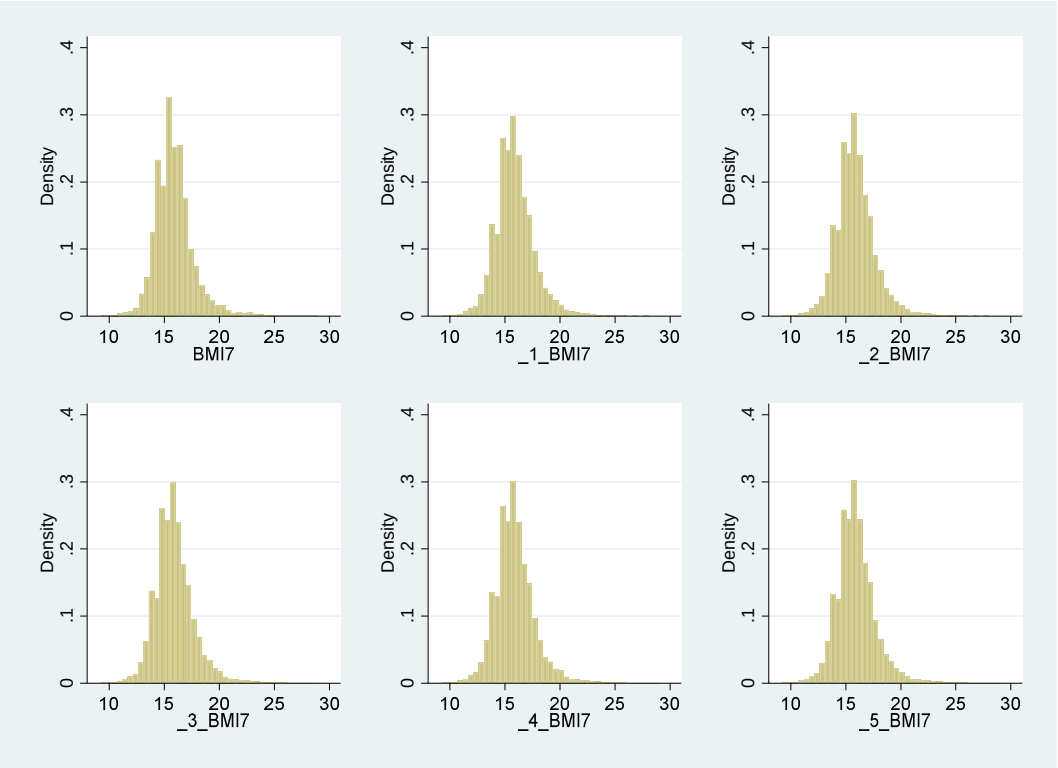


The below code will loop through each continuous variable in the imputation model in turn, producing histograms of the values in the observed data and then in each imputed dataset separately, before combining them at the end. The `pause` command is used as above.

```
. 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 DiffucultiesHousing DiffucultiesFinancial 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 [18]. 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 DiffucultiesHousing DiffucultiesFinancial 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 those 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 those 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 DiffucultiesHousing DiffucultiesFinancial 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 those 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 NCDS analyses 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 fitting 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 Sweeps 1-10 in the Appendix).

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.

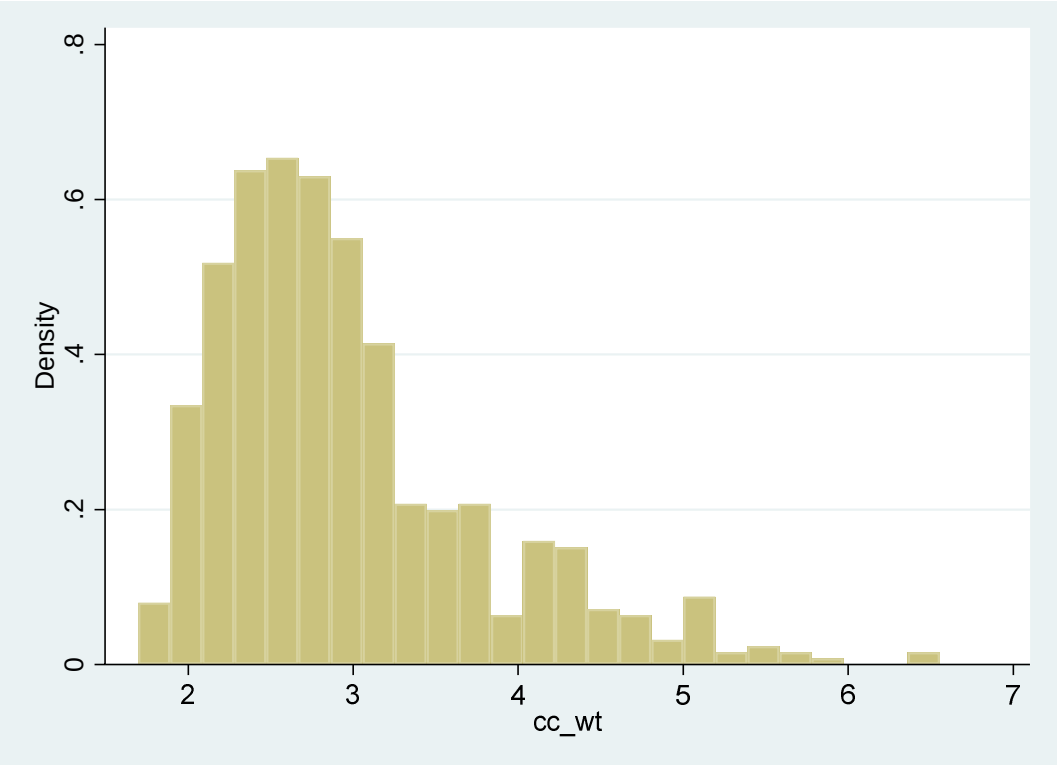
```
. quietly regress log_total_income55 i.ParStat42R n622 n553 BMI7 CogAbil7 MedExSum7
  a16_totalscore  Log_total_income42 dmdisab lsiany2 SepMore1Month MotherNeverReads7 NoIntEdu
  smpreg maw5 HousingTenure_7 DiffucultiesHousing DiffucultiesFinancial DivBy7 LBW bfever
  enuresis7 MumNotMarried PsychoMed 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_wt2 |        645    2.967592    .8079965   1.695903   6.555767

. hist cc_wt
```



The resultant weights vary between 1.7 and 6.6, with no 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 DiffucultiesHousing DiffucultiesFinancial 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

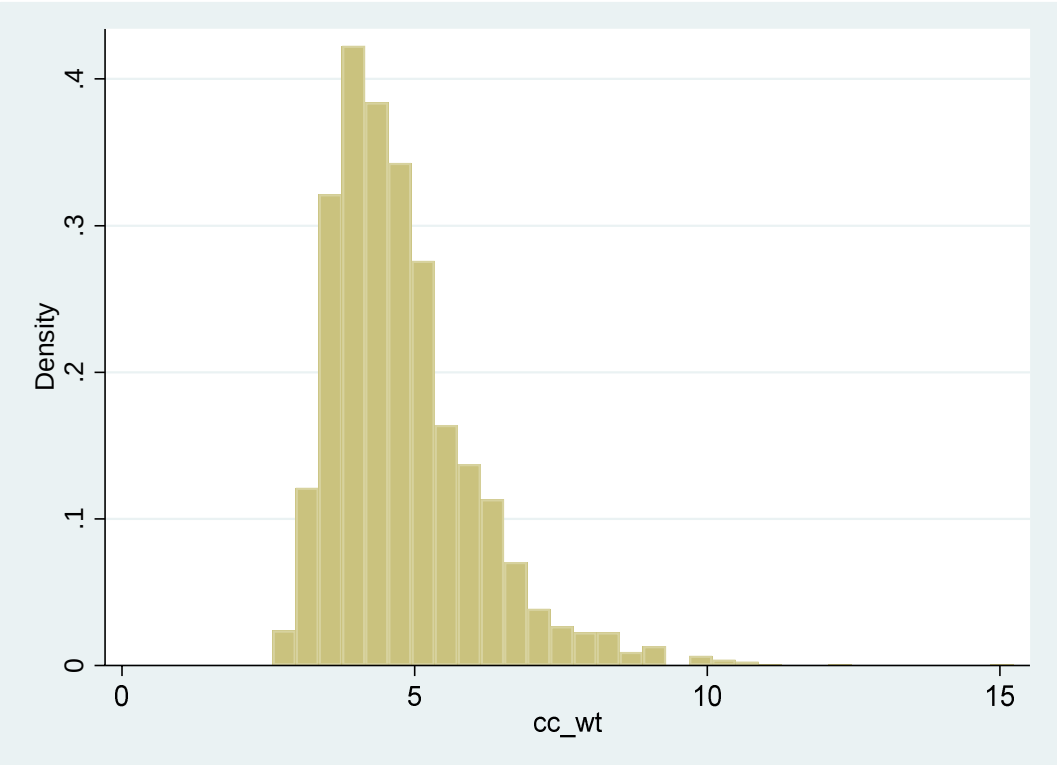
used as predictors of being a complete case (here, the 21 predictors of non-response at age 55).

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
```



The resultant weights vary between 2.6 and 15.2, with no extreme values. 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 DiffucultiesHousing DiffucultiesFinancial 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 comparable. 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, Kenward MG. Multiple Imputation and its Application. Chichester, UK: John Wiley & Sons, Ltd; 2013.
- 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, D.B., *Inference and missing data.* Biometrika, 1976. **63**: p. 581–592.
2. Little, R.J.A. and D.B. Rubin, *Statistical Analysis with Missing Data. Third Edition*. 2020, Hoboken, NJ: Wiley.
3. White, I.R. and J.B. Carlin, *Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values.* Statistics in Medicine, 2010. **29**(28): p. 2920-2931.
4. Little, R.J.A. and D.B. Rubin, *The analysis of social-science data with missing values.* Sociological Methods & Research, 1989. **18**(2-3): p. 292-326.
5. Carpenter, J.R. and M.G. Kenward, *Multiple Imputation and its Application*. 2013, Chichester, UK: John Wiley & Sons, Ltd.
6. Wooldridge, J.M., *Inverse probability weighted estimation for general missing data problems.* Journal of Econometrics, 2007. **141**(2): p. 1281-1301.
7. Seaman, S.R. and I.R. White, *Review of inverse probability weighting for dealing with missing data.* Stat Methods Med Res, 2013. **22**(3): p. 278-95.
8. Enders, C.K., *Applied missing data analysis*. 2010, New York: Guilford.
9. Enders, C.K., *The Performance of the Full Information Maximum Likelihood Estimator in Multiple Regression Models with Missing Data.* Educational and Psychological Measurement, 2001. **61**(5): p. 713-740.
10. Mostafa, T., et al., *Understanding non-response in the 1958 British birth cohort: A data driven approach.* Under review.
11. Carpenter, J.R., H. Goldstein, and M.G. Kenward, *REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types.* 2011, 2011. **45**(5): p. 14.
12. Quartagno, M., S. Grund, and J. Carpenter, *jomo: A Flexible Package for Two-level Joint Modelling Multiple Imputation.* The R Journal, In press.
13. 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.
14. Harel, O., et al., *Multiple Imputation for Incomplete Data in Epidemiologic Studies.* Am J Epidemiol, 2018. **187**(3): p. 576-584.
15. Tibshirani, R., *Regression Shrinkage and Selection Via the Lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-288.
16. Wagstaff, D.A., S. Kranz, and O. Harel, *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): p. 1025-30.
17. Graham, J.W., A.E. Olchowski, and T.D. Gilreath, *How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory.* Prevention Science, 2007. **8**(3): p. 206-213.
18. Von Hippel, P.T., *Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data.* Sociological Methodology, 2007. **37**(1): p. 83-117.

Appendix

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

Sweep	Variable description	Derived variable name	Original variable name
Sweep 0 (age 0)	Region	acatnn0region	n0region
	Number of persons per room	aconnn512	n512
	Social class of mother's husband	acatnn236	n236

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

Sweep	Variable description	Derived variable name	Original variable name
Sweep 0 (age 0)	Mother's present marital status	abinnn545	n545
Sweep 1 (age 7)	Number of kids under 21 in the household, including living away	bconnn99	n99
	Common difficulties age 7 (mother)	bconnage7dv1	age7dv1
	Hospital admissions	bconnage7dv5	age7dv5
	Cognitive ability summary	CogAbil7	CogAbil7
	Non-response at sweep 1	NR01priorNR	OUTCME01

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

Sweep	Variable description	Derived variable name	Original variable name
Sweep 0 (age 0)	Region	acatnn0region	n0region
Sweep 1 (age 7)	Number of kids under 21 in the household, including living away	bconnn99	n99
	Mother worked birth to 5	maw5	maw5 (from n197/n198)
	Ever breastfed	bfever	n222
Sweep 2 (age 11)	Non-response at sweeps 1-2	NR02priorNR	OUTCME01-OUTCME02

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

Sweep	Variable description	Derived variable name	Original variable name
Sweep 0 (age 0)	Region	acatnn0region	n0region
	Number of persons per room	aconnn512	n512
	Sex of child	bingender	n622
	Social class of mother's husband	acatnn236	n236
Sweep 1 (age 7)	Family moves since child's birth	bcatnn95	n95
	Cognitive ability summary	CogAbil7	CogAbil7
	Dad reads to child	DadNeverReads	n180
Sweep 2 (age 11)	Area of world in which mother born	ccatnn1434	n1434
	Number of family moves since child's birth	ccatnn1150	n1150
	Cognitive ability summary	genability11	genability11
	Number of household amenities	Amens	Amens
Sweep 3 (age 16)	Number of family moves since child's birth	dconnn2492	n2492
	Sum of favourable learning environments/outcomes re sex educ etc)	dconnage16dv46	age16dv46
	Conduct problems	Ext16MTDec18	Ext16MTDec18
	Non-response at sweeps 1-3	NR03priorNR	OUTCME01-OUTCME03

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

Sweep	Variable description	Derived variable name	Original variable name
Sweep 0 (age 0)	Number of persons per room	aconnn512	n512
	Sex of child	Bingender	n622
	Social class of mother's husband	acatnn236	n236
Sweep 1 (age 7)	Family moves since child's birth	bcatnn95	n95
	Social problems (alcoholism etc.)	bconnage7dv10	age7dv10
	Cognitive ability summary	CogAbil7	CogAbil7
	Summary of medical conditions	MedExSum7	MedExSum7
	Ever breastfed	bfever	n222
Sweep 2 (age 11)	Child's positive activities outside school	cconnage11dv32	age11dv32
	Cognitive ability summary	genability11	genability11
	Number of household amenities	Amens	Amens
Sweep 3 (age 16)	Number of family moves since child's birth	dconnn2492	n2492
	How long since child drank alcohol	dcatnn2888	n2888
	Test 2 – mathematics comprehension	dconnn2930	n2930
	Sum of favourable learning environments/outcomes re sex educ etc)	dconnage16dv46	age16dv46
Sweep 4 (age 23)	Type of current accommodation	ecatnn5318	n5318
	Voted in 1979 general election	ebinnn5960	n5960

Sweep	Variable description	Derived variable name	Original variable name
	Economic status	ecatneconstrg	econstrg
	Number of voluntary activities (youth club, church etc.)	econndv5	dv5
	Non-response at sweeps 1-4	NR04priorNR	OUTCME01-OUTCME04

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

Sweep	Variable description	Derived variable name	Original variable name
Sweep 0 (age 0)	Number of persons per room	aconnn512	n512
	Sex of child	bingender	n612
	Social class of mother's husband	acatnn236	n236
Sweep 1 (age 7)	Cognitive ability summary	CogAbil7	CogAbil7
Sweep 2 (age 11)	Area of world in which father born	ccatnn1436	n1436
	Child's positive activities outside school	cconnage11dv32	age11dv32
	Cognitive ability summary	genability11	genability11
Sweep 3 (age 16)	How long since child drank alcohol	dcatnn2888	n2888
	Sum of good activities performed outside school	dconnage16dv47	age16dv47
	Conduct problems [per unit]	Ext16MTDec18	Ext16MTDec18
Sweep 4 (age 23)	Legal marital status	ecatnn5113	n5113
	Voted in 1979 general election	ebinnn5960	n5960
Sweep 5 (age 33)	Type of accommodation	fcatnn502940_cat	n502940
	Current member of a Trade Union/Staff Association	fbinnn504646_bin	n504646
	Social capital score (people turn to for advice, support)	fconndvsoccapital_cont	Dvsoccapital
	Life contentment score	fconndvcontentmnt_cont	Dvcontentmnt
	Non-response at sweeps 1-5	NR05priorNR	OUTCME01-OUTCME05

Table 7. Predictors of non-response at Biomedical Sweep (age 44).

Sweep	Variable description	Derived variable name	Original variable name
Sweep 0 (age 0)	Number of persons per room	aconnn512	n512
	Abnormality during pregnancy	abinnn522	n522
	Social class of mother's father when she left school	acatnn660	n660
	Sex of child	bingender	n622
	Social class of mother's husband	acatnn236	n236
Sweep 1 (age 7)	Dad stayed on at school after minimum age	bbinnn194	n194
	Attendance	bcatnn458	n458
	Social problems (alcoholism etc.)	bconnage7dv10	age7dv10
	Cognitive ability summary	CogAbil7	CogAbil7
	Body mass index	bmi7	bmi7
Sweep 2 (age 11)	Cognitive ability summary	genability11	genability11
Sweep 3 (age 16)	Emotional or behavioural problem	dbinnn2021	n2021
	How long since child drank alcohol	dcatnn2888	n2888
	Test 2 – mathematics comprehension	dconnn2930	n2930
	Conduct problems	Ext16MTDec18	Ext16MTDec18
Sweep 4 (age 23)	Voted in 1979 general election	ebinnn5960	n5960
Sweep 5 (age 33)	Any work related training course since March 1981	fbinn501237	n501237

Sweep	Variable description	Derived variable name	Original variable name
	Number of hospital admissions since March 1981	fconnn504215_cont	n504215
	Driven/ridden after drinking alcohol in last 7 days	fcatnn504427_cat	n504427
	Social capital score (people turn to for advice, support)	fconndvsoccapital_cont	dvsoccapital
Sweep 6 (age 42)	Normally has access to a car or van	gcatncaracces	Caracces
	Participated in NCDS V	gbindmpart	Dmpart
	Intends to move in near future	gbinwantmove	wantmove
	Has a computer at home	gbinpchome	Pchome
	Non-response at sweeps 1-6	NR06priorNR	OUTCME01-OUTCME06

Table 8. Predictors of non-response at Sweep 7 (age 46).

Sweep	Variable description	Derived variable name	Original variable name
Sweep 0 (age 0)	Number of persons per room	aconnn512	n512
	Sex of child	bingender	N622
	Social class of mother's husband	acatnn236	n236
Sweep 1 (age 7)	Dad stayed on at school after minimum age	bbinnn194	n194
	Attendance	bcatnn458	n458
	Social problems (alcoholism etc.)	bconnage7dv10	age7dv10
	Cognitive ability summary [per unit]	CogAbil7	CogAbil7
Sweep 2 (age 11)	Source of family income last year	cbinnn1176	n1176
	Child's positive activities outside school	cconnage11dv32	age11dv32
	Cognitive ability summary	genability11	genability11
Sweep 3 (age 16)	Local Authority & voluntary schools	dcatnn2102	n2102
	Wish could leave school at 15 – study child	dcatnn2741	n2741
	How long since child drank alcohol	dcatnn2888	n2888
	Test 2 – mathematics comprehension	dconnn2930	n2930
Sweep 4 (age 23)	Number of accidents since 16th birthday	econnn5819	n5819
	Voted in 1979 general election	ebinnn5960	n5960
Sweep 5 (age 33)	Voted in 1987 general election	fbinnn504636_bin	n504636

Sweep	Variable description	Derived variable name	Original variable name
	Social capital score (people turn to for advice, support)	fconndvsoccapital_cont	dvsoccapital
Sweep 6 (age 42)	Participated in NCDS V	gbindmpart	dmpart
	Intends to move in near future	gbinwantmove	wantmove
	Membership in organisations	Org42	Org42
Biomedical Sweep (age 44)	Current legal marital status	mcatnmarital	marital
	Is current accommodation owned or rented?	OwnBM	Ownhome
	Non-response at sweeps 1-biomedical	NRBMpriorNR	OUTCME01-OUTCMEBM

Table 9. Predictors of non-response at Sweep 8 (age 50).

Sweep	Variable description	Derived variable name	Original variable name
Sweep 0 (age 0)	Number of persons per room	aconnn512	n512
	Sex of child	bingender	n622
	Social class of mother's husband	acatnn236	n236
Sweep 1 (age 7)	Social problems (alcoholism etc.)	bconnage7dv10	age7dv10
	Cognitive ability summary	CogAbil7	CogAbil7
	Summary of medical conditions	MedExSum7	MedExSum7
Sweep 2 (age 11)	Cognitive ability summary	genability11	genability11
	Conduct problems	Ext11Dec18	Ext11Dec18
Sweep 3 (age 16)	Child's school attendance	dconnn1721	n1721
	How long since child drank alcohol	dcatnn2888	n2888
	Test 2 – mathematics comprehension	dconnn2930	n2930
	Conduct problems	Ext16MTDec18	Ext16MTDec18
Sweep 4 (age 23)	Legal marital status	ecatnn5113	n5113
	Voted in 1979 general election	ebinnn5960	n5960
	Economic status	ecatneconstrg	econstrg
Sweep 5 (age 33)	Voted in 1987 general election	fbinnn504636_bin	n504636
	Social capital score (people turn to for advice, support)	fconndvsoccapital_cont	Dvsoccapital

Sweep	Variable description	Derived variable name	Original variable name
	Life contentment score	fconndvcontentmnt_cont	Dvcontentmnt
Sweep 6 (age 42)	Frequency of eating biscuits and cakes of all kinds	gconncakes	cakes
	Is current accommodation owned or rented?	gbinntenure2	tenure2
	Participated in NCDS V	gbindmpart	dmpart
	Ever wanted improve your maths?	gbinmthimp	mthimp
	Membership in organisations	Org42	Org42
Biomedical Sweep (age 44)	Consent to access NHS records	mbinnhsok	nhsok
	How many children do you have living with you aged 18 or less	mconnchildnow	childnow
	How many natural (biological) children have you ever had	mconnchildnum	childnum
Sweep 7 (age 46)	Non-response at sweeps 1-7	NR07priorNR	OUTCME01-OUTCME07

Table 10. Predictors of non-response at sweep 9 (age 55).

Sweep	Variable description	Derived variable name	Original variable name
Sweep 0 (age 0)	Mother's age	aconnn553	n553
	Number of persons per room	aconnn512	n512
	Parity	aconnn504	n504
	Social class of mother's father when she left school	acatnn660	n660
	Sex of child	bingender	n622
	Social class of mother's husband	acatnn236	n236
Sweep 1 (age 7)	Dad stayed on at school after minimum age	bbinnn194	n194
	Social problems (alcoholism etc.)	bconnage7dv10	age7dv10
	Cognitive ability summary	CogAbil7	CogAbil7
	Ever breastfed	bfever	n222
Sweep 2 (age 11)	Cognitive ability summary	genability11	genability11
	Conduct problems	Ext11Dec18	Ext11Dec18
Sweep 3 (age 16)	Child receiving help at school – backwardness	dbinnn2250	n2250
	Child's school attendance	dconnn1721	n1721
	How long since child drank alcohol	dcatnn2888	n2888
	Test 2 – mathematics comprehension	dconnn2930	n2930
	Conduct problems	Ext16MTDec18	Ext16MTDec18

Sweep	Variable description	Derived variable name	Original variable name
Sweep 4 (age 23)	Legal marital status	ecatnn5113	n5113
	Voted in 1979 general election	ebinnn5960	n5960
Sweep 5 (age 33)	Telephone in home	fbinn502977	n502977
	How much physical effort in job	fconnn504361_cont	n504361
	Voted in 1987 general election	fbinnn504636_bin	n504636
	Housing tenure	fbinntenure91_bin	tenure91
	Social capital score (people turn to for advice, support)	fconndvsoccapital_cont	dvsoccapital
Sweep 6 (age 42)	Participated in NCDS V	gbindmpart	dmpart
	Membership in organisations	Org42	Org42
Biomedical Sweep (age 44)	Self-rated general health	mconngenhlth	genhlth
Sweep 7 (age 46)	Marital status - de facto	hcatnnd7ms_cat	nd7ms
Sweep 8 (age 50)	Total number of natural children	iconnnd8nchtt_cont	nd8nchtt
	Employer provided pension scheme	ibinn8j2101	n8j2101
	Non-response at sweeps 1-8	NR08priorNR	OUTCME01-OUTCME08