# Handling missing data in Stata – a whirlwind tour

2012 Italian Stata Users Group Meeting

Jonathan Bartlett
www.missingdata.org.uk

20th September 2012

# Outline

# Outline

The problem of missing data and a principled approach

Missing data assumptions

Complete case analysis

Multiple imputation

Inverse probability weighting

Conclusions

# The problem of missing data

- Missing data is a pervasive problem in epidemiological, clinical, social, and economic studies.
- Missing data always cause some loss of information which cannot be recovered.
- But statistical methods can often help us make best use of the data which has been observed.
- More seriously, missing data can introduce bias into our estimates.

# Untestable assumptions

- Whether missing data cause bias depends on how missingness is associated with our variables.
- Crucially, with missing data we cannot empirically verify the required assumptions.
- e.g. consider the following distribution of smoking status (for males in THIN from [1]):

| Smoking status | n (% of sample) | (% of those observed) |
|---|---|---|
| Non | 82,479 (36) | (48) |
| Ex | 30,294 (13) | (18) |
| Current | 57,599 (25) | (34) |
| Missing | 56,661 (25) | n/a |

- Are the %s in the last column unbiased estimates?

# A principled approach to missing data

- We cannot be sure that the required assumptions are true given the observed data.
- Data analysis and contextual knowledge should be used to decide what assumption(s) are plausible about missingness.
- We can then choose a statistical method which is valid under this/these assumption(s).

# Outline

The problem of missing data and a principled approach

Missing data assumptions

Complete case analysis

Multiple imputation

Inverse probability weighting

Conclusions

# Rubin's classification

- Rubin developed a classification for missing data 'mechanisms' [2].
- We introduce the three types in a very simple setting.
- We assume we have one fully observed variable $X$ (age), and one partially observed variable $Y$ (blood pressure (BP)).
- We will let $R$ indicate whether $Y$ is observed ($R = 1$) or is missing ($R = 0$).

# Missing completely at random

- The missing values in BP ($Y$) are said to be missing completely at random (MCAR) if missingness is independent of BP ($Y$) and age ($X$).
- i.e. those subjects with missing BP do not differ systematically (in terms of BP or age) to those with BP observed.
- In terms of the missingness indicator $R$, MCAR means

$$P(R = 1|X, Y) = P(R = 1)$$

- e.g. 1 in 10 printed questionnaires were mistakenly printed with a page missing.

# Example - blood pressure (simulated data)

We assume age has been categorised into 30-50 and 50-70.

$n = 200$, but only 99 subjects have BP observed:

| Age | n | Mean (SD) BP |
|---|---|---|
| 30-50 | 72 | 129.7 (10.3) |
| 50-70 | 27 | 160.6 (11.7) |

# Checking MCAR

- With the observed data, we could investigate whether age $X$ is associated with missingness of blood presure ($R$).
- If it is, we can conclude the data are not MCAR.
- If it is not, we cannot necessarily conclude the data are MCAR.
- It is possible (though arguably unlikely in this case) that BP is associated with missingness in BP, even if age is not.

# Example - blood pressure (simulated data)

We compare the distribution of age in those with BP observed and those with BP missing:

```
. tab age r, chi2 row
```

| Key |
|---|
| *frequency* <br> *row percentage* |

| age | r: 0 | r: 1 | Total |
|---|---|---|---|
| 30-50 | 28<br>28.00 | 72<br>72.00 | 100<br>100.00 |
| 50-70 | 73<br>73.00 | 27<br>27.00 | 100<br>100.00 |
| Total | 101<br>50.50 | 99<br>49.50 | 200<br>100.00 |

```
Pearson chi2(1) =  40.5041   Pr = 0.000
```

$p < 0.001$ from chi2 test, shows we have strong evidence that missingness is associated with age.

# Missing at random

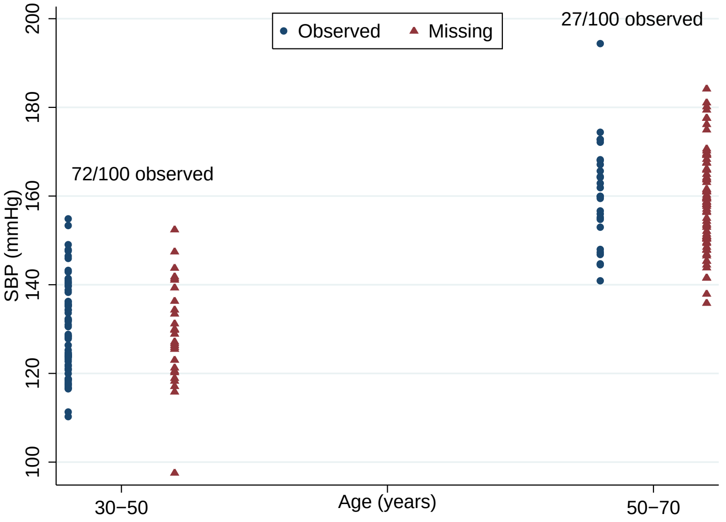
- BP ($Y$) is missing at random (MAR) given age ($X$) if missingness is independent of BP ($Y$) given age ($X$).
- This means that amongst subjects of the same age, missingness in BP is independent of BP.
- In terms of the missingness indicator $R$, MAR means

$$P(R = 1|X, Y) = P(R = 1|X)$$

# Checking MAR

- We cannot check whethe MAR holds based on the observed data.
- To do this we would need to check whether, within categories of age, those with missing BP had higher/lower BP than those with it observed.

# BP MAR given age



# A different representation of MAR

- We have defined MCAR and MAR in terms of how $P(R = 1|Y, X)$ depends on age ($X$) and BP ($Y$).
- From the plot, we see that MAR can also be viewed in terms of the conditional distribution of BP ($Y$) given age ($X$).
- MAR implies that

$$f(Y|X, R = 0) = f(Y|X, R = 1) = f(Y|X)$$

- That is, the distribution of BP ($Y$), given age ($X$), is the same whether or not BP ($Y$) is observed.
- This key consequence of MAR is directly exploited by multiple imputation.

# Missing not at random

- If data are neither MCAR nor MAR, they are missing not at random (MNAR).
- This means the chance of seeing $Y$ depends on $Y$, even after conditioning on $X$.
- Equivalently, $f(Y|X, R=0) \neq f(Y|X, R=1)$.
- MNAR is much more difficult to handle. Essentially the data cannot tell us how the missing values differ to the observed values (given $X$).
- We are thus led to conducting sensitivity analyses.

# Outline

# Complete case analysis

- Complete case (CC) (or complete records) analysis involves using only data from those subjects for whom all of the variables involved in our analysis are observed.
- CC is the default approach of most statistical packages (including Stata) when we have missing data.
- By only analysing a subset of records, our estimates will be less precise than had there been no missing data.
- Arguably more importantly, our estimates may be biased if the complete records differ systematically to the incomplete records.
- However, CC can be unbiased in certain situations in which the complete records are systematically different.

# Validity of complete case analysis

- CC analysis is valid provided the probability of being a CC is independent of outcome, given the covariates in the model of interest [3].
- Note that this condition has nothing to do with which variable(s) have missing values.
- This condition does not ‘fit’ into the MCAR/MAR/MNAR classification.
- It is not true, as is sometimes stated, that CC is always biased if data are not MCAR!

# The complete case assumption

- The validity of the assumption required for CC analysis to be unbiased depends on the model of interest.
- Returning to the example of estimating mean BP, we can think of this as the following linear model with no covariates:

$$BP_i = \alpha + \epsilon_i$$

  with $\epsilon_i \sim N(0, \sigma_\epsilon^2)$.
- Here CC analysis is unbiased only of missingness is independent of BP ($Y$), i.e. $P(R = 1|Y) = P(R = 1)$.

# Estimating mean BP - complete case analysis

```
. reg sbp
      Source |       SS       df       MS              Number of obs =      99
-------------+------------------------------           F(  0,    98) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |  29924.3689    98  305.350703           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  29924.3689    98  305.350703           Root MSE      =  17.474

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   138.1012   1.756232    78.63   0.000      134.616    141.5864
------------------------------------------------------------------------------
```

- The estimated mean (138.1) is biased downwards (truth=145).
- This is because missingness is associated with BP (higher BP → more chance of BP missing).

# A model for which CC is unbiased

```
. reg sbp age

      Source |       SS       df       MS              Number of obs =      99
-------------+------------------------------           F(  1,    97) =  163.17
       Model |  18767.6873     1  18767.6873           Prob > F      =  0.0000
    Residual |  11156.6816    97  115.017336           R-squared     =  0.6272
-------------+------------------------------           Adj R-squared =  0.6233
       Total |  29924.3689    98  305.350703           Root MSE      =  10.725

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    30.9154   2.420199    12.77   0.000     26.11197    35.71882
       _cons |   129.6697   1.263908   102.59   0.000     127.1612    132.1782
------------------------------------------------------------------------------
```

- This CC analysis is unbiased, because we condition on the cause of missingness (BP).
- Of course this alternative model does not (by itself) give an estimate of mean BP.

# Outline

The problem of missing data and a principled approach

Missing data assumptions

Complete case analysis

Multiple imputation

Inverse probability weighting

Conclusions

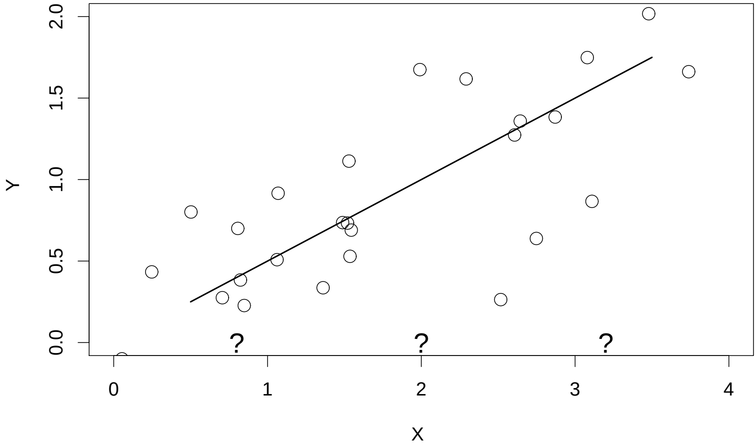
# Multiple imputation

- Multiple imputation (MI) involves ‘filling in’ each missing values multiple times.
- This results in multiple completed datasets.
- We then analyse each completed dataset separately, and combine the estimates using formulae developed by Rubin (‘Rubin’s rules’).
- By using observed data from all cases, estimates based on MI are generally more efficient than from CC.
- And, in some settings, MI may remove bias present CC estimates.

# MI in a very simple setting

- There are many different imputation methods.
- We describe one (the 'classic') in the context of a very simple setting.
- Suppose we have two continuous variables $X$ and $Y$.
- $X$ is fully observed, but $Y$ has some missing values.
- Our task is to impute the missing values in $Y$ using $X$.

# Imputing $Y$ from $X$



# Linear regression imputation

1. Fit the linear regression of $Y$ on $X$ using the complete cases:

$$Y = \alpha + \beta X + \epsilon$$

   where $\epsilon \sim N(0, \sigma^2)$.
2. This gives estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$.
3. To create the $m$th imputed dataset:
   3.1 Draw new values $\alpha_m$, $\beta_m$ and $\sigma_m^2$ based on $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$.
   3.2 For each subject with observed $X_i$ but missing $Y_i$, create imputation $Y_{i(m)}$ by:

$$Y_{i(m)} = \alpha_m + \beta_m X_i + \epsilon_{i(m)}$$

   where $\epsilon_{i(m)}$ is a random draw from $N(0, \sigma_m^2)$.

# The end result

| | Data | | Imputation 1 | | Imputation 2 | | Imputation 3 | | Imputation 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Subject | $Y$ | $X$ | $Y$ | $X$ | $Y$ | $X$ | $Y$ | $X$ | $Y$ | $X$ |
| 1 | 1.1 | 3.4 | 1.1 | 3.4 | 1.1 | 3.4 | 1.1 | 3.4 | 1.1 | 3.4 |
| 2 | 1.5 | 3.9 | 1.5 | 3.9 | 1.5 | 3.9 | 1.5 | 3.9 | 1.5 | 3.9 |
| 3 | 2.3 | 2.6 | 2.3 | 2.6 | 2.3 | 2.6 | 2.3 | 2.6 | 2.3 | 2.6 |
| 4 | 3.6 | 1.9 | 3.6 | 1.9 | 3.6 | 1.9 | 3.6 | 1.9 | 3.6 | 1.9 |
| 5 | 0.8 | 2.2 | 0.8 | 2.2 | 0.8 | 2.2 | 0.8 | 2.2 | 0.8 | 2.2 |
| 6 | 3.6 | 3.3 | 3.6 | 3.3 | 3.6 | 3.3 | 3.6 | 3.3 | 3.6 | 3.3 |
| 7 | 3.8 | 1.7 | 3.8 | 1.7 | 3.8 | 1.7 | 3.8 | 1.7 | 3.8 | 1.7 |
| 8 | **?** | 0.8 | **0.2** | 0.8 | **0.8** | 0.8 | **0.3** | 0.8 | **2.3** | 0.8 |
| 9 | **?** | 2.0 | **1.7** | 2.0 | **2.4** | 2.0 | **1.8** | 2.0 | **3.5** | 2.0 |
| 10 | **?** | 3.2 | **2.7** | 3.2 | **2.5** | 3.2 | **1.0** | 3.2 | **1.7** | 3.2 |

# The analysis stage

- For each imputation, we estimate our parameter of interest $\theta$, and records its standard error.
- e.g. $\theta = E(Y)$, the average value of $Y$.
- Let $\hat{\theta}_m$ and $Var(\hat{\theta}_m)$ denote the estimate of $\theta$ and its variance from the $m$th imputation.
- Our overall estimate of $\theta$ is then the average of the estimates from the imputed datasets

$$\hat{\theta}_{MI} = \frac{\sum_{m=1}^{M} \hat{\theta}_m}{M}$$

where $M$ denotes the number of imputations used.

# Variance estimation

- The 'within-imputation variance' is given by

$$\frac{\sum_{m=1}^{M} Var(\hat{\theta}_m)}{M}.$$

  This quantifies uncertainty due to the fact we have a finite sample (the usual cause of uncertainty in estimates).

- The 'between-imputation variance' is given by

$$\frac{\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta}_{MI})^2}{M-1}.$$

  This quantifies uncertainty due to the missing data.

- The overall uncertainty in our estimate $\hat{\theta}$ is then given by

$$Var(\hat{\theta}_{MI}) = \sigma_w^2 + \left(1 + \frac{1}{M}\right)\sigma_b^2.$$

# Inference

- The MI estimate and its variance can be used to form confidence intervals and performs hypothesis test.
- Implementations of MI in statistical packages like Stata automate the process of analysing each imputation and combining the results.

# Assumptions for MI

- MI gives unbiased estimates provided data are MAR and the imputation model(s) is correctly specified.
- To be correctly specified, we must include all variables involved in our model of interest in the imputation model(s).
- The plausibility of MAR can be guided by data analysis and contextual knowledge.
- Often we have variables which are associated with missingness and the variable(s) being imputed, but which are not in the model of interest.
- Including these in the imputation model increases likelihood of MAR holding.

# Specification of imputation models

- We should also ensure as best as possible that our imputation models are reasonably well specified.
- e.g. if a variable has a highly skewed distribution, imputing using normal linear regression is probably not a good idea.
- Various diagnostics can be used to aid this process, e.g. comparing distributions of imputed and observed

# MI in Stata

- Historically the only imputation command in Stata was Patrick Royston's `ice` command, which performed ICE/FCS imputation (more on this later).
- Stata 11 included imputation using the multivariate normal model.
- Stata 12 adds ICE/FCS imputation functionality.

# Imputing missing BP values in Stata

## Step 1 - `mi set` the data

- e.g. `mi set wide`
- Alternatives include `mlong`, `flong`.
- This only affects how Stata organises the imputed datasets.

# Imputing missing BP values in Stata

## Step 2 - `mi register` variables

- At a minimum, we must `mi register` variables with missing values we want to impute.
- e.g. `mi register imputed sbp`

# Imputing missing BP values in Stata

## Step 3 - imputing the missing values

- We are now ready to impute the missing values.
- Since we have only missing values in one continuous variable, we shall impute using a linear regression imputation model:

```
. mi impute reg sbp age, add(10) rseed(5123)
Univariate imputation                    Imputations =       10
Linear regression                              added =       10
Imputed: m=1 through m=10                    updated =        0
```

| | Observations per *m* | | | |
|---|---|---|---|---|
| Variable | Complete | Incomplete | Imputed | Total |
| sbp | 99 | 101 | 101 | 200 |

```
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

# Imputing missing BP values in Stata

## Step 4 - analysing the imputed datasets

- We are now ready to analyse the imputed datasets.
- This is done by Stata's `mi estimate` command, which supports most of Stata's estimation commands.

```
. mi estimate: reg sbp

Multiple-imputation estimates                   Imputations     =         10
Linear regression                               Number of obs   =        200
                                                Average RVI     =     0.7163
                                                Largest FMI     =     0.4420
                                                Complete DF     =        199
                                                DF:     min     =      35.63
                                                        avg     =      35.63
DF adjustment:   Small sample                           max     =      35.63
                                                F(   0,      .) =          .
Within VCE type:          OLS                   Prob > F        =          .

------------------------------------------------------------------------------
         sbp |      Coef.  Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   145.3263   1.747398    83.17   0.000     141.7811    148.8715
------------------------------------------------------------------------------
```

- The estimate is quite close to the true value (145).

# Other MI imputation methods in Stata

In addition to linear regression Stata's `mi` command offers imputation using:

- Logistic, ordinal logistic, and multinomial logsitic models
- Predictive mean matching
- Truncated normal regression for imputing bounded cts variables
- Interval regression for imputing censored cts variables
- Poisson regression for imputing count data
- Negative binomial regression for imputing overdispersed count data

# MI with more than one variable

- So far we have considered setting with one variable partially observed.
- Often we have datasets with multiple partially observed variables.
- Stata 11/12 supports imputation with the multi-variate normal model.
- What if we have categorical or binary variables with missing values?
- More on this in tomorrow's course...

# Outline

# Inverse probability weighting

- Inverse probability weighting (IPW) for missing data takes a different approach [4].
- We perform a CC analysis, but weight the complete cases by the inverse of their probability of having data observed (i.e. not being missing).
- Those who had a small chance of being observed are given increased weight, to compensate for those similar subjects who are missing.
- This requires us to model how missingness depends on fully observed variables.

# Using IPW to estimate mean BP

- Recall our previous analysis of missingness in BP and age:

```
. tab age r, chi2 row

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
+-------------------+
```

| age | r = 0 | r = 1 | Total |
|---|---|---|---|
| 30-50 | 28<br>28.00 | 72<br>72.00 | 100<br>100.00 |
| 50-70 | 73<br>73.00 | 27<br>27.00 | 100<br>100.00 |
| Total | 101<br>50.50 | 99<br>49.50 | 200<br>100.00 |

Pearson chi2(1) = 40.5041 Pr = 0.000

- The probability of observing BP is 0.72 for 30-50 year olds, and 0.27 for 50-70 year olds.
- So the 'weight' for 30-50 year olds is $1/0.72 = 1.39$ and for 50-70 year olds is $1/0.27 = 3.7$.

# The IPW estimator

- Since we are interested in estimating a simple parameter (mean BP), we can manually calculate the IPW estimate:

$$\frac{72 \times 129.7 \times 1.39 + 27 \times 160.6 \times 3.7}{72 \times 1.39 + 27 \times 3.7} = 145.1$$

- IPW appears has removed the bias from the simple CC estimate of mean BP.

# IPW more generally

## Step 1 - Constructing weights

- With multiple fully observed variables, we can use logistic regression to model missingness:

```
. logistic r age

Logistic regression                             Number of obs   =        200
                                                LR chi2(1)      =      42.00
                                                Prob > chi2     =     0.0000
Log likelihood = -117.62122                     Pseudo R2       =     0.1515
```

| r | Odds Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | .1438356 | .0455618 | -6.12 | 0.000 | .0773103 | .2676059 |
| _cons | 2.571428 | .5727026 | 4.24 | 0.000 | 1.661869 | 3.9788 |

```
. predict pr, pr

. gen wgt=1/pr
```

# IPW more generally

## Step 2 - parameter estimation

- We can then pass the constructed weights to our estimation command:

```
. reg sbp [pweight=wgt]
(sum of wgt is   2.0000e+02)

Linear regression                               Number of obs =      99
                                                F(  0,     98) =    0.00
                                                Prob > F      =       .
                                                R-squared     =  0.0000
                                                Root MSE      =  19.008
```

| sbp | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| _cons | 145.1274 | 2.162726 | 67.10 | 0.000 | 140.8356 | 149.4193 |

- Notice that the SE is larger (2.16) compared to the MI SE (1.75).

# Outline

# Problems caused by missing data and a principled approach

- Missing data reduce precision and potentially parameter bias estimates and inferences.
- Producing valid estimates requires additional assumptions about the missingness to be made.
- Ad-hoc methods should generally be avoided.
- Both data analysis and contextual knowledge should guide us in thinking about missingness in a given setting.
- We can then choose a statistical method which accommodates missing data under our chosen assumption (e.g. MAR).

# Complete case analysis

- Complete case (CC) analysis is the default method of most software packages, including Stata.
- CC analysis is generally biased unless data are MCAR.
- But it can be unbiased in certain non-MCAR settings when the model of interest is a regression model.
- Even when it is unbiased, CC may be inefficient compared to other methods.

# Multiple imputation

- Multiple imputation is a flexible approach to handling missing data under the MAR assumption [5].
- Stata 12 now includes a comprehensive range of MI commands, including ICE/FCS MI.
- In settings where both CC and MI are unbiased, MI will generally give more precise estimates.
- We must carefully consider the plausibility of the MAR assumption and whether imp. models are correctly specified.

# Inverse probability weighting

- IPW involves performing a weighted CC analysis.
- Rather than model the partially observed variable, we model the observation/missingness indicator $R$.
- The weights based on this model are then passed to our estimation command, and most Stata estimation commands support weights.
- Sometimes modelling missingness may be easier than modelling the partially obs. variable (e.g. if the partially observed variable has a tricky distribution).
- However, IPW estimators can be quite inefficient compared to MI or maximum likelihood.
- IPW is also difficult (or impossible) to use in settings with complicated patterns of missingness.

# Sensitivity to the MAR assumption

- Since we can never definitively our assumptions (e.g. MAR) hold, we should consider sensitivity analysis.
- MI can also be used to perform MNAR sensitivity analyses [6].
- If you want to learn more, come on our missing data short course at LSHTM in June.
- And/or visit our website www.missingdata.org.uk

# References I


[1] L. Marston, J. R. Carpenter, K. R. Walters, R. W. Morris, I. Nazareth, and I. Petersen.

Issues in multiple imputation of missing data for large general practice clinical databases.

*Pharmacoepidemiology and Drug Safety*, 19:618–626, 2010.

[2] D B Rubin.

Inference and missing data.

*Biometrika*, 63:581–592, 1976.

[3] I. R. White and J. B. Carlin.

Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values.

*Statistics in Medicine*, 28:2920–2931, 2010.

# References II

[4] S. R. Seaman and I. R. White.
Review of inverse probability weighting for dealing with missing data.
*Statistical Methods in Medical Research*, 2011.

[5] J A C Sterne, I R White, J B Carlin, M Spratt, P Royston, M G Kenward, A M Wood, and J R Carpenter.
Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.
*British Medical Journal*, 339:157–160, 2009.

[6] J R Carpenter, M G Kenward, and I R White.
Sensitivity analysis after multiple imputation under missing at random — a weighting approach.
*Statistical Methods in Medical Research*, 16:259–275, 2007.