Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2018/2019

Multiple imputation

- In the firsts lectures we have seen several *ad hoc* methods to deal with missing values, with the majority of them (exceptions being complete and available case analysis) falling in the category of the so-called *single imputation* techniques.

- Specifically, we have seen:

  1. Mean or unconditional imputation.
  2. Conditional mean/regression imputation.
  3. Stochastic regression imputation.
  4. Hot deck imputation (including predictive mean matching!).

- All these methods replace the missing value by *one* imputed value. As we have seen, the most promising approach was stochastic regression imputation (hot deck imputation was also fairly good).

Multiple imputation

- ↪ Single imputation, regardless of the method used, fails to satisfy statistical objectives concerning the validity of resulting inferences based on the filled-in data.

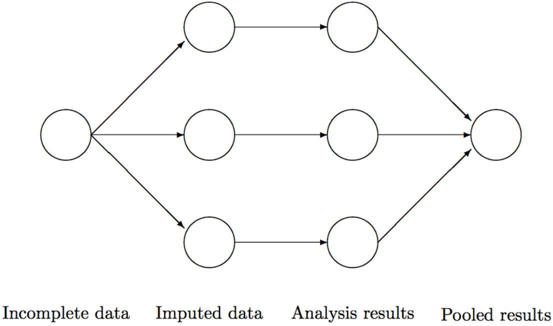
- ↪ Because a single imputed value cannot reflect any of the uncertainty about the true underlying value, analyses that treat imputed values just like observed values systematically underestimate uncertainty.

- ↪ Consequently, imputing a single value for each missing datum and then analysing the filled-in data using standard techniques for complete data will result in standard errors estimates that are too small, confidence intervals that fail to attain their nominal coverage, and *p*-values that are too significant.

Multiple imputation

- ↪ The problems of single imputation are largely overcome by the use of *multiple imputation* (MI), which is an approach to the missing values problem that allows the investigator to obtain valid assessments of the uncertainty.
- ↪ The basic idea of multiple imputation is to impute each missing value several times, thus creating $M > 1$ complete datasets.
- ↪ Each of these datasets is analysed by applying the analytic method that we would have used had the data been complete.
- ↪ The M results are pooled into a final point estimate plus standard error by simple pooling rules (‘Rubin’s rules’, see later).

Multiple imputation

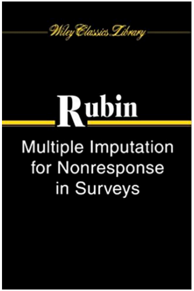


Scheme of the main steps in multiple imputation. Here *M* is three. Figure from van Buuren (2012), page 17.

Multiple imputation

- Multiple imputation was developed by Donald Rubin and first described by him in a 1977 manuscript prepared for the United Stated Social Survey Administration.

- This is reproduced as an Appendix in his 1987 book called *Multiple Imputation for Nonresponse in Surveys* (Rubin, 1987).



Wiley Classics Library

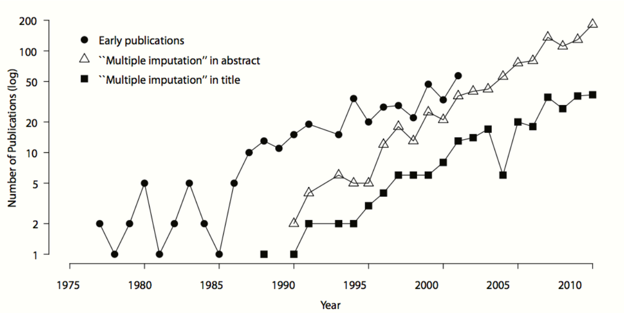
Rubin

Multiple Imputation
for Nonresponse
in Surveys

Multiple imputation

- ↪ The origins of multiple imputation clearly lie in the analysis of survey data, and the dominant (but not exclusive) paradigm for developing and justifying the approach, as set out in both the original manuscript and subsequent book, is that of sample surveys.
- ↪ In the decades that followed the use of MI has extended to many other application areas, most notably in biostatistics.
- ↪ In fact, there is an enormous expanding literature in multiple imputation.

Multiple imputation



Number of publications (log) on multiple imputation during the period 1977–2010 according to three counting methods. Data source: `www.scopus.com`. Figure from van Buuren (2012), page 28.

Multiple imputation

- ↪ The figure in the previous slide contains three time series with (log) counts on the number of publications on multiple imputation during the period 1977–2010.

- ↪ The right most series corresponds to the number of publications per year that featured the search term ‘multiple imputation’ in the title.

- ↪ The search was done in Scopus on July 11, 2011.

- ↪ These are often methodological articles in which new adaptations are being developed.

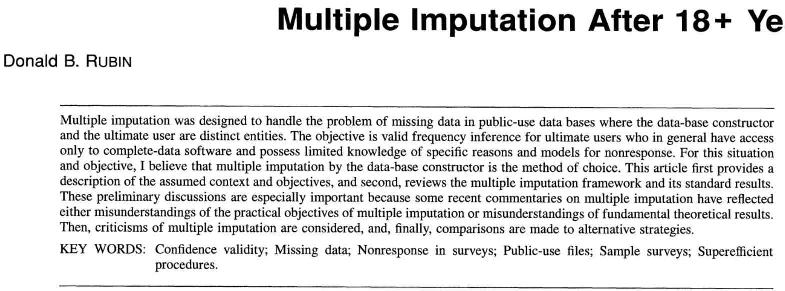
- ↪ The series in the middle is the number of publication that featured ‘multiple imputation’ in the title, abstract or key words in Scopus on the same search data. This set includes a growing group of papers that contain applications.

Multiple imputation

- ↪ Scopus does not go back further than 1988 on this topic.
- ↪ The leftmost series is the number of publications in a collection of early publications.
- ↪ This collection covers essentially everything related to multiple imputation from its inception in 1977 up to the year 2001.
- ↪ Note that the vertical axis is set in the logarithm. Perhaps the most interesting series is the middle series counting the applications. The pattern is approximately linear, meaning that the number of applications is growing at an exponential rate.

Multiple imputation

↪ Below I leave the title and abstract of an interesting article published by Donald Rubin in the *Journal of the American Statistical Association*, volume 91, pages 473–489, in 1996.



Multiple Imputation After 18+ Years

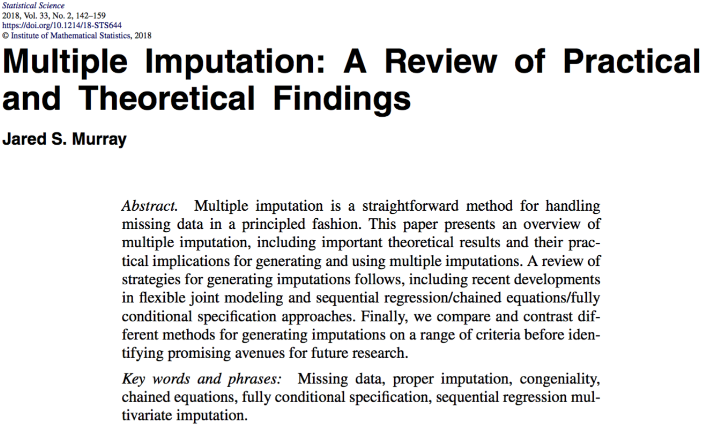
Donald B. RUBIN

Multiple imputation was designed to handle the problem of missing data in public-use data bases where the data-base constructor and the ultimate user are distinct entities. The objective is valid frequency inference for ultimate users who in general have access only to complete-data software and possess limited knowledge of specific reasons and models for nonresponse. For this situation and objective, I believe that multiple imputation by the data-base constructor is the method of choice. This article first provides a description of the assumed context and objectives, and second, reviews the multiple imputation framework and its standard results. These preliminary discussions are especially important because some recent commentaries on multiple imputation have reflected either misunderstandings of the practical objectives of multiple imputation or misunderstandings of fundamental theoretical results. Then, criticisms of multiple imputation are considered, and, finally, comparisons are made to alternative strategies.

KEY WORDS: Confidence validity; Missing data; Nonresponse in surveys; Public-use files; Sample surveys; Superefficient procedures.

Multiple imputation

↪ Another very interesting article that came out this year in Statistical Science (available on Learn).



Statistical Science
2018, Vol. 33, No. 2, 142–159
https://doi.org/10.1214/18-STS644

Multiple Imputation: A Review of Practical and Theoretical Findings

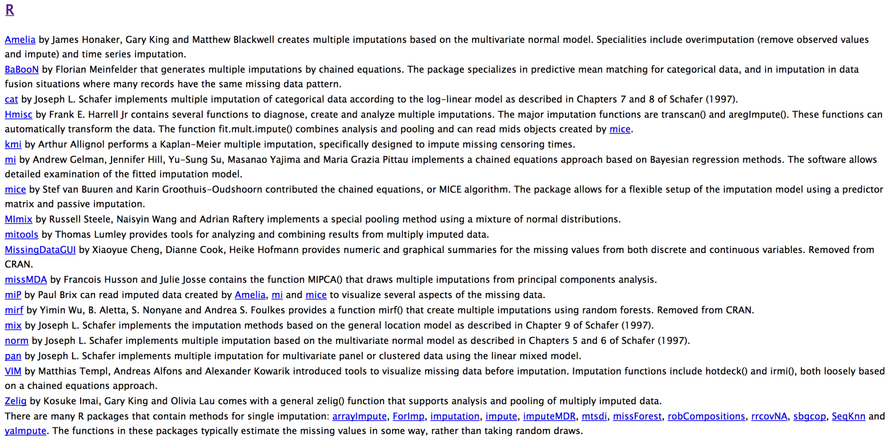
Jared S. Murray

Abstract. Multiple imputation is a straightforward method for handling missing data in a principled fashion. This paper presents an overview of multiple imputation, including important theoretical results and their practical implications for generating and using multiple imputations. A review of strategies for generating imputations follows, including recent developments in flexible joint modeling and sequential regression/chained equations/fully conditional specification approaches. Finally, we compare and contrast different methods for generating imputations on a range of criteria before identifying promising avenues for future research.

Key words and phrases: Missing data, proper imputation, congeniality, chained equations, fully conditional specification, sequential regression multivariate imputation.

Multiple imputation

↪ List of software available to conduct multiple imputation:



R

Amelia by James Honaker, Gary King and Matthew Blackwell creates multiple imputations based on the multivariate normal model. Specialities include overimputation (remove observed values and impute) and time series imputation.

BaBooN by Florian Meinfelder that generates multiple imputations by chained equations. The package specializes in predictive mean matching for categorical data, and in imputation in data fusion situations where many records have the same missing data pattern.

cat by Joseph L. Schafer implements multiple imputation of categorical data according to the log-linear model as described in Chapters 7 and 8 of Schafer (1997).

Hmisc by Frank E. Harrell Jr contains several functions to diagnose, create and analyze multiple imputations. The major imputation functions are transcan() and aregImpute(). These functions can automatically transform the data. The function fit.mult.impute() combines analysis and pooling and can read mids objects created by mice.

kmi by Arthur Allignol performs a Kaplan-Meier multiple imputation, specifically designed to impute missing censoring times.

mi by Andrew Gelman, Jennifer Hill, Yu-Sung Su, Masanao Yajima and Maria Grazia Pittau implements a chained equations approach based on Bayesian regression methods. The software allows detailed examination of the fitted imputation model.

mice by Stef van Buuren and Karin Groothuis-Oudshoorn contributed the chained equations, or MICE algorithm. The package allows for a flexible setup of the imputation model using a predictor matrix and passive imputation.

MImix by Russell Steele, Naisyin Wang and Adrian Raftery implements a special pooling method using a mixture of normal distributions.

mitools by Thomas Lumley provides tools for analyzing and combining results from multiply imputed data.

MissingDataGUI by Xiaoyue Cheng, Dianne Cook, Heike Hofmann provides numeric and graphical summaries for the missing values from both discrete and continuous variables. Removed from CRAN.

missMDA by Francois Husson and Julie Josse contains the function MIPCA() that draws multiple imputations from principal components analysis.

miP by Paul Brix can read imputed data created by Amelia, mi and mice to visualize several aspects of the missing data.

mirf by Yimin Wu, B. Aletta, S. Nonyane and Andrea S. Foulkes provides a function mirf() that create multiple imputations using random forests. Removed from CRAN.

mix by Joseph L. Schafer implements the imputation methods based on the general location model as described in Chapter 9 of Schafer (1997).

norm by Joseph L. Schafer implements multiple imputation based on the multivariate normal model as described in Chapters 5 and 6 of Schafer (1997).

pan by Joseph L. Schafer implements multiple imputation for multivariate panel or clustered data using the linear mixed model.

VIM by Matthias Templ, Andreas Alfons and Alexander Kowarik introduced tools to visualize missing data before imputation. Imputation functions include hotdeck() and irmi(), both loosely based on a chained equations approach.

Zelig by Kosuke Imai, Gary King and Olivia Lau comes with a general zelig() function that supports analysis and pooling of multiply imputed data.

There are many R packages that contain methods for single imputation: arrayImpute, ForImp, imputation, impute, imputeMDR, mtsdi, missForest, robCompositions, rrcovNA, sbgcop, SeqKnn and yaImpute. The functions in these packages typically estimate the missing values in some way, rather than taking random draws.

Multiple imputation

The three steps

- Let us now describe in detail the three main steps (or tasks in Rubin's terminology) of MI.
  - **Step 1**: Create a number (M) of copies of the incomplete dataset, and use an appropriate procedure to impute (fill in) the missing values in each of these copies. The imputed datasets are composed of a fixed portion – the observed data– and a varying portion – the filled-in values. Since we do not know the true values that are missing it seems reasonable that the imputed values used in each copy should in general differ from each other. The choice of M is discussed later.
  - **Step 2**: For each completed/imputed copy of the dataset, carry out the standard analysis that would have been performed in the absence of missing values, and store the parameter estimates of interest, along with their estimated variances (or variance-covariance matrix in the case of a multivariate parameter of interest). For now, we focus on a single (univariate) parameter of interest, which we denote by θ. The estimate of θ obtained from the mth ($m = 1, \ldots, M$) complete dataset is denoted by $\widehat{\theta}^{(m)}$ and its (estimated) variance by $U^{(m)}$.

Multiple imputation

The three steps

- ↪ (Continued)
    - ↪ **Step 3**: The results of the *M* analyses are combined into a single analysis that takes into account the imputation.
- ↪ As we shall see, the only complex part of this procedure is step one: formulate a good imputation process, to be used *M* times.
- ↪ The specification of an appropriate imputation model is the key issue, since if this is misspecified, there is the potential for bias.
- ↪ The second step, producing the final estimate, is straightforward as it treats each imputed dataset as if it were a real dataset, we just have to do it *M* times.
- ↪ As Schafer (1997) says, multiple imputation works by 'solving an incomplete-data problem by repeatedly solving the complete-data version'.
- ↪ The third step involves simple arithmetic.

Multiple imputation

Rubin's rules

↪ Let us now turn attention to the third step and learn how to pool the M estimates together.

↪ Let θ be our (scalar) parameter of interest. From step 2 we have $\widehat{\theta}^{(1)}, \ldots, \widehat{\theta}^{(M)}$, the M estimated parameters. We also have $U^{(m)}$, the estimated variances of $\widehat{\theta}^{(m)}$, $m = 1, \ldots, M$.

↪ According to Rubin's rules, the combined estimate of the parameter is the average of the M separate estimates

$$\widehat{\theta}^{\mathrm{MI}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\theta}^{(m)}.$$

Multiple imputation

Rubin's rules

- The total variance, V^{MI} is computed by combining the between imputation variability

$$B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\theta}^{(m)} - \widehat{\theta}^{\mathrm{MI}}\right)^2,$$

and the within imputation variability

$$\bar{W} = \frac{1}{M}\sum_{m=1}^{M} U^{(m)}.$$

- It is tempting to conclude that the total variance V^{MI} is equal to the sum of $\bar{W}$ and B, but that would be incorrect.

- We need to incorporate the fact that $\widehat{\theta}^{\mathrm{MI}}$ itself is estimated using finite M, and thus only approximates $\widehat{\theta}^{\mathrm{MI}}_{\infty}$, the estimator that would have been obtained for an infinitely large number of imputations $M = \infty$.

Multiple imputation

Rubin's rules

↪ Rubin (1987, eq. 3.3.5) shows that the contribution to the variance of this factor is systematic and equal to B_∞/M. Since B approximates B_∞ (estimated between imputation variance for infinitely many imputations), we may write:

$$\begin{aligned} V^{\text{MI}} &= \bar{W} + B + \frac{B}{M} \\ &= \bar{W} + \left(1 + \frac{1}{M}\right) B, \end{aligned}$$

for the total variance of $\widehat{\theta}^{\text{MI}}$.

Multiple imputation

Rubin's rules

↪ In summary, the total variance $\widehat{\theta}^{\text{MI}}$ stems from three sources:

1. $\bar{W}$, the variance caused by the fact that we are taking a sample rather than observing the entire population. This is the conventional measure of variability.
2. B, the extra variance caused by the fact that there are missing values in the sample.
3. B/M, the extra simulation variance caused by the fact that $\widehat{\theta}^{\text{MI}}$ itself is estimated for finite M.

↪ The addition of the later term is critical to make the multiple imputation work at low values of M. Note including it would result in p-values that are too low, or confidence intervals that are too short.

Multiple imputation

Rubin's rules–toy example

- Suppose we take a survey of five people, measuring their height and weight. Only three of them give their weight; the other two don't give it just because of random chance. The data are:

Height (inches)	Weight (pounds)
65	130
68	140
70	150
72	NA
75	NA

- The aim of the analysis (step 2) is to regress the weight on the height.

- Suppose that five plausible values for each missing weight have been generated to create five complete datasets.

Multiple imputation

Rubin's rules–toy example

Height	Weight–1	Weight–2	Weight–3	Weight–4	Weight–5
65	130	130	130	130	130
68	140	140	140	140	140
70	150	150	150	150	150
72	157	166	155	157	156
75	171	169	167	171	168
Estimated slope	4.12	4.26	3.71	4.12	3.83
(Variance)	(0.025)	(0.346)	(0.024)	(0.025)	(0.018)

Multiple imputation

Rubin's rules–toy example

↪ The final estimate for the slope is

$$\widehat{\theta}^{\text{MI}} = \frac{1}{5}(4.12 + 4.26 + 3.71 + 4.12 + 3.83) = 4.008$$

↪ The within imputation variance is

$$\bar{W} = \frac{1}{5}(0.025 + 0.346 + 0.024 + 0.025 + 0.018) = 0.0876$$

↪ The between imputation variance is

$$\begin{aligned} B &= \frac{1}{4}\{(4.12-4.008)^2 + (4.26-4.008)^2 + (3.71-4.008)^2 + (4.12-4.008)^2 + (3.83-4.008)^2\} \\ &= 0.05227 \end{aligned}$$

↪ Thus, the final estimate of the variance is

$$V^{\text{MI}} = 0.0876 + \left(1 + \frac{1}{5}\right) \times 0.05227 = 0.150324$$

Multiple imputation

Rubin's rules

↪ Extensions to the case where the parameter of interest is multivariate, say $\boldsymbol{\theta}$, are straightforward.

↪ With a single parameter, the within imputation variance is the arithmetic average of the M sampling variances.

↪ In the multivariate context, the within imputation covariance matrix is the average of the M covariance matrices, namely

$$\bar{\mathbf{W}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{U}^{(m)},$$

where $\bar{\mathbf{W}}$ is the average within imputation covariance matrix, and $\mathbf{U}^{(m)}$ is the covariance matrix from the completed dataset m.

Multiple imputation

Rubin's rules

- ↪ Filling in the data with different sets of imputed values causes the parameter estimates to vary across the M analyses, and this between imputation variability is an important component of the total sampling error.

- ↪ The between imputation covariance matrix quantifies this variability as follows

$$\mathbf{B} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\boldsymbol{\theta}}^{(m)} - \widehat{\boldsymbol{\theta}}^{\text{MI}}\right)\left(\widehat{\boldsymbol{\theta}}^{(m)} - \widehat{\boldsymbol{\theta}}^{\text{MI}}\right)^{T},$$

where $\mathbf{B}$ is the between imputation covariance matrix, $\widehat{\boldsymbol{\theta}}^{(m)}$ contains the parameter estimates from the mth imputed dataset, and $\widehat{\boldsymbol{\theta}}^{\text{MI}}$ is the vector of pooled point estimates (i.e., the arithmetic average of the $\widehat{\boldsymbol{\theta}}^{(m)}$ vectors).

Multiple imputation

Rubin's rules

- ↪ The diagonal elements of **B** contain the between imputation variance estimate for individual parameters, and the off-diagonal elements quantify the extent to which the between imputation fluctuation in one parameter is related to the between imputation fluctuation in another parameter.

- ↪ Considered as a whole, the between imputation covariance matrix represents the additional sampling fluctuation that results from the missing data.

- ↪ Finally, the total covariance matrix combined the within and between imputation covariance matrices as follows

$$\begin{aligned} \mathbf{V}^{\mathrm{MI}} &= \bar{\mathbf{W}} + \mathbf{B} + \frac{1}{M}\mathbf{B}, \\ &= \bar{\mathbf{W}} + \left(I + \frac{1}{M}\right)\mathbf{B}. \end{aligned}$$

Multiple imputation

How many imputations?

- The traditional choices for M are $M = 3$, $M = 5$, and $M = 10$.

- Such a small number of imputations are often judged to be acceptable because on another result due to Rubin (1987).

- He showed that the relative efficiency of using M imputations compared to the maximum theoretical number of imputations if infinitely, is $M/(M + \lambda)$, where λ is the fraction of missing information.

- Even when λ is as much as one half, the relative efficiency when $M = 5$ is $5/(5 + 0.5) = 0.91$, i.e., this is 91% efficient.

Multiple imputation

How many imputations?

- ↪ Even so, averaging processes, such as the one involved in $\widehat{\theta}^{\mathrm{MI}}$ and V^{MI} are intuitively more reliable when the number of items averaged over is higher, whilst modern computing power should be able to handle many imputations in a reasonable amount of time for all but the largest datasets.

- ↪ Further, as seen in Rubin's expression for the total variance, the larger M gets, the smaller the effect of simulation error on the total variance.

Multiple imputation

Proper MI

- The validity of MI rests on how imputations are created and how that procedure relates to the model used to subsequently analyse the data.

- Creating MIs often require special algorithms (Schafer 1997).

- In general, they should be drawn from a distribution for the missing data that reflects uncertainty about the parameters of the data model.

- Recall that in stochastic regression imputation we've imputed from the conditional distribution $f(y_{\text{mis}} \mid y_{\text{obs}}, \widehat{\theta})$, where $\widehat{\theta}$ is an estimate from the observed data. More specifically, we've imputed y_{mis} from $N(y_{\text{obs}}^T \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$, where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ were the least squares estimates computed from the observed data.

Multiple imputation

Proper MI

- In MI, imputations are simulated from the Bayesian posterior predictive distribution of the missing data given the observed data, rather from the density $f(y_{\text{mis}} \mid y_{\text{obs}}, \widehat{\theta})$.

- Assuming MAR data, the predictive posterior distribution is obtained by integrating (averaging) out parameters from the likelihood using the posterior distribution of the parameters
$$f(y_{\text{mis}} \mid y_{\text{obs}}) = \int f(y_{\text{mis}} \mid y_{\text{obs}}, \theta) f(\theta \mid y_{\text{obs}}) \mathrm{d}\theta$$

- Therefore, imputed values are sampled from a distribution that incorporates uncertainty in estimating model parameters and uncertainty due to sampling data (missing values) from the estimated model.

- So, in MI we first simulate parameters $\theta^{(1)}, \ldots, \theta^{(M)}$ from the posterior distribution $f(\theta \mid y_{\text{obs}})$ and then we draw the missing data $y_{\text{mis}}^{(m)}$ from $f(y_{\text{mis}} \mid y_{\text{obs}}, \theta^{(m)})$, for $m = 1, \ldots, M$.

Multiple imputation

Proper MI

- For instance, a MI version of stochastic regression imputation (i.e., stochastic regression imputation repeated M times) that accounts for uncertainty in the parameter estimates would be

$$\boldsymbol{\theta}^{(m)} = \left(\boldsymbol{\beta}^{(m)}, (\sigma^{(m)})^2\right) \sim f(\boldsymbol{\beta}, \sigma^2 \mid y_{\text{obs}}),$$

$$y_{\text{mis}}^{(m)} \sim \text{N}\left(y_{\text{obs,partial}}^T \boldsymbol{\beta}^{(m)}, (\sigma^{(m)})^2\right),$$

  where $f(\boldsymbol{\beta}, \sigma^2 \mid y_{\text{obs}})$ is the posterior distribution of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ and $y_{\text{obs,partial}}$ corresponds to the observed variables for the subjects with missing values.

- As can be observed, what makes it slightly different than just stochastic regression imputation repeated M times is that it also integrates uncertainty about the parameters used to predict the missing values.

Multiple imputation

Proper MI

- ↪ It is useful to consider the consequences of creating M complete datasets using only the second step (i.e., without resampling from the posterior of θ).
- ↪ In such approach, point estimates would still be valid. However, the total variance V^{MI} computed using Rubin's rules would be too small, because the between imputation variability would not include the uncertainty due to parameter estimation.
- ↪ As a result, the confidence intervals based on V^{MI} would be too narrow.
- ↪ Such imputation models have been termed improper (Rubin, 1987).
- ↪ Treating parameters as random rather than fixed is an essential part of MI.

Multiple imputation

Proper MI

- ↪ A few remarks apply. In a Bayesian analysis, all of data's evidence about parameters is summarised with a likelihood function.

- ↪ As with maximum likelihood, the assumed parametric form of the model may be crucial, if the model is inaccurate, then the posterior distribution may provide an unrealistic view of the state of knowledge about θ.

- ↪ Bayesian analyses require a prior distribution for the unknown parameters. In some problems, prior distributions can be formulated to reflect a state of relative ignorance about the parameters.

Multiple imputation

Proper MI

- ↪ An alternative to sampling from the posterior distribution to reflect uncertainty about the parameter estimates in step 1 is to use the *bootstrap* (see lecture 8).

- ↪ Specifically, we can resample from the observed data, estimate the parameters of the imputation model from the resampled data and then drawing the missing values from the specified normal linear model using these parameter estimates.

Multiple imputation

Choosing the imputation model

- ↪ The imputation model is not intended to provide a parsimonious description of the data, nor does it represent structural or causal relationships among variables.
- ↪ The model is merely a device to preserve important features of the joint distribution of the variables (means, variances, and correlations) in the imputed values.
- ↪ Distinctions between dependent and independent variables and substantive interpretations should be left to postimputation analyses (step 2).
- ↪ Although it is not necessary to have a scientific theory underlying an imputation model, it is crucial for that model to be general enough to preserve effects of interest in later analyses.
- ↪ For instance, suppose that one is interested in examine differences in mean responses between a treatment and a control group. For differences to be preserved in the imputed values, some indicator of group membership should enter the imputation model.