AStA Wirtsch Sozialstat Arch (2014) 8:249–267
DOI 10.1007/s11943-014-0151-8

ORIGINALVERÖFFENTLICHUNG

Multiple Imputation: an attempt to retell the evolutionary process

Florian Meinfelder

Received: 15 October 2014 / Accepted: 6 November 2014 / Published online: 2 December 2014

Abstract Multiple Imputation describes a strategy for analyzing incomplete data that accounts for uncertainty in the missing data by replacing (imputing) each missing value by several 'candidates'. The actual implementation of any Multiple Imputation method is typically computationally expensive which is why the concept has only really caught on around the verge of the new millennium, when the first algorithms for Multiple Imputation had become accessible.

In this article, we are going to give a rough overview of the shortcomings of methods for handling missing data prior to Rubin's work in the late 1970s, and we explore the conceptual innovations that might have lead to Multiple Imputation based on an example, where mean imputation is the steppingstone for more advanced methods. The general concept of Multiple Imputation is explained using a simulated trivariate data set, and the imputation model is based on the standard Bayesian linear model, in order to explain the method as illustrative as possible.

Keywords Missing data · Multiple Imputation

1 Introduction

Multiple Imputation (Rubin 1978, 1987) has come a long way. Initially considered as theoretically intriguing, but almost impossible to implement, it is now widely regarded as one of the most sophisticated and most popular ways of handling missing-data problems. The objective of Multiple Imputation (MI) is to incorporate the uncertainty

F. Meinfelder (✉)
Lehrstuhl für Statistik und Ökonometrie,
Feldkirchenstraße 21, 96052 Bamberg, Germany
e-mail: florian.meinfelder@uni-bamberg.de

in the data that stems from the missing information, while at the same time using all the information available from the observed data as efficiently as possible.

The advent of ready-to-use algorithms—at first as stand-alone software (NORM, Schafer 1999) or as packages in R (e.g. mice, van Buuren and Oudshoorn 1999)—but later on also in commercial software such as SAS (Raghunathan et al. 2001) or Stata (Royston 2004) has played an important role, because the implementation of MI algorithms in statistical standard software made the method more accessible to data analysts. On a side note the term 'method' might already be too specific, as it is rather a concept that comprises various methods which we will discuss later.

Although Multiple Imputation has become very popular, there exist methods not based on the MI framework which yield valid inference for the analysis of incomplete data. Some of these methods are likelihood approaches, such as the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) or Full Information Maximum Likelihood (FIML). Other approaches are based on single imputation in combination with variance correction via resampling methods (e.g. Rao 1996), and an overview of these approaches is given in Münnich (2007). If (co-)variances of parameter estimates are of interest, the Maximum Likelihood variants have to implement them into the algorithm (e.g. supplemented EM, Meng and Rubin 1991). If the likelihood is (too) hard to derive, an alternative is to use a Bayesian Markov Chain Monte Carlo (MCMC) approach, such as Data Augmentation (Tanner and Wong 1987), where random draws from the conditional predictive distribution of the missing data are performed with interchanging random draws from the complete data posterior distribution.

The key difference between the methods listed above and Multiple Imputation is that all of them require prior knowledge about the analysis model (i.e. the parameters to be estimated), whereas under the MI framework the 'imputer' and the 'analyst' can be different entities. A multiply imputed data set can be used for multiple analyses if no so-called *uncongeniality* problem occurs (Meng 1994) which—loosely speaking—means that the scope of the imputer's model does not comprise the scope of the analyst's model. In this case inferences based on the multiply imputed data will be biased if the analyst (rightfully) includes terms into his model which the imputer omitted.

The following section makes an attempt to successively derive more and more sophisticated methods from naïve methods, and leads to stochastic components in imputation settings which might sound counterintuitive to people whose objective in a 'good' imputation is to get as close as possible to the true value—a point we will reconsider in the final section.

Section 3 introduces a simulation design which we use to illustrate the 'evolutionary' process behind MI, by conceptually improving step-by-step upon the previous methods. We investigate the preliminary candidate method within a small Monte Carlo study.

Section 4 gives an overview of the theoretical foundations which were needed to derive the Bayesian model framework behind the MI process, and introduce the relevant distributions for the MI method used throughout this paper (which is also a standard method implemented in numerous MI algorithms).

In Sect. 5 we describe Rubin's combining rules, and explain how MI analyses are conducted in general, before we resume the Monte Carlo study from Sect. 3—this time using Multiple Imputation as the final evolutionary step.

Up to this point we will have neglected some crucial general assumptions regarding the mechanism that governs the missing-data, because at first we want to explore the concept behind MI using the easiest-to-handle (and therefore unrealistic) missingness to convey the basic idea in principle. Section 6 deals with the introduction of the relevant assumptions and completes the notation needed for the discussion of missing-data problems.

The final section addresses advanced issues in Multiple Imputation, and discusses how the algorithms which are nowadays implemented in statistical software also underwent some evolutionary process, and how it affects today's application of MI routines to missing-data problems.

2 Multiple Imputation: an evolutionary approach to explain the concept

Multiple Imputation has not come out of the blue, but is—as basically everything in science—a consequence of logical innovations. We try to recapture some of the cognitive steps behind MI, without claiming this overview to be complete or even to be the actual foundations of Rubin's idea. The most sophisticated method described in this section still leaves quite a large conceptual gap to Multiple Imputation, but we hope our explanations will help readers to understand and appreciate the methods behind MI which have become black boxes inside of algorithms performing the actual imputation.

2.1 Notation

For the remainder of this article we assume a data situation, where the first $k-1$ variables are completely observed, and the last of the k variables is only partially observed. We will let $i = 1, \ldots, n_{obs}$ denote the complete cases over all k variables, whereas $Y_{j,obs}$ (with $j = 1, \ldots, k$) denotes the part of the data, where variable Y_k is observed, and $Y_{j,mis}$ denotes the (bottom) part of the data, where Y_k is missing (observations $n_{obs}+1, \ldots, n$). While capital letters are used to denote univariate or multivariate variables, bold font upper-case letters are used to denote data matrices, and bold font lower-case letters are used to denote data or parameter vectors (even when they are treated as random variables in a Bayesian context). We deliberately refrain from using X as variable names for the observed and Y as variable name for the incomplete variable, as we try to avoid the typical 'regression jargon' of using terms such as dependent or endogenous variable which is not appropriate in the context of imputation modeling. Moreover, algorithms based on chained equations are continuously switching the 'roles' of the involved variables.

2.2 Unconditional mean imputation

There is one quick and dirty way to impute missing data which most data analysts might at least have felt tempted to apply at some point—at least for minor rates of missing data: mean imputation.

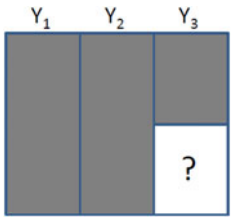

Fig. 1 Missing-data pattern for the trivariate data situation

Mean imputation simply means replacing missing values by the mean of the observed values of the corresponding variable, i.e. $y_{k,imp,i} = \bar{y}_{k,obs}$, for $i = n_{obs} + 1, \ldots, n$. This method will yield only unbiased point estimates for the expectational value of Y_k, and only if the data are missing in a purely random manner (in Sect. 6 we will introduce a more sensible definition for missingness). It is obvious that (co-)variance estimates will be biased as the (co-)variance for the imputed part is zero. Likewise, quantile estimates are affected, depending on the percentage of missing values.

In the following we will use for illustrational purposes a small trivariate data set with $Y = [Y_1, Y_2, Y_3]$ and $n = 500$ observations, where the last 50 % of the observations for variable Y_3 are missing. Because the observations are in random order, the missingness in Y_3 is purely random again.

We assume a data-generating process of the following functional form:

$$\begin{aligned} Y_1 &= N(\mu = 8, \sigma^2 = 9) \\ Y_2 &= 10 - 0.5Y_1 + N(0, 9) \\ Y_3 &= 5 + 0.6Y_1 + 0.5Y_2 + N(0, 2) \end{aligned} \tag{1}$$

The schematic overview for the missing data pattern is given in Fig. 1.

As stated above, mean imputation is no sensible imputation method, but we want to illustrate in detail, why almost all kinds of estimators based on mean imputation are biased. We compare the distribution of Y_3 before the bottom half was removed, and after these values were imputed. Figure 2 shows that Y_3 has become a semi-continuous variable, with 50 % of the observations now being equal to a single value. The sampling variance of Y_3 is halved, because we imputed the missing values with a constant.

And since (unconditional) mean imputation does not account for associations among variables, any multivariate estimator will be biased as well. This is true, among others, for correlations or regression model parameter estimators which can be seen in Fig. 3.

2.3 Conditional mean imputation

The next evolutionary step would be to find a remedy for mean imputation's lack of variance and its blindness to associations with other variables. *Regression imputation* does address these problems. Suppose

$$Y_{k,i} = {Y_{-k,i}}^T \beta + U_i = \beta_0 + \beta_1 Y_{1,i} + \cdots + \beta_{k-1} Y_{k-1,i} + U_i$$

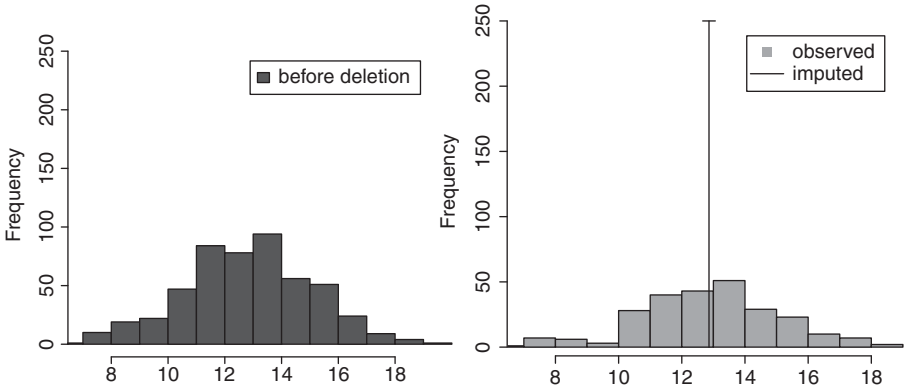

Fig. 2 Mean imputation: histogram of Y_3 before deletion and after imputation

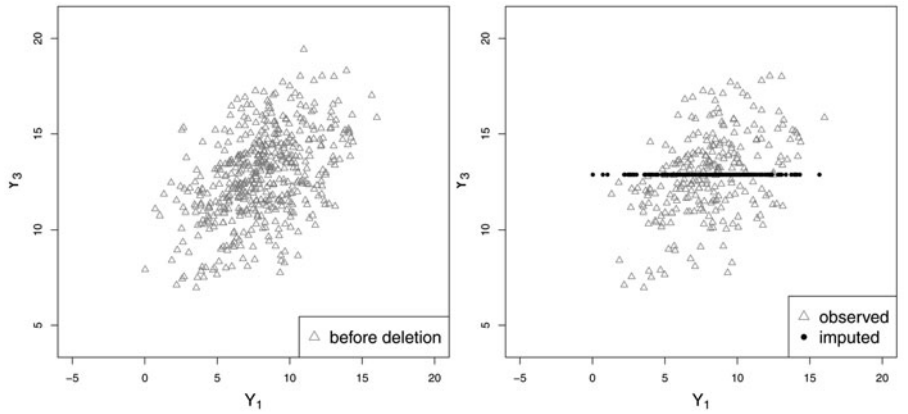

Fig. 3 Mean imputation: bivariate scatterplot of Y_1 and Y_3 before deletion and after imputation

for $i = 1, \ldots, n$, and $U_i \overset{iid}{\sim} N(\mu, \sigma^2)$. Then the method is implemented by estimating the imputation model parameters based on the complete cases to get

$$\widehat{\beta} = (\mathbf{Y_{-k,obs}}^T \mathbf{Y_{-k,obs}})^{-1} \mathbf{Y_{-k,obs}}^T Y_{k,obs},$$

where $\mathbf{Y_{-k,obs}}$ is the design matrix based on the complete cases. Imputations are then simply generated by imputing missing values via regression prediction:

$$y_{k,imp,i} = \mathbf{Y_{-k,mis,i}}^T \widehat{\beta}$$

Unless the imputation model explains Y_k perfectly, the variance of Y_k will still be biased towards zero—after all, it is a *conditional mean* imputation—, but at least the part of the variance explained by the covariates will be preserved.

For our trivariate data example the (Gaussian) kernel density plots of the complete (before the bottom 50 % were removed) and the regression-imputed Y_3 only hints at this problem, as can be seen in Fig. 4. Even the bivariate scatterplot of Y_1 and Y_3 with the partially regression-imputed values is misleading, because the second covariate Y_2 disguises that the imputed part of Y_3 is a deterministic function of Y_1 and Y_2. Looking at these graphical diagnostics might lead to the misconception that the imputed data 'mix' well with the observed data, but Fig. 5 reveals the inherent weaknesses of regression imputation in this trivariate data situation by changing the perspective until we can see that the imputed data can be found in a plane. The graphical diagnostics confirm the presumed shortcomings of regression imputation, and it has become obvious that any imputation algorithm that is purely deterministic will yield biased inferences if the underlying imputation model does not fit the incomplete data perfectly (which is almost a given in any empirical data situation).

Therefore, the next step on our imaginary evolutionary ladder is conceptually a big one: The deliberate introduction of randomness.

2.4 Stochastic regression imputation

The invention of this method is a direct consequence of the observed weaknesses of the regression imputation approach. In order to avoid too strong associations, a stochastic component is added to the predicted values which yields

$$y_{k,imp,i} = \mathbf{y}_{-\mathbf{k},\mathbf{mis},\mathbf{i}}{}^{T}\,\widehat{\beta} + \tilde{u}_i,$$

where $\tilde{u}_i$ is a random draw from a $N(0, \widehat{\sigma}^2)$ distribution with

$$\widehat{\sigma}^2 = \frac{1}{n_{obs} - k} \sum_{i=1}^{n_{obs}} \left(y_{k,obs,i} - \hat{y}_{k,obs,i}\right)^2,$$

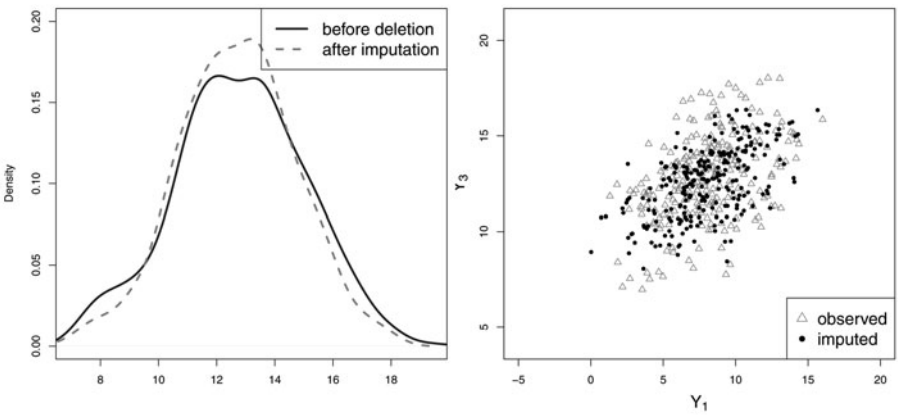

Fig. 4 Regression imputation: univariate kernel density of Y_3 and two-dimensional scatterplot for Y_1 and Y_3

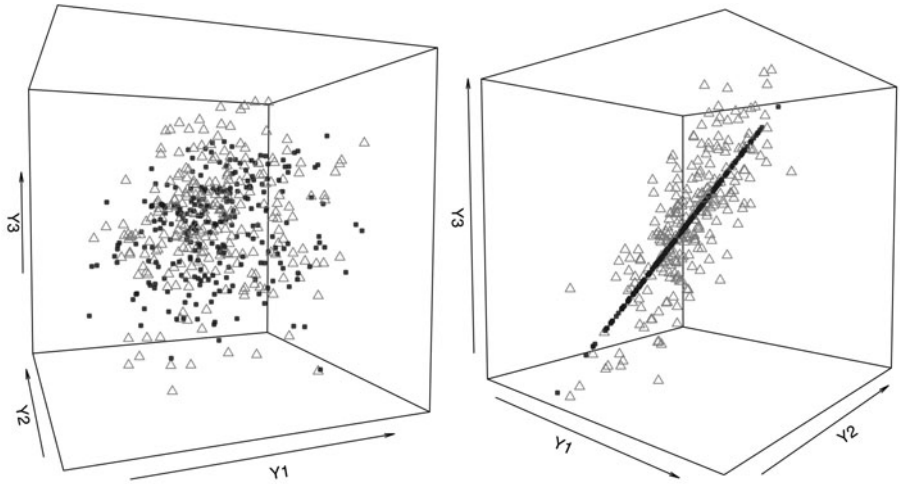

Fig. 5 Regression imputation: three-dimensional scatterplot from two different angles (imputed values in *dark grey*)

with $\hat{y}_{k,obs,i} = \mathbf{y}_{-\mathbf{k},\mathbf{obs},\mathbf{i}}{}^T \widehat{\beta}$. Note that this can also be directly expressed as a random draw of the form

$$y_{k,imp,i} \sim N\left(\mathbf{y}_{-\mathbf{k},\mathbf{mis},\mathbf{i}}{}^T \widehat{\beta}, \widehat{\sigma}^2\right). \tag{2}$$

The univariate kernel density plot and three-dimensional scatterplot which revealed the problems of regression imputation now look inconspicuous, as we can see in Fig. 6, where the same angle was used that illustrated the perfect dependency in Fig. 5. Since this method looks like a sensitive approach for handling missing data without any obvious weaknesses, we will investigate stochastic regression imputation further within a simulation study.

3 A little Monte Carlo study based on stochastic regression imputation

3.1 Setup

The purpose of this Monte Carlo study is to examine whether or not stochastic regression generally yields unbiased inferences. It is centered around the trivariate data set introduced in Sect. 2. As quantities of interest of our analysis serve the expectational value of Y_3, and the regression parameters of the data-generating model for Y_3. Since

$$\mu_3 = E(Y_3) = 5 + 0.6 \cdot E(Y_1) + 0.5 \cdot E(Y_2) = 12.8,$$

we get as true values for our four quantities of interest

$$\theta = [\mu_3, \beta_0, \beta_1, \beta_2]^T = [12.8, 5, 0.6, 0.5]^T.$$

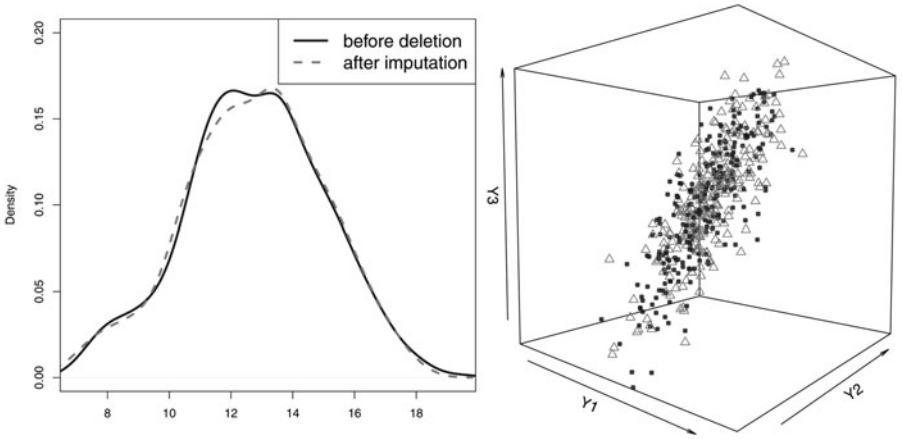

Fig. 6 Stochastic regression imputation: univariate kernel density and three-dimensional scatterplot

Table 1 Overview of results from MC study

	$C(\mu_3)$	$b(\widehat{\mu_3})$	$C(\beta_0)$	$b(\widehat{\beta_0})$	$C(\beta_1)$	$b(\widehat{\beta_1})$	$C(\beta_2)$	$b(\widehat{\beta_2})$
BD	0.94	− 0.00	0.94	− 0.02	0.94	0.00	0.93	0.00
CC	0.95	− 0.00	0.95	− 0.03	0.95	0.00	0.94	0.00
SR	**0.87**	− 0.01	**0.78**	− 0.03	**0.77**	0.00	**0.79**	0.00

We compare stochastic regression (SR) with inferences based on the data before 50 % of Y_3 were deleted (before deletion—BD), and with the complete cases (CC), i.e. after deletion, but before imputation. This method is also known as *listwise deletion*, and yields unbiased (but inefficient) estimators, if the missingness is purely random.

3.2 Results

We expect for both methods (CC and SR) unbiased estimates, but we control for the bias $b(\widehat{\theta})$ anyway (the deviation of the parameter from its estimator, averaged over the 1000 MC cylces), and we also include results for the BD situation as benchmark. The focus, however, is on the coverage rate $C(\theta)$ (with $\alpha = 0.05$) as diagnostic to evaluate stochastic regression, i.e the proportion of the 1000 MC cycles, where the calculated 95 % confidence intervals around $\widehat{\theta}$ covered the true parameter θ.

Table 1 displays the results we get after 1000 MC cycles. As expected, bias is not an issue for any method, but while the coverages for the complete cases (and, of course, for the before-deletion-data) look good, the stochastic regression imputation yields coverages well below the expected 0.95 for all four parameters. But what causes this undercoverage? To answer this question we have to examine the imputation model parameters. For instance, the imputation model parameter vector for our last cycle was $[\beta_0, \beta_1, \beta_2, \sigma^2]^T = [4.75046, 0.62470, 0.49283, 2.00014]^T$. The problem of stochastic regression is that these are just parameter *estimates* based not even on a complete sample but on $\mathbf{Y_{obs}}$. Had we known and used the true parameter vector $[5, 0.6, 0.5, 2]^T$,

stochastic regression imputation would have yielded approximately correct coverage rates.

Apparently, we have to additionally account for randomness in the imputation model parameters, and this leads us to the Bayesian framework behind Multiple Imputation.

4 Drawing parameters from a distribution—the Bayesian framework of MI

4.1 Observed-data posterior and posterior predictive distribution

Without going into detail about Bayesian inference, an important aspect is that parameters are considered to be random variables. Suppose some current data Y are given. Future observations Y^* are then drawn from the so-called posterior predictive distribution

$$f(Y^*|Y) = \int_\theta f(Y^*|\theta) f(\theta|Y) d\theta, \tag{3}$$

where the second part of the integral denotes the posterior distribution of the parameter θ given the current data. If we now replace 'future' by 'missing', and 'current' by 'observed', and if we treat the data as one multivariate random variable Y that consists of an observed and a missing part, i.e. $Y = (Y_{obs}, Y_{mis})$, we can rewrite (3) as

$$f(Y_{mis}|Y_{obs}) = \int_\psi f(Y_{mis}|\psi, Y_{obs}) f(\psi|Y_{obs}) d\psi \tag{4}$$

which is referred to as the posterior predictive distribution of the missing data given the observed data. By applying MC integration techniques we can realize draws from this distribution by alternately drawing from the *observed-data posterior* $f(\psi|Y_{obs})$, and the *conditional predictive distribution of the missing data* (given the observed data) $f(Y_{mis}|\psi, Y_{obs})$. At this stage we have deliberately omitted an important aspect (and, in turn, an important implicit assumption): The mechanism that governs the missingness. We will deal with this complex in detail in Sect. 6.

4.2 The Bayesian linear model

How do the above formulas relate to our findings from the stochastic regression MC study? As a core problem of the method we identified the usage of an imputation model based on *estimators* for the parameters of the imputation model. While we cannot know the true imputation model parameters, we can account for the uncertainty regarding our knowledge by drawing these parameters from a distribution. So instead of working with $\widehat{\beta}$ and $\widehat{\sigma}^2$, we make a random draw for $\psi|Y_{obs}$, where $\psi = [\beta, \sigma^2]^T$. Since the joint observed-data posterior distribution is not available in closed form, we decompose it and get

$$f(\psi|Y_{obs}) = f(\sigma^2|Y_{obs}) f(\beta|\sigma^2, Y_{obs}).$$

Applying the standard Bayesian linear model with an uninformative prior $f(\psi) \propto \sigma^{-2}$, we obtain

$$\sigma^2 | Y_{obs} \sim \frac{\left(\mathbf{y_{3,obs}} - \hat{\mathbf{y}}_{3,obs}\right)^T \left(\mathbf{y_{3,obs}} - \hat{\mathbf{y}}_{3,obs}\right)}{\chi^2_{n_{obs}-3}} \tag{5}$$

and

$$\beta | \sigma^2, Y_{obs} \sim N_3 \left(\mathbf{Y_{-3,obs}}\widehat{\beta}, (\mathbf{Y_{-3,obs}}^T \mathbf{Y_{-3,obs}})^{-1}\sigma^2\right), \tag{6}$$

We implement (5) and (6) within an imputation algorithm by drawing σ^2 first from a scaled-inverse χ^2-distribution, and subsequently drawing β from the trivariate Normal distribution, with the random draw for σ^2 as part of the variance for β. These two steps together form the so-called *posterior* step, i.e. a draw from the *observed-data posterior* $f(\psi | Y_{obs})$.

4.3 Generating imputations

The draw from the *conditional predictive distribution of the missing data* $f(Y_{mis} | \psi, Y_{obs})$ is virtually identical with (2), except that now the imputation model parameters are the draws from the observed-data posterior(s). Plugging in the random draws for β and σ^2 thus yields

$$\mathbf{y_{3,imp}} | \beta, \sigma^2, \mathbf{y_1}, \mathbf{y_2}, \mathbf{y_{3,obs}} \sim N(\mathbf{Y_{-3,mis}}\beta, \sigma^2). \tag{7}$$

While this is conceptually another big evolutionary step towards MI, this method will not improve upon the results we got for the coverages using stochastic regression imputation. We need one more ingredient to utilize the additionally created variation.

5 Multiple Imputation in a nutshell

Donald B. Rubin's outrageous idea was to replace a missing value by several candidates. The final evolutionary step was made in the 1970s while he was working at the US Census Bureau and the concept was first published within a technical report in 1977. Scheuren (2005) provides a good overview of the 'early days', when imputation was referred to as 'allocation', and describes the beginnings of Multiple Imputation.

The idea was way ahead of its time, as Bayesian inference was frowned upon by the majority of statisticians, computing power for the implementation of the needed algorithms was scarce, and statistical analysis software could not handle the idea of using several versions of a data set for analysis purposes (and neither could most of his contemporaries). In Rubin (1987) he derived most of the theoretical proofs for the formulas needed for the application of MI, and in the same year the first edition of the standard compendium 'Statistical Analysis with Missing Data' was published (Little and Rubin 1987).

As mentioned in the introduction, the advent of computing power in combination with flexible algorithms to perform MI played a key role in the proliferation of the

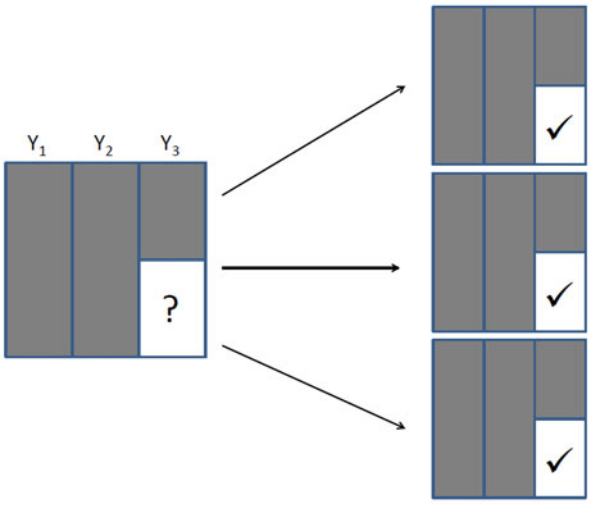

Fig. 7 Multiply imputed data (M = 3)

method. However, the analysis of the multiply imputed data posed another problem, because the combination of the analysis results still requires some work on the analyst's behalf. Nowadays, MI standard analyses such as generalized linear regression or standard hypothesis tests are implemented in most of the statistical software packages (e.g. Stata, SPSS), and they are also integral part of some MI *R* packages e.g. *mi*, (Su et al. 2011).

In the following we will present a general form of what has now become known as *Rubin's Combining Rules*, see e.g. Rässler et al. (2007).

5.1 Rubin's combining rules

Suppose we have carried out the steps from (5), (5) and (7) $M > 1$ times, and we have now M different data sets at our disposal which are identical for Y_{obs}, but (can) differ for Y_{mis}. Figure 7 illustrates this data situation. If we let θ now define the parameter vector of our analysis model, we can use the technique of iterated expectations to show that

$$E(\theta|Y_{obs}) = E(E[\theta|Y_{obs}, Y_{mis}]|Y_{obs}). \tag{8}$$

The averaging of the M posterior expectations over the missing data can be estimated by averaging our estimates for θ from the M different data sets.

From (8) we thus get

$$\widehat{\theta}_{MI} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}^{(m)}, \tag{9}$$

which is referred to as the *MI estimator* for θ. The variance of $\widehat{\theta}_{MI}$ is defined as

$$V(\widehat{\theta}_{MI}) = E(V[\theta|Y_{obs}, Y_{mis}]|Y_{obs}) + V(E[\theta|Y_{obs}, Y_{mis}]|Y_{obs}). \tag{10}$$

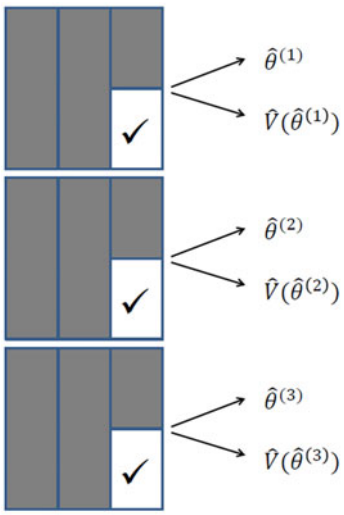

Fig. 8 Running M separate analyses (M = 3)

This is actually the decomposed variance consisting of a *within*-component (the first term) and a *between*-component (the second term). Estimating this *total* variance term yields

$$\begin{aligned} T &= W + (1 + M^{-1})B = \\ &= \tfrac{1}{M} \sum_{m=1}^{M} \widehat{V}\left(\hat{\theta}^{(m)}\right) + (1 + M^{-1}) \tfrac{1}{M-1} \sum_{m=1}^{M} \left(\widehat{\theta}^{(m)} - \widehat{\theta}_{MI}\right)^2 \end{aligned} \tag{11}$$

which is the average of the sampling variances of $\widehat{\theta}^{(m)}$ plus the sampling variance of the M estimators. The term $(1 + M^{-1})$ is required as a correction for $M < \infty$. This means we carry out an MI analysis by extracting the estimator for our quantity of interest θ and the corresponding sampling variance for each of the M imputed data sets independently (as shown in Fig. 8), before we apply the combining rules to get the necessary variance correction. Via a Satterthwaite approximation (Rubin and Schenker 1986) it can be shown that

$$(\widehat{\theta}_{MI} - \theta)/\sqrt{T} \sim t_v \tag{12}$$

with $v = (M - 1)\left(1 + \frac{W}{(1+M^{-1})B}\right)^2$ degrees of freedom.

This ANOVA decomposition allows us to create various measures, the most important one is the *fraction of missing information*

$$\gamma = \frac{r + 2/(v + 3)}{1 + r}, \text{ with } r = \frac{\left(1 + M^{-1}\right) B}{W}.$$

Table 2 MC study results for MI

	$C(\mu_3)$	$\overline{R}(\mu_3)$	$C(\beta_0)$	$\overline{R}(\beta_0)$	$C(\beta_1)$	$\overline{R}(\beta_1)$	$C(\beta_2)$	$\overline{R}(\beta_2)$
BD	0.94	0.41	0.94	1.09	0.95	0.09	0.95	0.08
CC	0.93	0.57	0.95	1.54	0.95	0.13	0.94	0.12
MI	0.93	0.47	0.93	1.50	0.94	0.13	0.93	0.11

An asymptotically identical result is given by $\lambda = \frac{(1+M^{-1})B}{T}$ which is more intuitive: The larger the share of the between variance in the total variance, the less we apparently know about the parameter we are estimating based on the incomplete data set.

5.2 Complementing the MC study

We rerun the simulation study from Sect. 3, but this time we impute the missing values in Y_3 multiply ($M = 10$), each time using the three random draw steps described in (5), (6) and (7). As described above, we apply Rubin's combining rules to the 10 imputations, and calculate the 95 % CI's based on the t_v distribution. Since the previous results confirmed that none of the methods produces any substantial bias, we replace the bias diagnostic with the average CI width $\overline{R}(\theta)$ which is the difference of the 97.5 % quantile and 2.5 % quantile for the distribution of $\widehat{\theta}$, averaged over the 1000 Monte Carlo cycles Table 2 .

While the coverages now get close to the expected value of 0.95, the gain in efficiency expressed by $\overline{R}(\theta)$ is very different for the expectational value of Y_3 on the one hand and the regression parameters on the other hand. While the average CI width for μ_3 is considerably smaller for the MI CI's (in spite of the added between variance) than the complete cases counterpart, we cannot find the same increase in efficiency for the regression parameters. The reason for this can be mathematically proven, and is well documented (see e.g. von Hippel 2007), but it also makes intuitively sense: A data situation with missing values for the 'dependent' variable 'Y' and observed values for the covariates 'X' is identical to a hypothetical situation, where predictions based on some given vectors of covariates are made. But this does not add any additional information to the model (it would be nice, if it did...). It is *exactly* the same situation for a data analysis, where we have only missing values in 'Y' (here: Y_3).

But what happens if we 'reverse' the analysis model and declare Y_1 as the 'dependent' variable, i.e. $Y_{1,i} = \gamma_0 + \gamma_1 Y_{2,i} + \gamma_2 Y_{3,i} + V_i$, with $V_i \sim N(0, \tau^2)$? This mimics a data situation, where one of the covariates (again: Y_3) has missing values. Calculation of the new true model parameters is slightly tedious, but based on the data generating model from 1 it can be shown that the covariance matrix of $Y = [Y_1, Y_2, Y_3]$ is given by

$$\Sigma = \begin{pmatrix} 9.00 & -4.500 & 3.1500 \\ -4.50 & 11.250 & 2.9250 \\ 3.15 & 2.925 & 5.3525 \end{pmatrix},$$

and thus we can derive the new true model parameters

$$\gamma = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} -0.1742160 \\ -0.6445993 \\ 0.9407666 \end{pmatrix}.$$

Table 3 MC study results for the 'reversed' analysis model

	$C(\gamma_0)$	$\overline{R}(\gamma_0)$	$C(\gamma_1)$	$\overline{R}(\gamma_1)$	$C(\gamma_2)$	$\overline{R}(\gamma_2)$
BD	0.95	1.75	0.94	0.10	0.95	0.15
CC	0.94	2.48	0.96	0.14	0.95	0.21
MI	0.93	2.13	0.94	0.12	0.94	0.18

Re-analysing the data from the MC study now yields the results displayed in Table 3. For all three estimators the average MI CI width is almost exactly half-way in between the average CI width for the *before deletion* and the *complete cases* data situation. The reason for this gain in efficiency is that now the associations between $Y_{1,mis}$ and $Y_{2,mis}$ are utilized in the multiply imputed data.

This small simulation study demonstrates that Multiple Imputation leads in general to more efficient inference, compared to inference based on the complete cases, but there are exceptions to the rule, whenever the estimator based on the imputed data does not use more information to estimate the model than the complete-case estimator.

6 Relevant assumptions and theoretical concepts

So far we have not dealt yet with one important aspect of incomplete data: What governs the process of missingness? A neat 'trick' to describe the missing-data mechanism is to introduce an additional response indicator R variable (see e.g. Little and Rubin 2002) which is defined as

$$R_{i,j} = \begin{cases} 1, \text{ if } Y_{i,j} \text{ is observed.} \\ 0, \text{ if } Y_{i,j} \text{ is missing.} \end{cases}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

Thus, the joint distribution of Y and R is given by

$$f(Y, R|\theta, \xi) = f(Y|\theta) f(R|Y, \xi, \theta), \quad (\theta, \xi) \in \Omega_{\theta,\xi}$$

The *observed-data posterior* $f(\psi|Y_{obs})$ from (4) is itself based on the *observed-data likelihood* $f(Y_{obs}|\psi)$ Therefore, the completely denoted observed-data likelihood is defined as

$$\begin{aligned} f(Y_{obs}, R|\theta, \xi) = L(\theta|Y_{obs}, R, \xi) &= \int f(Y_{obs}, Y_{mis}, R|\theta, \xi) dY_{mis} \\ &= \int f(Y_{obs}, Y_{mis}|\theta) f(R|Y_{obs}, Y_{mis}, \xi, \theta) dY_{mis}, \end{aligned} \tag{13}$$

where ξ are unknown parameters pertaining to the missing-data mechanism.

6.1 Missingness mechanisms and ignorability

The now famous classification of missing-data mechanisms was first described by (Rubin 1976), when he introduced the terms *missing at random* and *observed at random*, and was refined into its final version by Little and Rubin (1987). If we assume

that ξ and θ are independent, i.e. from a Bayesian perspective the prior distribution $f(\theta, \xi) = f(\theta)f(\xi)$, we can write $f(R|Y, \theta, \xi) = f(R|Y, \xi)$. This assumptions is called *distinctness* in the missing-data literature (see e.g. van Buuren 2012).

The first of three levels of missing-data mechanisms which we described above as 'purely random' is actually labeled as Missing Completely at Random (MCAR), and means that the missing-data mechanism only depends on the unknown parameters ξ, i.e.

$$f(R|Y_{obs}, Y_{mis}, \xi) = f(R|\xi). \tag{14}$$

This missing-data mechanism typically only occurs in missing-by-design data situations, such as matrix sampling in questionnaires.

A slightly more restrictive version is called *missing at random* (MAR), which is somewhat misleading, as the missingness mechanism is only conditionally on the observed part of Y independent, such that

$$f(R|Y_{obs}, Y_{mis}, \xi) = f(R|Y_{obs}, \xi). \tag{15}$$

A frequently used example for this mechanism is that we have missing values for 'income', and that older people have a higher propensity of refusing to answer a question about it (e.g. for fear of burglary). If 'income' and 'age' are positively correlated, an estimator of the average income based on the complete cases would be biased towards zero, but conditioning on the variable 'age' makes the missingness random again.

If, however, the propensity of refusing to answer the income question depends on 'income' itself (e.g. because high-income earners are more afraid of burglary), i.e. the missing-data mechanism depends on Y_{mis} as well, the missing-data mechanism is labeled *Missing Not at Random* (MNAR), and cannot be (completely) 'healed'. Inference based on incomplete data with an MNAR mechanism will be biased.

The *observed-data likelihood under ignorability* requires the MAR as well as the *distinctness* assumption, and allows us to split (13) into two components, such that

$$L(\theta, \xi|Y, R) = \int f(Y_{obs}, Y_{mis}|\theta) f(R|Y_{obs}, Y_{mis}, \xi, \theta) dY_{mis} = f(R|Y_{obs}, \xi) L(\theta|Y_{obs}),$$

where the first part describes the model for the missing-data mechanism (which is usually not relevant for the analysis), and the second part is the observed-data likelihood under ignorability that leads to the observed-data posterior that features in (4).

Some researchers claim that they do not use MI, because of its "unrealistic assumptions" , but while MAR might be in some data situations debatable, it is—without access to auxiliary information—a necessary assumption for *any* method applied to analyzing incomplete data (even using complete-case analysis is a 'method', and it requires the stronger MCAR assumption to get only unbiased, but not efficient estimators). MAR or distinctness cannot be tested (we can only test against MCAR), which is why we have to make these assumptions, but even if ignorability does not hold (because the mechanism is MNAR), wrongfully assuming it and applying MI will still yield less biased estimators (e.g. because 'income' can be partially explained by 'age') than using the complete cases.

7 Discussion

Stochastic regression and multiple imputation share the concept of introducing randomness in order to preserve the actual data structure, unlike regression imputation which tries to retrieve each missing value with the highest possible precision. Unfortunately, this is a valid approach if the imputation model fits the data perfectly—but as soon as some stochastic component is involved, trying to predict each missing value rather than to preserve distributions (and to account for missingness-related additional uncertainty) will yield biased inferences. But the hypothetical situation of a perfect imputation model might explain why some researchers have tried to evaluate imputation methods using cross-validation schemes (typically based on so-called 'hitrates') which are suited for classification tasks but not for inferential statistical analysis with missing data.

One unresolved issue is our differentiating usage of imputation model parameters ψ, and the parameters pertaining to the data-generating model θ. This distinction is sometimes omitted in the literature, and again, some sort of evolutionary process might be one of the reasons behind it.

To the best of our knowledge, the first working stand-alone MI software was NORM by Joseph Schafer, and the underlying algorithms are described in detail in Schafer (1997). Like virtually all of the algorithms described in the early publications during the 1980's and early 1990's, NORM assumes a multivariate joint distribution for the data. By modeling the whole data jointly, usually no distinction between ψ and θ has to be made, because the joint (e.g. the multivariate normal in NORM) distribution governs the analysis, and all variables in the data are part of the imputation model. While these algorithms are mathematically elegant, the implementation becomes computationally challenging for data sets with many variables, and the assumed joint distribution which is used for the data model can not be sustained if the variables have different measurement levels. Approaches to adapt to mixed scale-type data, while preserving the benefits of this so-called *joint modeling* (JM) approach, still encounter dimensionality problems in data sets with many variables (e.g. the 'mix' algorithm by Schafer 2010). One solution for this problem is to model each variable conditionally on (all the) other variables. Incidentally, this is also what we have implemented in Sect. 3. The benefit is straightforward: We can model each variable separately using the link function $f(y)$ of the corresponding generalized linear model. For instance, a Bayesian version of the binomial logit model can be used to model dichotomous variables, or, as demonstrated in the MC study, a linear model for continuous data. The basic idea to implement this approach into an algorithm is to use the concept of chained or switching regressions, where each variable is modeled on all the other variables, i.e. imputations are conditional and univariate. The first algorithm to adapt this idea has been *Multivariate Imputation by Chained Equations* (MICE, van Buuren and Oudshoorn 1999), and meanwhile, most MI algorithms are based on this concept (e.g. IVEware, Raghunathan et al. 2002). Even most of the major statistical software packages like Stata (StataCorp LP 2013) feature MI algorithms based on these *fully conditional specification* (FCS) models. A comprehensive overview can be found in van Buuren (2012).

But FCS comes at a prize: Modeling conditionally means that we do not know whether or not the joint distribution actually exists, or if we encounter an incompatible Gibbs problem (see Rubin 2003), although so far no empirical application has been published, where incompatibility had caused problems. Additionally, FCS algorithms can handle incomplete high-dimensional data, but the imputation model might have to be reduced. This brings us back to the distinction between ψ and θ, and the problem of uncongeniality which we briefly addressed in the introductory section: The evolution of more sophisticated and flexible MI algorithms in combination with easier accessability has brought higher demands to the functionalities of MI algorithms, and sometimes the imputer's model no longer can anticipate (or incorporate) every analysis which might be carried out at a later stage on the incomplete data.

One general answer to imputing large-scale data is the introduction of MI algorithms based on data-mining procedures such as CART (Burgette and Reiter 2010), or the usage of model selection methods (e.g. Koller-Meinfelder 2009). However, in both cases the imputation model is a compromise of automatically including the variables into the imputation model which jointly explain the incomplete variable itself (and/or the missingness for this variable) reasonably well. Such non-tailored MI solutions can cause problems for sophisticated analyses. For instance, Bernardini Papalia et al. (2013) analyze small area estimation situations, where the fraction of missing information γ is high, and the usually suggested 'M = 5' imputations lead to inefficient standard errors, because the efficiency of an estimator based on a finite number of imputations relative to the fully efficient infinite-M imputation estimator is given by $RE = \left(1 + \frac{\gamma}{M}\right)^{-1/2}$ (see Rubin 1987).

While new challenges have occurred in missing-data related research, the evolution of Multiple Imputation has helped to develop a general perception for the problems missing data can inflict on statistical inference, and the development of powerful algorithms has triggered the search for ever more complicated tasks to which the concept could be applied to.

MI example

We want to illustrate the application of Rubin's combining rules by using a little hypothetical example: Suppose we have monthly income information (in thousand Euros) for 4 men and 5 women, where the income information was missing for observation one and three of the female subgroup, and thus multiply (m = 3) imputed: The analysis objective is to investigate if men have a higher average income than women, and therefore a test against the null hypothesis $H_0 : \mu_{women} - \mu_{men} \leq 0$ is conducted. We run

Table 4 Hypothetical example: income of five women and four men (imputations in bold font)

Observation number	1	2	3	4	5
Income men	2.50	4.90	3.60	2.80	–
Income women (m=1)	**1.80**	3.90	**4.20**	3.20	2.40
Income women (m=2)	**2.60**	3.90	**4.00**	3.20	2.40
Income women (m=3)	**2.70**	3.90	**4.30**	3.20	2.40

Table 5 Results for the 3 independent analyses

$\bar{x}_2$	$D = \bar{x}_1 - \bar{x}_2$	$s^2_{x_2}$	$\widehat{V}(D)$
3.100	0.350	1.010	0.482
3.220	0.230	0.532	0.359
3.300	0.150	0.635	0.385

an unpaired t test assuming unknown, but equal variances, where $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_{12}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ and $S_{12} = \sqrt{\frac{(n_1-1)s^2_{x_1} (n_2-1)s^2_{x_2}}{n_1+n_2-2}}$. For the men we get $\widehat{\mu}_{men} = \bar{x}_1 = 3.45$ and $s^2_{x_1} = 1.15$. The results for the three imputed versions for women and the combined results are given in Table 5. Note that the estimator for our quantity of interest is $\widehat{\theta} = D = \bar{x}_1 - \bar{x}_2$, and that its sampling variance is defined as $\widehat{V}(D) = S^2_{12}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.

The MI estimator is given by $D^{MI} = \frac{1}{3}\sum_{m=1}^{3} D^{(m)} = 0.243$, and the within and between variance are given by $W = \frac{1}{3}\sum_{m=1}^{3} \widehat{V}(D)^{(m)} = 0.408$, and $B = \frac{1}{3-1}\sum_{m=1}^{3}\left(D^{(m)} - D^{MI}\right)^2 = 0.010$. The total variance according to (11) yields $T = 0.422$, and the degrees of freedom according to (12) yield $\nu = 1950.1$ which makes the statistic virtually normally distributed for all practical purposes, and we get $t = \frac{0.243-0}{\sqrt{0.422}} = 0.374$ which means that we cannot reject the null for any meaningful α.

In this example, the total variance T is completely dominated by the within variance, because the 3 imputations yield comparatively similar results.

Acknowledgements The author would like to thank the editor and two reviewers, who discovered mistakes of all sorts, added valuable suggestions, asked for clarifications, and thus, helped to improve the quality and readability of this paper considerably.

References

Bernardini Papalia R, Bruch C, Enderle T, Falorsi S, Fasulo A, Hernandez-Vazquez E, Ferrante M, Kolb J-P, Münnich R, Pacei S, Priam R, Righi P, Schmid T, Shlomo N, Volk F, Zimmermann T (2013) Best practice recommendations on variance estimation and small area estimation in business surveys: software code/BLUE-ETS. http://www.blue-ets.istat.it/

Burgette LF, Reiter JP (2010) Multiple imputation for missing data via sequential regression trees. Am J Epidemiol 172(9):1070–1076

Dempster AP, Laird N, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B 39:1–38

Koller-Meinfelder F (2009) Analysis of incomplete survey data: multiple imputation via Bayesian bootstrap predictive mean matching, opus, Bamberg. http://opus4.kobv.de/opus4-bamberg/frontdoor/index/index/docId/200

Little RJ, Rubin DB (1987) Statistical analysis with missing data, Wiley series in probability and mathematical statistics. Wiley, New York

Little RJ, Rubin DB (2002) Statistical analysis with missing data, 2 edn. Wiley, New York.

Meng X-L (1994) Multiple-imputation inferences with uncongenial sources of input. Statis Sci 9:538–558
Meng X-L, Rubin DB (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. J Am Statis Assoc 86(416):899
Münnich R (2007) Discussion on non-Bayesian multiple imputation. J Off Statis 23(4):455–461
Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Methodol 27:85–95
Raghunathan TE, Solenberger PW, van Hoewyk J (2002) IVEware: imputation and variance estimation software
Rao JN (1996) On variance estimation with imputed survey data. J Am Statis Assoc 91:499–506
Rässler S, Rubin DB, Zell ER (2007) 19 incomplete data in epidemiology and medical statistics. *in* Epidemiology and medical statistics, Vol 27 of handbook of Statistics. Elsevier, pp 569–601
Royston P (2004) Multiple imputation of missing values. Stata J 4(3):227–241
Rubin DB (1976) Inference and missing data, Biometrika 63:581–592
Rubin DB (1978) Multiple imputation in sample surveys—a phenomological Bayesian approach to nonresponse. in: Proceedings of the survey research method section of the American Statistical Association, pp 20–40
Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
Rubin DB (2003) Nested multiple imputation of NMES via partially incompatible MCMC. Statistica Neerlandica 57(1):3–18
Rubin DB, Schenker N (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J Am Statis Assoc 81:366–374
Schafer JL (1997) Analysis of incomplete multivariate data. Chapman and Hall, London
Schafer JL (1999) NORM—multiple imputation under a normal model, version 2.03
Schafer JL (2010) mix: estimation/multiple imputation for mixed categorical and continuous data. http://CRAN.R-project.org/package=mix
Scheuren F (2005) Multiple imputation. Am Stat 59(4):315–319
StataCorp LP (2013) Stata multiple-imputation reference manual: release 13. http://www.stata.com/manuals13/mi.pdf
Su Y-S, Gelman A, Hill J, Yajima M (2011) Multiple imputation with diagnostics (mi) in R: opening windows into the black box. J Statis Software 45(2):1–31. http://www.jstatsoft.org/v45/i02/
Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation (with discussion). J Am Statis Assoc 82:528–550
van Buuren S (2012) Flexible imputation of missing data, Chapman & Hall/CRC interdisciplinary statistics series. CRC Press, Boca Raton
van Buuren S, Oudshoorn K (1999) Flexible multivariate imputation by MICE
von Hippel PT (2007) Regression with missing YS: an improved strategy for analyzing multiply imputed data. Sociol Methodol 37(1):83–117