Generating missing values for simulation purposes: a multivariate amputation procedure

Rianne Margaretha Schouten, Peter Lugtig & Gerko Vink

Taylor & Francis
Taylor & Francis Group

Generating missing values for simulation purposes: a multivariate amputation procedure

Rianne Margaretha Schouten, Peter Lugtig and Gerko Vink

Department of Methodology and Statistics, University of Utrecht, Utrecht, Netherlands

ABSTRACT

Missing data form a ubiquitous problem in scientific research, especially since most statistical analyses require complete data. To evaluate the performance of methods dealing with missing data, researchers perform simulation studies. An important aspect of these studies is the generation of missing values in a simulated, complete data set: the amputation procedure. We investigated the methodological validity and statistical nature of both the current amputation practice and a newly developed and implemented multivariate amputation procedure. We found that the current way of practice may not be appropriate for the generation of intuitive and reliable missing data problems. The multivariate amputation procedure, on the other hand, generates reliable amputations and allows for a proper regulation of missing data problems. The procedure has additional features to generate any missing data scenario precisely as intended. Hence, the multivariate amputation procedure is an efficient method to accurately evaluate missing data methodology.

1. Introduction

Missing data form a ubiquitous problem in scientific research, especially since most statistical analyses require complete data. Of all possible methods for dealing with missing data (e.g. weighting procedures, likelihood-based methods), multiple imputation is an appealing solution since it results in complete data sets without losing information. This is because multiple imputation replaces the missing values with estimated values based on the observed data [1–3]. Because multiple imputation is now available in several statistical software packages such as SPSS, Stata, SAS and R, the application of multiple imputation has become straightforward.

For a proper usage of methods dealing with missing data, a thorough and valid evaluation of the performance of these methods is vital. Because scientific research increasingly

CONTACT Rianne Margaretha Schouten ✉ r.m.schouten@uu.nl 📧 Department of Methodology and Statistics, University of Utrecht, Sjoerd Groenman building, Padualaan 14, 3584 CHUtrecht, Netherlands

relies on missing data methodologies, it is important to understand under which circumstances a certain technique can and should be used. In addition, new missing data methods are continuously being developed and improved (e.g. [4–6]). An accurate evaluation procedure is essential to know whether such a methodology functions properly. In this paper, we will focus on the process of generating missing values; a key procedure in the evaluation of the performance of missing data techniques.

In general, a missing data methodology is evaluated by means of simulations. Such simulation studies generally have four steps:

(1) A multivariate, complete data set is simulated and considered the population of interest.
(2) This data set is then made incomplete.
(3) The incomplete data are estimated by means of the missing data method.
(4) Statistical inferences are obtained for the original, complete data set and after dealing with the missing values. A comparison of these inferences gives an indication of the performance of the missing data method.

A key part in the evaluation of a missing data methodology is step 2; the generation of missing values. We refer to the process where missingness is induced in complete data as *amputation*.

In current simulation studies, missing values are generated one variable at a time. With this univariate amputation approach, it can be difficult to appropriately control the characteristics of a missing data problem. For instance, [7,p.11] intended to generate missing values for 70% of the cases, but as a result of their approach 77% of the cases were made incomplete. Another issue, that we will return to, is that the missing data structures do not reflect what was intended by the researcher. As a result, missing data methods may be evaluated under the wrong conditions and invalid conclusions about their inferential performance may be drawn.

A valid amputation procedure is critical for the evaluation of missing data methods as it defines the missing data problem. Hence, we will put the methodological validity and the statistical nature of the current amputation practice under close investigation. Further, we will outline an alternative approach for generating any kind of missingness in any data set. Because our amputation procedure generates missing values in a multivariate way, issues with unreliable and inconsistent missing data generation can be overcome. Additionally, to make the multivariate amputation procedure available to a broad audience, we implemented the methodology as the function `ampute` [8] in R-package `mice` [9].

In the next section, we will explore the way researchers currently generate missing values (Section 2.1) and the reason why this approach may be inappropriate for generating reliable missing data problems (2.2). We will then explain the methodology of the multivariate amputation procedure (3.1) and detail our argument that the multivariate amputation method overcomes the problems with the current practice (3.2). Furthermore, we will demonstrate the performance of `ampute` with simulation research (Section 4) and make its improvement on the current practice evident (Section 5). We will conclude and discuss additional applications of `ampute` in Section 6.

2. Univariate amputation

2.1. Univariate amputation

In current evaluations of missing data methodologies, every researcher has to develop an ad-hoc approach to generating missing values. Most of these amputation methods are designed for a particular data set. For instance, Sullivan et al. [10,p.4] simulated data 'from the model $Y_i = 0.30T_i + \beta_2 X_i + e_i$, with X and $e \sim N(0, 1)$' while others use real data such as 'the complete SHS data set [with variables] age, gender, history of diabetes, CVD status, and three serial measures of Scr [i.e. serum creatinine]' [11,p.3]. Because the variables of these data sets have a specific covariance structure and tailor-made amputation approaches are used, the missing data situations these researchers study are unique.

All ad-hoc amputation techniques have in common that missing values are generated in a univariate way. Univariate amputation means that missingness is generated one variable at a time. For example, Sullivan et al. [10] generated missing values in variable Y. For Shara et al. [11,p.3], 'the outcome variable of interest was CVD' and variable Scr was the incomplete variable.

In almost all real datasets, missingness is present for multiple variables. When such a situation needs to be mimicked in a simulation study, researchers repeatedly use a univariate amputation procedure for every variable that needs missing values. Missing values are generated first in Y_1, second in Y_2, and so on. In other words, the missingness on one variable is generated separately and independently from the missingness on any other variable. An example is the simulation study of [12,p.3] who describe that 'In the MAR setting, the probability to be missing for each of X, Y or Y' was independently set to be conditional on the exposure E with OR = 5.' Other examples of studies where this procedure is followed are [13–15]. From here on, we will continue to refer to this class of approaches as *stepwise univariate amputation*.

2.2. Problems with stepwise univariate amputation

Probabilities to be missing on one or multiple variables can be independent, be related to each other, depend on values of covariates, or depend on the variables themselves. The situation where the missingness does not depend on any other variable in the dataset is referred to as Missing Completely At Random i.e. MCAR, [16]. Alternatively, the probability of being missing in a certain variable, say Y_1, can be based on the values of another variable, say X_1. This situation is called Missing At Random [i.e. MAR, 16] [3,p.63,64] advised to relate the missingness on variable Y_1 to the values of X_1 by means of a continuous probability function, such as a logistic distribution function. However, in practice it is common to use the X_1 values to divide the participants into groups. Then, a specific probability value is specified for each of these groups (e.g. [12,17,18]). A third type of missing data problem is created when the missingness in a variable is based on the variable values themselves. In that situation, the probability of a case being missing in variable Y_1 depends on its value on Y_1. We call this a Missing Not At Random mechanism i.e. MNAR, [1,3,p.31,32]. In practice, researchers aim at generating at least one of the three missingness mechanisms to evaluate missing data methodology.

Problems with stepwise univariate amputation appear especially – but not only – when multiple variables are amputed. The stepwise univariate amputation method is problematic

for a proper regulation of the characteristics of missing data problems. For instance, consider a simple characteristic of any missing data problem: the percentage of cases with missing values. When a researcher generates 50% MCAR missing values in variable Y_1, each case has a probability of 0.5 of being missing in Y_1. When the same procedure is followed for a second variable, say Y_2, the probability that a case has missing values on both Y_1 and Y_2 is $0.5 \cdot 0.5 = 0.25$. Hence, to reach a certain overlapping missingness percentage, a different individual probability has to be used. Fortunately, with MCAR missingness it is quite straightforward to estimate that probability (see Appendix 1).

With MAR missingness, on the other hand, it is challenging to determine how the probabilities should be adapted. For example, consider a situation where we induce missingness in variables Y_1 and Y_2 and base our amputations on the values of a normally distributed variable X_1. When we use the standard right-tailed logistic distribution function of Figure 1, we will induce 50% missingness in variable Y_1 and 50% in variable Y_2. However, the joint missingnes percentage (i.e. the percentage of cases with missing values on both Y_1 and Y_2) will become 29% instead of the intuitive 25% (Appendix 2). To obtain the desired missingness percentage, we can move the logistic distribution function horizontally. For instance, the right-tailed distribution function can be shifted with 0.922 (Figure 1) which will result in 50% joint missingness (Appendix 2). Unfortunately, this calculation requires extensive mathematical knowledge and becomes even more complicated when missingness is created in three or more variables.

Moreover, any shift of the logistic distribution function also influences other parameters of the missing data problem, such as the impact of the missing values on the statistical outcome. Since the logistic distribution function has an asymptote at $y = 1$, a certain part of the distribution of Y will almost definitely be amputed. With the right-tailed distribution function, cases with extreme X_1 values will have $P(\text{missing}) = 0.99$. Depending on the correlation between X_1 and Y_1, this means that high Y_1 values will almost certainly be made missing. When the logistic distribution function is shifted, not only will more values be
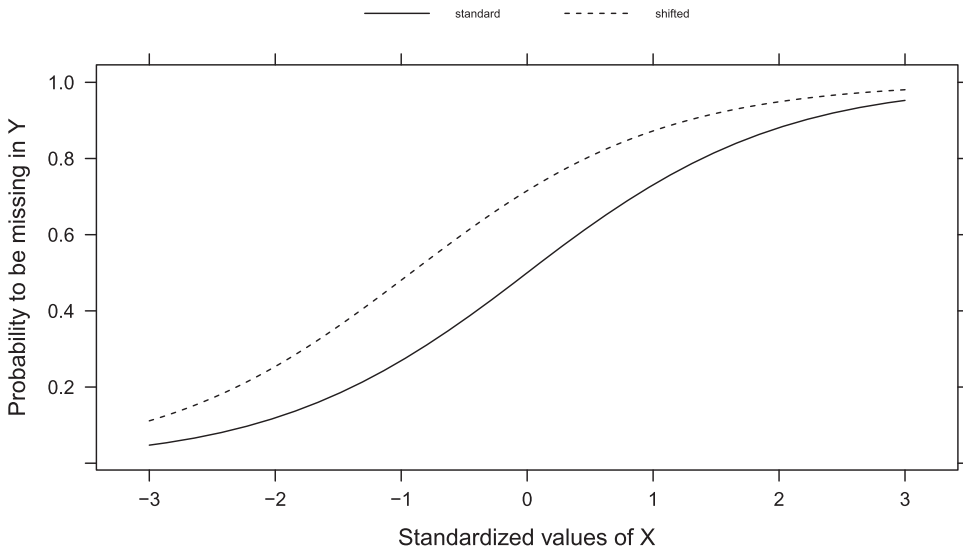

Figure 1. Standard ($b = 0$) and shifted ($b = 0.922$) right-tailed logistic distribution functions.

amputed, but the largest part of these missings will occur in the right tail of the distribution of Y_1. This causes the observed distribution of Y_1 to become more skewed than initially intended.

When researchers are not aware of the consequences of using a stepwise univariate amputation procedure, simulation results may not be interpreted in the right context. For instance, when missingness is created in a data set with low correlations, intended MAR missingness patterns may more reflect a MCAR mechanism [19]. In such situations, the performance of the missing data method may be evaluated as sufficient when in fact the evaluation procedure – and not the missing data method – leads to the finding. The opposite could be true as well, when a researcher shifts the right-tailed logistic distribution function to the left to obtain a certain multivariate missingness percentage, but is not aware that this shift increases the severity of the MAR mechanism. Hence, the missing data method may be falsely categorized as underperforming.

3. Multivariate amputation

3.1. Multivariate amputation

In an ideal situation, a missing data methodology is tested with a realistic missing data problem. For the generation of multivariate missing data scenarios that mimic real-life missingness, [20,p.110–113] proposed an idea for a multivariate amputation that has only been briefly discussed twice [21,22]. We build upon this method and in this paper develop an amputation procedure that can generate sophisticated missing data problems.

Figure 2 shows a schematic overview of the resulting multivariate amputation procedure. The procedure requires a multivariate, complete data set of n participants and m variables. The result of the procedure, as can be seen on the right-hand side of Figure 2, consists of multiple subsets with either incomplete or complete data. These subsets are merged to obtain an incomplete version of the original data set.

The amputation procedure starts with the researcher deciding what kind of missing data patterns he desires to establish. A missing data pattern is a particular combination of variables with missing values and variables remaining complete. For instance, a researcher may wish for some participants to have missing values on the first two variables, while for other participants values should be missing on variables 1 and 3. A practical example of a missing data pattern is a situation where some participants do not show up for the second round of a longitudinal study. Hence, part of the cases should have missing values on specifically the variables of the second wave.

Based on the number of missing data patterns k, the complete data set is randomly divided into k subsets. The size of these subsets may vary. For example, we can assign a frequency value of $\frac{2}{3}$ to a first missing data pattern and a value of $\frac{1}{3}$ to a second missing data pattern. As a result, $\frac{2}{3}$ of the cases becomes part of subset 1. These participants are then called *candidates* for missing data pattern 1. Since all cases are allocated to a subset and every case is candidate for one pattern only, the sum of the frequency values should be 1. Note that at this stage it is not yet decided whether the participants will receive missing values; they are merely candidates for one of the k missing data patterns.

Before we explain how to calculate a so-called weighted sum score for each candidate, it is important to know that the weighted sum scores are used to determine whether a data
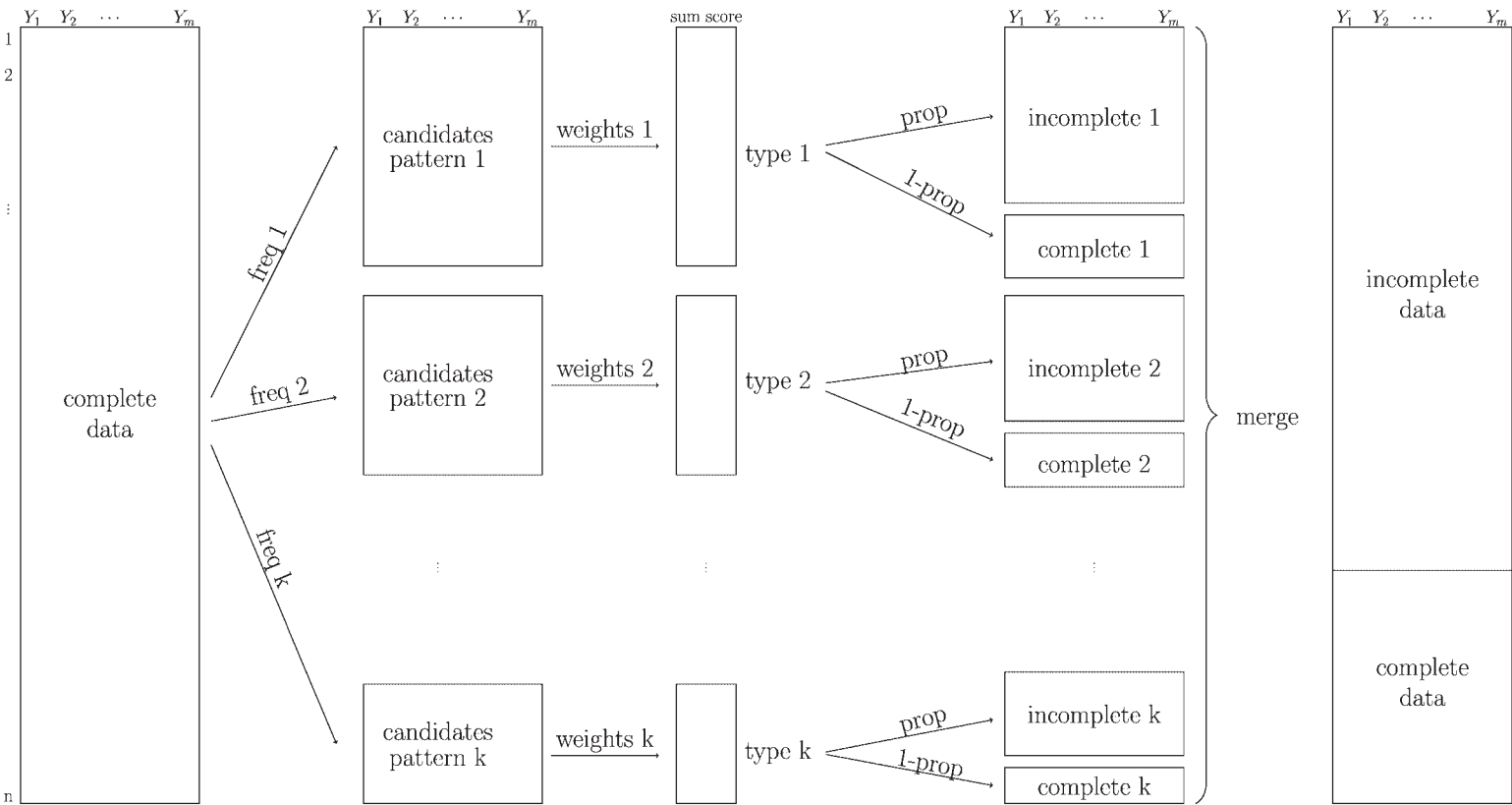
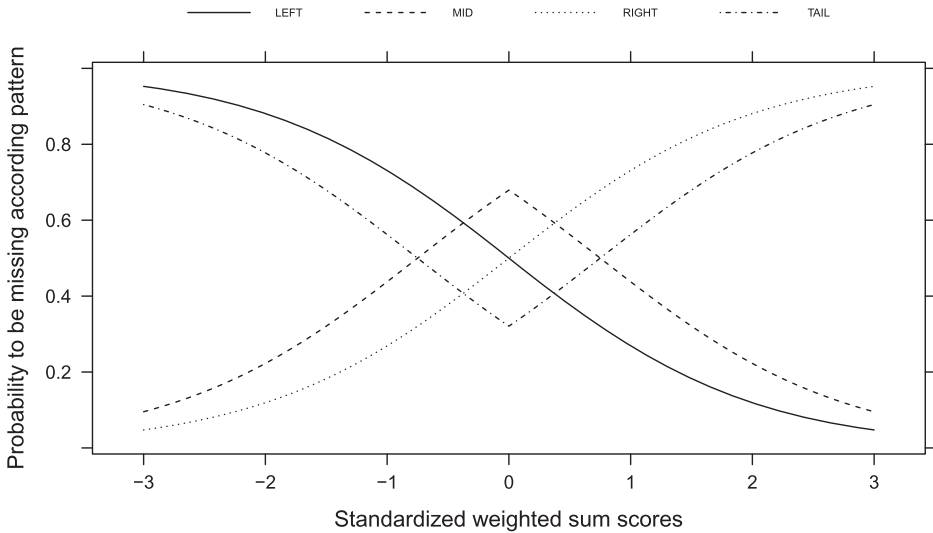**Figure 2.** Schematic overview of the multivariate amputation procedure.

Figure 3. Four variants of the logistic distribution function based on [3,p.64].

cell becomes missing or not: based on his weighted sum score, each candidate receives a probability of being missing for a given variable. For the allocation of these probabilities, we apply one of four possible logistic distribution functions on the weighted sum scores (Figure 3). It is possible to choose a different logistic distribution function for every missing data pattern. For instance, if a right-tailed (RIGHT) type of missingness is used, candidates with high weighted sum scores will receive a high probability of being missing. With a left-tailed (LEFT), centred (MID) or both-tailed (TAIL) missingness type, higher probability values are given to the candidates with low, average or extreme weighted sum scores respectively. Since the weighted sum scores determine the missingness probabilities, the calculation of weighted sum scores is an important part of the amputation procedure.

Weighted sum scores are calculated as the outcome of a linear regression equation where the coefficients are determined by the researcher. The weighted sum score of case i is calculated as follows:

$$wss_i = w_1 \cdot y_{1i} + w_2 \cdot y_{2i} + \cdots + w_m \cdot y_{mi},$$

where $\{y_{1i}, y_{2i}, \ldots, y_{mi}\}$ is the set of variable values of case i and $\{w_1, w_2, \ldots, w_m\}$ are the corresponding pre-specified weights. One weight per variable governs the impact of that variable on the formation of the sum score. Variables with higher weights will have a larger influence on the size of the weighted sum score than variables with lower weights. For instance, if variables Y_1 and Y_2 have weight values of 4 and 2, respectively, the importance of Y_1 is twice as large as that of Y_2. The sign of the weight values influences whether a weighted sum score increases or decreases. A positive weight will increase the weighted sum score while a negative weight will have a decreasing effect. Furthermore, the weight values can value by missing data pattern. For example, variable Y_1 can have a weight value of 4 in the first pattern, but a weight value of -2 in the second pattern.

An important feature of the multivariate amputation procedure lies in the possibility to choose a weight value of zero. A zero weight indicates that the values of that variable play no role in the calculation of the weighted sum scores. Consequently, the generation of

MCAR missingness can be seen as the situation where all weight values are zero. Since the probability of being missing with MAR missingness by definition depends on the values of observed variables, we can generate MAR missingness by assigning weights of zero to all variables that will be amputed. In contrast, if we choose to give a non-zero weight to one or more of the variables that will be amputed, the generated missingness mechanism becomes MNAR. In the last step of the multivariate amputation procedure, the assigned probabilities are used to generate missing values. Because we use a probabilistic model, some cases will remain complete, while other will receive missing values (Figure 2).

3.2. Benefits of multivariate amputation

With the multivariate amputation approach, we have a procedure to accurately mimic realistic missingness problems. It is straightforward to define various missing data patterns, the relative occurrence of these patterns (by defining the size of the subsets) and to manipulate the total missingness percentage for different variables. By offering the possibility to specify weight values, the multivariate amputation procedure also enables the fine-tuning of missing data mechanisms.

With the availability of R-function `ampute`, we deliver a means to overcome potential problems that may occur with the current way of practice. `ampute`'s default settings will generate 50% MAR-RIGHT missingness. Further, all aspects of the multivariate amputation methodology as discussed above can be regulated for different types of data (categorical data will be made numeric). Moreover, a smart use of the weight matrix enables the generation of both MAR and MNAR mechanisms within the same data set. These and other features of `ampute` ensure the generation of complex – but realistic – missing data problems. For a detailed explanation about `ampute`'s arguments, we refer to [8].

The next two sections will answer two research questions. First, we will evaluate the performance of `ampute`. By means of a simulation study, we will study whether `ampute` generates legitimate missingness mechanisms. Next, Section 5 will compare the performance of stepwise univariate amputation and multivariate amputation.

4. Evaluation of `ampute`

4.1. Simulation conditions

We perform a model-based simulation by repeatedly drawing $N = 1000$ cases from a multivariate normal distribution with `mvrnorm` in the package `MASS` [23] in R [24]. We use mean structure

$$\mu = \begin{matrix} Y_1 \\ Y_2 \\ X_1 \end{matrix} \begin{pmatrix} 5 \\ 5 \\ 10 \end{pmatrix},$$

and covariance structure

$$\Sigma = \begin{matrix} Y_1 \\ Y_2 \\ X_1 \end{matrix} \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}.$$

Table 1. Five simulation conditions for the evaluation of `ampute`.

condition	pattern			weights		
	Y_1	Y_2	X_1	Y_1	Y_2	X_1
mcar	0	0	1	0	0	0
mar	0	0	1	0	0	1
weak mar	0	0	1	0	0	0
	0	0	1	0	0	1
mnar	0	0	1	1	0	0
weak mnar	0	0	1	1	0	5

Notes: The pattern column displays whether variables are amputed (0) or remain complete (1). The weights column displays whether the variables influence the amputations.

We vary the correlation $\rho \in \{0.1, 0.2, \ldots, 0.9\}$. The number of Monte Carlo simulations is 1000.

We will examine the five different missing data scenarios from Table 1. For each scenario, we will generate joint missingness in variables Y_1 and Y_2 for 20% of the cases. Note that this means that there are no cases with missing values on either Y_1 or Y_2. Rather, a case is complete or has missing values on the two Y variables together. We leave X_1 an observed covariate (Table 1).

We categorize the five missing data scenarios as a MCAR, MAR, weak MAR, MNAR and weak MNAR kind of missingness. With MCAR missingness, the probability of being missing is 0.2 for each case. For the MAR and MNAR scenarios, the probabilities of being missing are based on the values of X_1 and Y_1, respectively, indicated by 1 in weights column in Table 1. In case of weak MAR missingness, one part of the cases will be MCAR while the other part is based on the values of X_1. For the weak MNAR simulation condition, the weighted sum scores are determined by variables Y_1 and X_1 with ratio 1:5. For all scenarios, we generate right-tailed (RIGHT) missingness as displayed in Figure 3.

For reasons of brevity, we focus our evaluation on the mean value of Y_1 ($\mathbb{E}_{Y_1}$). We calculate $\mathbb{E}_{Y_1}$ for every incomplete data set with complete case analysis (CCA) and after multiple imputation (MI) with predictive mean matching as the imputation technique (PMM; `mice.impute.pmm`). This last step, of imputing the variables again, is crucial to see whether exisiting missing data methodologies can recover the missing data we generate. For every replication, we generate $m = 5$ multiply imputed data sets and combine the m completed data means into a single inference following Rubin's rules [1,p.76,77].

We study the coverage rate of the 95% confidence intervals for the mean (i.e. in how many of the 1000 replications does the 95% confidence interval of $\mathbb{E}_{Y_1}$ contain the true population value of $\mathbb{E}_{Y_1} = 5$), the average width of that confidence interval and the average bias with respect to the true population value of $\mathbb{E}_{Y_1} = 5$.

4.2. Hypotheses

When MCAR missingness is generated, the observed distribution of Y_1 will be thinned. Consequently, CCA will yield unbiased estimates and the coverage rates will be around 0.95. When the MAR and MNAR mechanisms are generated as intended, CCA estimates

will be negatively biased and the coverage rates will be under 0.95. Because we induced right-tailed missingness, the right side of the distribution of Y_1 will be amputed. As a result, the observed distribution of Y_1 becomes skewed which leads to a negative bias in the CCA estimates of $\mathbb{E}_{Y_1}$. The same holds for the weak MAR and weak MNAR simulation conditions.

Because we know that MI yields valid inference when the observed data contains the information about the missing part, we expect that after imputation, estimates of $\mathbb{E}_{Y_1}$ are unbiased and efficient for the MCAR, MAR and weak MAR simulation conditions. In case of MNAR missingness, the information about the true data model is not available from the observed data. Thus, a MNAR missingness mechanism will still give biased estimates and coverage rates will be under 0.95. For the same reason, weak MNAR missingness will yield invalid statistical inferences, but because the observed information of the weak MAR condition partly contains the information about the missing values, the deviation from the true population value may be smaller than with MNAR missingness.

4.3. Results

Figure 4 shows the coverage rate, average bias and average confidence interval width of the estimates of $\mathbb{E}_{Y_1}$ for CCA and MI. We will discuss the results by examining Figure 4(a)/4(c), 4(d)/4(d) and 4(e)/4(f), respectively. Note that a numerical representation of this information is available from Table A1 in Appendix 3.

As expected, MCAR missingness results in unbiased estimates (Figure 4(a)) and coverage rates around 0.95 (Figure 4(c)). In line with our hypotheses, all MAR and MNAR missingness mechanisms produce negatively biased estimates (Figure 4(a)) and cause a drop in the coverage rates (Figure 4(c)). It is interesting to see that for missing data scenarios where X_1 governs the amputation procedure (i.e. with MAR, weak MAR and weak MNAR missingness), the size of the bias has a linear relation with data correlation ρ. This is not the case for MCAR and MNAR missingness, as these mechanisms do not depend on the values of X_1.

Figure 4(b, d) show that when the observed data holds the information about the missingness (i.e. with MCAR, weak MAR and MAR missingness), MI results in unbiased estimates and coverage rates around 0.95. In contrast, when the missingness is based on the missing values themselves (i.e. with weak MNAR and MNAR missingness), the estimates of $\mathbb{E}_{Y_1}$ are biased and coverage rates are under 0.95. These results are in line with what we know from theory. Furthermore, Figure 4(d) shows that when the correlations between X_1, Y_1 and Y_2 are high, the coverage rates of these mechanisms increase towards 0.95. It is promising to see that MI can utilize the observed information to solve (part of) the missing data problem.

Because the percentage of cases with missing values is similar in all simulation conditions (20%), the width of the 95% confidence interval should be seen as a measure of the variance of the observed distribution of Y_1. For MCAR missingness, this variance is larger than for MNAR missingness (see Figure 4(e)). Besides, we see that compared with CCA, the confidence interval widths after MI start with a higher value. This effect is expected because MI produces between imputation variance on top of the data variance [1]. Nonetheless, when data correlation ρ increases, MI decreases the confidence interval
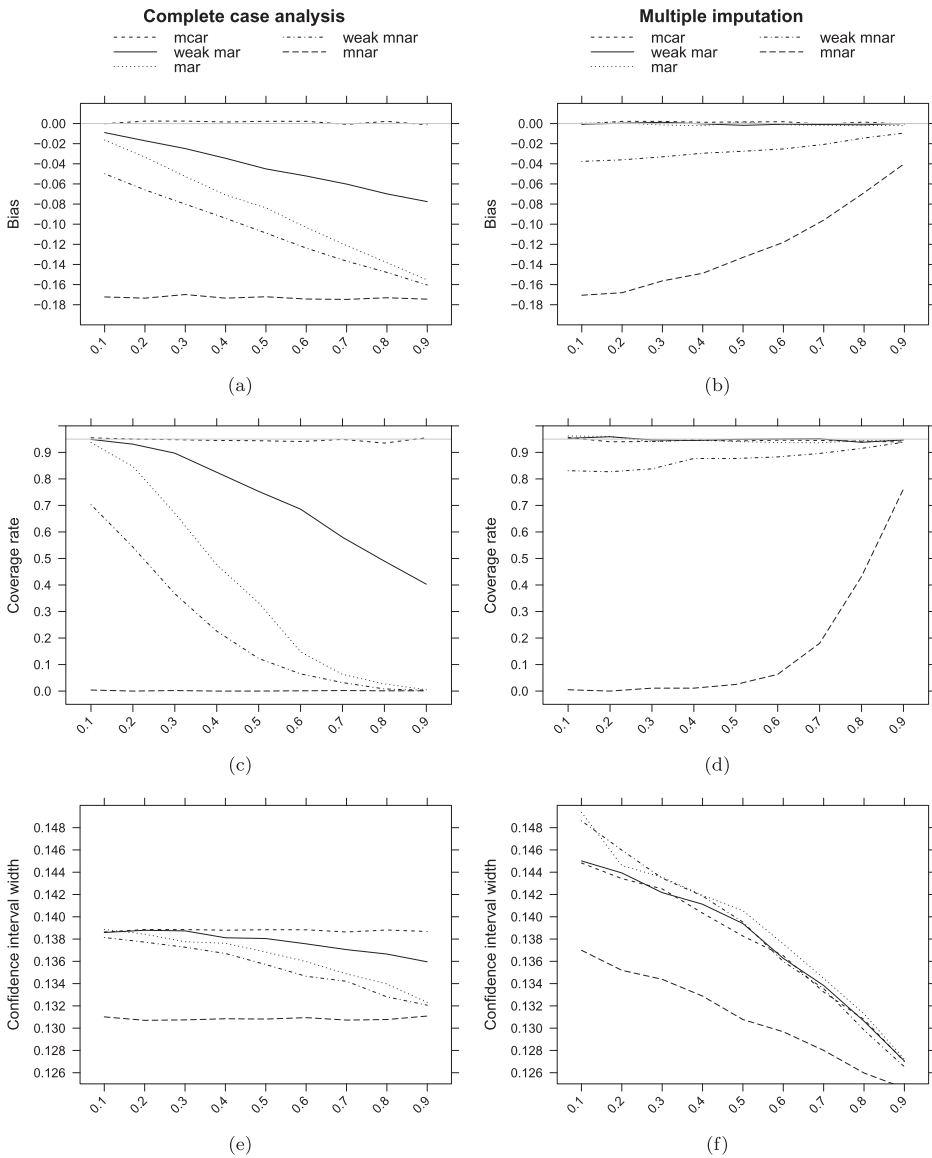
Figure 4. Evaluation of R-function `ampute`. Coverage rate, bias and confidence interval width of $\mathbb{E}_{Y_1}$ estimates are shown for CCA (a, c, e) and multiple imputation by predictive mean matching (b, d, f). The *x*-axis shows the data correlation $\rho \in \{0.1, 0.2, \dots, 0.9\}$.

widths (Figure 4(f)). In fact, the competence of MI to simultaneously reduce the bias and the variance of statistical estimates makes this missing data method the fantastic method we know it is.

In sum, we find that `ampute` produces the five evaluated missing data scenarios exactly according our hypotheses. Without any doubt we can conclude that `ampute` produces legitimate MCAR, MAR and MNAR missingness and any desired variation thereof.

5. Comparison of amputation procedures

5.1. Simulation conditions

To compare the performance of stepwise univariate amputation (SUA) and multivariate amputation (MA), we perform a model-based simulation study. Similar to section 4, we repeatedly draw $N = 1000$ cases from a multivariate normal distribution with the same mean and covariance structures. We vary the data correlation $\rho \in \{0.2, 0.5, 0.8\}$ and again, replicate the simulations 1000 times.

It is our intention to generate missing values for 50% of the cases. The incomplete cases will have missing values on both variables Y_1 and Y_2. Note that we will not permit cases to have missing values on either Y_1 or Y_2. Rather, 50% of the cases has joint missingness on variables Y_1 and Y_2 while the other 50% remains complete. Besides, we leave X_1 to be an always observed covariate.

We compare three simulation conditions: two SUA approaches and MA. For the first SUA approach, we repeat the *standard* right-tailed (RIGHT) type of logistic distribution function for variables Y_1 and Y_2 (cf. Figure 1). For the second simulation condition, we take the same approach but with a *shifted* probability distribution (cf. Figure 1). We determined b to be 0.922. This can be calculated with Equation (A6) in Appendix 2. Third, we jointly generate missing values for variables Y_1 and Y_2 with the implementation of MA in R-function `ampute`. The specifications of the arguments of the function are attached in Appendix 4.

We perform simulations for both MAR and MNAR missingness. For the two SUA conditions, we distinguish between MAR and MNAR missingness by choosing either X_1 (MAR) or Y_1 (MNAR) as the x-axis variable in Figure 1. With `ampute`, we made the according specifications for the weight values (see Appendix 4). In addition, we repeat the process for a data set with four variables (Y_1, Y_2, Y_3 and X_1) where we intent to generate 50% joint missingness on variables Y_1, Y_2 and Y_3.

Equivalent to the simulation study of Section 4, we evaluate the mean value of Y_1 ($\mathbb{E}_{Y_1}$) with CCA and after multiple imputation (MI) with predictive mean matching as the imputation technique (PMM; `mice.impute.pmm`). We choose for PMM because this imputation technique samples from the observed data and thus, may be more sensitive to amputing part of the true data distribution than other imputation methods [25]. By comparing the actual obtained missingness percentage, the average bias, the average width of the 95% confidence interval and the coverage rate of these intervals, we can evaluate and compare the performance of the three amputation approaches.

5.2. Hypotheses

Although we intent to generate missing values for 50% of the cases, we expect to see that standard SUA gives a missingness percentage smaller than 50. By definition, SUA generates the missing values in Y_1 separately from the missingness in Y_2. Therefore, it is difficult to regulate the percentage of joint missingness. In contrast, as we manually shifted the distribution function for the second simulation condition, we anticipate a joint missingness percentage of 50 for this simulation condition. With MA, regulation of the missingness percentage is no issue and we expect to see a neat 50% there as well.

When it is the intention to generate 50% right-tailed missingness in a normally distributed variable, the observed distribution of this variable is supposed to get a steady, fixed degree of skewness. With CCA, we can observe this level of skewness by evaluating the bias in the estimates of $\mathbb{E}_{Y_1}$. Because it is logical that the actual obtained missingness proportion is of influence on the size of this bias as well, we expect that the estimates of standard SUA will have smaller bias values than the other two simulation conditions. These small biases induce the generated missing data problem to be less severe. Furthermore, as shifting the distribution function will engender a ceiling effect (see Section 2.2), we expect that shifted SUA will generate too much bias in the estimates of $\mathbb{E}_{Y_1}$. Hence, shifted SUA will generate a too severe missing data problem. In contrast, MA works such that not only the intended 50% of the cases will be made incomplete but the procedure will also produce the legitimate degree of skewness.

We examine the consequences of the three amputation approaches for the evaluation of a missing data methodology by investigating the estimates of $\mathbb{E}_{Y_1}$ after imputation. The performance of MI depends on (1) whether the observed information contains the information about the missing part and (2) the severity of the generated missing data problem. As a result of the first, we expect that for MAR missingness MI returns unbiased and efficient estimates. With a MNAR mechanism, the information about the missingness lies outside the observed data and therefore, we expect to see biased estimates. Fortunately, MI uses the observed data to reduce the bias as much as possible. Furthermore, because standard SUA produces too small biases, shifted SUA produces too large biases and MA returns the intended biases, an evaluation of the performance of MI will be too optimal for standard SUA, too pessimistic for shifted SUA and legitimate for MA.

5.3. Results

First of all, the obtained missingness percentages are precisely as expected. Furthermore, 2 shows that although the missingness percentages of shifted SUA and MA are similar, there is a difference in the generated bias. For instance, when the correlations between X_1, Y_1 and Y_2 are 0.2, the average bias with shifted SUA is -0.092 while it is -0.081 with MA. These findings show that although shifting the logistic distribution function results in the desired missingness percentage, the severity of the missing data problem has changed as well. In addition, the higher the correlation in the simulated data, the more apparent this effect becomes. For instance, compare the difference between -0.233 and -0.207 for $\rho = 0.5$ with the difference between -0.372 and -0.331 for $\rho = 0.8$. All these findings are in line with our hypotheses. Furthermore, 2 displays a relatively small confidence interval width for standard SUA (0.147 for $\rho = 0.2$). The other two approaches have larger and comparable confidence interval widths because their missingness percentages are larger. However, due to their difference in bias, the coverage rates of shifted SUA and MA have become different. An example of this situation occurs for $\rho = 0.2$. There, shifted SUA results in a coverage rate of 0.455 while it is 0.552 for MA.

What we know from theory - MI yields valid inference for MAR missing data - can clearly be seen in Table 2. MI decreases the biases to zero. Interestingly, we notice that although the missingness proportions for shifted SUA and MA are similar, the 95% confidence intervals do not always have an equal width. Probably, our finding that shifted SUA increases the severity of the missing data problem affects the performance of MI. Based on

Table 2. Generation of MAR missingness on two variables with standard and shifted stepwise univariate amputation (SUA) and multivariate amputation (MA).

cor	condition	%mis		cca			mi		
		int	obt	bias	ciw	cov	bias	ciw	cov
	standard SUA	50	29	−0.061	0.147	0.630	−0.004	0.166	0.947
0.2	shifted SUA	50	50	−0.092	0.175	0.455	0.001	0.223	0.938
	MA with `ampute`	50	50	−0.081	0.175	0.552	−0.001	0.219	0.936
	standard SUA	50	29	−0.146	0.144	0.028	−0.002	0.156	0.940
0.5	shifted SUA	50	50	−0.233	0.172	0.000	−0.007	0.204	0.917
	MA with `ampute`	50	50	−0.207	0.172	0.002	−0.005	0.193	0.936
	standard SUA	50	29	−0.235	0.139	0.000	−0.007	0.137	0.936
0.8	shifted SUA	50	50	−0.372	0.164	0.000	−0.013	0.157	0.913
	MA with `ampute`	50	50	−0.331	0.166	0.000	−0.010	0.157	0.912

Notes: Estimates of $\mathbb{E}_{Y_1}$ are evaluated with complete case analysis (cca) and multiple imputation (mi). Comparisons are done for the intended (int) and obtained (obt) missingness percentage (%mis), the bias, confidence interval width (ciw) and coverage rate (cov) for correlations $\rho \in \{0.2, 0.5, 0.8\}$ (cor).

these results, a direct consequence for the coverage rates cannot be justified. However, we do see that it is possible that the performance of MI is different for shifted SUA than for MA. For instance, the coverage rate after MI is 0.917 for shifted SUA and 0.936 for MA when $\rho = 0.5$.

Table 3 displays the results for MNAR missingness. Our findings are in line with our hypotheses and comparable with the trends discussed above. Further, the results presented in Table 4 confirm our hypothesis that MA enables the generation of reliable missing data problems. That is to say, when the required adaptation of the logistic distribution function is not executed, the missingness percentages of SUA quickly drop below the intended percentage. Table 4 shows that standard SUA generates missing values for merely 19% of the cases and the shift of $b = 0.922$ induces 39% joint missingness. In contrast, MA exactly generates the intended 50% missingness. Moreover, we see that with MA the average bias with CCA is similar to the bias we saw earlier. Namely, when $\rho = 0.8$, the bias is −0.331 in Tables 2 and 4. This finding shows that MA allows for the adjustment of one feature without disturbing other missing data characteristics. In sum, MA performs the amputation procedure according the intentions.

Table 3. Generation of MNAR missingness on two variables with standard and shifted stepwise univariate amputation (SUA) and multivariate amputation (MA).

cor	condition	%mis		cca			mi		
		int	obt	bias	ciw	cov	bias	ciw	cov
	standard SUA	50	29	−0.291	0.134	0.000	−0.282	0.142	0.000
0.2	shifted SUA	50	50	−0.465	0.158	0.000	−0.451	0.178	0.000
	MA with `ampute`	50	50	−0.415	0.160	0.000	−0.401	0.182	0.000
	standard SUA	50	29	−0.289	0.134	0.000	−0.227	0.137	0.000
0.5	shifted SUA	50	50	−0.465	0.158	0.000	−0.369	0.169	0.000
	MA with `ampute`	50	50	−0.412	0.160	0.000	−0.325	0.169	0.000
	standard SUA	50	29	−0.292	0.134	0.000	−0.121	0.129	0.064
0.8	shifted SUA	50	50	−0.466	0.158	0.000	−0.199	0.148	0.000
	MA with `ampute`	50	50	−0.415	0.160	0.000	−0.175	0.148	0.009

Notes: Estimates of $\mathbb{E}_{Y_1}$ are evaluated with complete case analysis (cca) and multiple imputation (mi). Comparisons are done for the intended (int) and obtained (obt) missingness percentage (%mis), the bias, confidence interval width (ciw) and coverage rate (cov) for correlations $\rho \in \{0.2, 0.5, 0.8\}$ (cor).

Table 4. Generation of MAR missingness on three variables with standard and shifted stepwise univariate amputation (SUA) and multivariate amputation (MA).

cor	condition	%mis		cca			mi		
		int	act	bias	ciw	cov	bias	ciw	cov
	standard SUA	50	19	−0.042	0.138	0.776	0.001	0.147	0.950
0.2	shifted SUA	50	39	−0.078	0.158	0.507	−0.003	0.195	0.947
	MA with `ampute`	50	50	−0.084	0.175	0.521	−0.004	0.221	0.937
	standard SUA	50	19	−0.108	0.135	0.128	−0.002	0.140	0.942
0.5	shifted SUA	50	39	−0.193	0.154	0.003	−0.005	0.173	0.940
	MA with `ampute`	50	50	−0.209	0.172	0.003	−0.008	0.196	0.939
	standard SUA	50	19	−0.172	0.131	0.001	−0.004	0.131	0.951
0.8	shifted SUA	50	38	−0.310	0.148	0.000	−0.015	0.145	0.901
	MA with `ampute`	50	50	−0.331	0.166	0.000	−0.017	0.155	0.904

Notes: Estimates of $\mathbb{E}_{Y_1}$ are evaluated with complete case analysis (cca) and multiple imputation (mi). Comparisons are done for the intended (int) and obtained (obt) missingness percentage (%mis), the bias, confidence interval width (ciw) and coverage rate (cov) for correlations $\rho \in \{0.2, 0.5, 0.8\}$ (cor).

6. Conclusion and discussion

6.1. Conclusion

A valid amputation procedure is essential for an accurate evaluation of the performance of missing data methodology. We showed that the approach of stepwise univariate amputation may not be appropriate for the generation of intuitive and reliable missing data problems. Namely, stepwise univariate amputation does not allow for a precise manipulation of missing data characteristics. For instance, one of the most basic features of missing data problems – the missingness percentage – requires the solving of complex integrals. These calculations can be done for situations with two or three variables, but for larger data sets a proper generation of any realistic missing data problem becomes impossible. In addition, we showed that an increase of the missing data proportion leads to an inflated observed data skewness. Hence, the evaluation of missing data methodology may happen under the wrong simulation conditions and missing data methodology can be falsely judged to under- or overperform.

We implemented a multivariate amputation procedure into R-function `ampute`. We showed that `ampute` generates any missing data scenario precisely as desired. With multivariate amputation, missing data characteristics such as the missingness percentage and the number of incomplete variables are easy to manipulate. Furthermore, the amputations induce a missing data problem which behaves exactly as intended. In other words, when MAR missingness is intended, MAR missingness is generated. Moreover, we showed that the degree of observed data skewness is not sensitive to the number of amputed variables. This indicates that multivariate amputation allows for a separate regulation of missing data parameters. All in all, the multivariate amputation procedure is an intuitive and reliable amputation method which enables valid evaluations of missing data methodology.

6.2. Discussion

Possible applications of multivariate amputation function `ampute` can be found in the field of survey research. Many studies describe the use of so-called planned missing data

designs (e.g. [26–28]), which are survey designs where 'missing data are used strategically to improve the validity of data collection.' [29,p.426]. For instance, in order to reduce the respondent's burden, a large questionnaire is divided into multiple parts and each participant receives one or two of these parts. Such a multi-form design is almost always based on random chance (i.e. MCAR) which may give weird combinations of questions. With `ampute`, it is possible to create a logical combination of variables and to investigate the most efficient way to shorten a questionnaire.

In addition, two-method measurements designs (e.g. [29,30]) could benefit from the availability of `ampute`. With this type of planned missing data designs, all participants fill in a questionnaire but only a small part of the participants is invited for a second measurement wave. Currently, the selection of these second wave participants is based on the values of one variable. As an alternative, `ampute` could combine the information of multiple variables. Possibly, this could enhance the use of observed information and reduce the number of participants without a loss of statistical power.

With the availability of a multivariate amputation procedure, systematic evaluations of the performance of missing data methods are feasible. With R-function `ampute`, it is easy to assess in which situations which missing data methods are preferred. Besides, any usage of `ampute` is not restricted to evaluations of imputation techniques. Evaluations of weighting procedures (e.g. [31]) and likelihood-based methods (e.g. [32,33]) can be done as well. Besides, scientific studies that examine the combination of multiple source data (e.g. [34]), different types of measurement error (e.g. [35]) or the application of missing data methods in software tools (e.g. [36]) could benefit from `ampute`.

We focused our simulations on right-tailed MCAR, MAR and MNAR missingness. Additionally, it would be interesting to know how these and variants of these mechanisms behave in data structures other than the multivariate normal distribution. For instance, what happens with the accuracy of statistical estimates when variables have a nonlinear relation? And how can `ampute` be used to generate a realistic multilevel missing data problem? Especially the effect of correlations on the outcome of the amputation procedure may need extra attention. Because correlations determine a large part of the effect of missing values on statistical inferences, and because correlations are important for the performance of missing data methods, it would be valuable to gain more insights into this aspect of multivariate amputation.

Overall, we emphasize the importance of a greater understanding of the behaviour of missing data mechanisms and their impact on the validity of both the generation and the handling of missing data. After all, for an accurate evaluation strategy, an efficient and reliable amputation procedure is just the beginning.

Acknowledgments

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Rubin D. Multiple imputation for nonresponse in surveys. New York: John Wiley and Sons; 1987.

[2] Rubin D. Multiple imputation after 18+ years. J Amer Statist Assoc. 1996;91(434):473–489.

[3] Van Buuren S. Flexible imputation of missing data. Boca Raton, FL: Chapman & Hall/CRC; 2012.

[4] Vink G, Van Buuren S. Multiple imputation of squared terms. Sociol Methods Res. 2013;42(4):598–607.

[5] Fang F, Shao J. Iterated imputation estimation for generalized linear models with missing response and covariate values. Comput Statist Data Anal. 2016;103:111–123.

[6] Kombo A, Mwambi H, Molenberghs G. Multiple imputation for ordinal longitudinal data with monotone missing data patterns. J Appl Stat. 2016;44(2):270–287.

[7] Seaman S, Bartlett J, White I. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Med Res Methodol. 2012;12:1–13.

[8] Schouten R, Lugtig P, Brand J, et al. Generate missing values with ampute; 2018; [cited 2018 June 28]. Available from: https://rianneschouten.github.io/mice_ampute/vignette/ampute.html

[9] Van Buuren S, Groothuis-Oudshoorn C. Mice: multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1–67.

[10] Sullivan T, White I, Salter A, et al. Should multiple imputation be the method of choice for handlig missing data in randomized trials? Stat Methods Med Res. 2016;1–17. DOI:10.1177/0962280216683570

[11] Shara N, Yassin S, Valaitis E, et al. Randomly and non-randomly missing renal function data in the strong heart study: a comparison of imputation methods. PloS one. 2015;10(9):e0138923

[12] Kontopantelis E, White I, Sperrin M, et al. Outcome-sensitive multiple imputation: a simulation study. BMC Med Res Methodol. 2017;17(2):1–13.

[13] Abrahantes J, Sotto C, Molenberghs G, et al. A comparison of various software tools for dealing with missing data via imputation. J Stat Comput Simul. 2011;18(11):1653–1675.

[14] Kalaycioglu O, Copas A, King M, et al. A comparison of multiple imputation methods for handling missing data in repeated measurements observational studies. Stat Soc Ser A. 2016;(1793):683–706.

[15] Pastor J. Methods for the analysis of explanatory linear regression models with missing data not at random. Qual Quant. 2003;37:363–376.

[16] Rubin D. Inference and missing data. Biometrika. 1976;63(3):581–592.

[17] Allison P. Multiple imputation for missing data. Sociol Methods Res. 2000;28(3):301–309.

[18] Yuan K, Yang-Wallentin F, Bentler P. ML versus MI for missing data with violation of distribution conditions. Sociol Methods Res. 2012;4(4):598–629.

[19] Vink G Towards a standardized evaluation of multiple imputation routines; 2016; [cited 2018 June 28]. Available from: https://www.gerkovink.com/publications/

[20] Brand J. Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. University Medical Center Rotterdam; 1999

[21] Van Buuren S, Brand J, Groothuis-Oudshoorn C, et al. Fully conditional specification in multivariate imputation. J Stat Comput Simul. 2006;76(12):1049–1064.

[22] Brand J, Van Buuren S, Groothuis-Oudshoorn C, et al. A toolkit in SAS for the evaluation of multiple imputation methods. Statist Neerlandica. 2003;57(1):36–45.

[23] Ripley B, Venables B, Bates D. R-package MASS; 2018; [cited 2018 June 28]. Available from: https://cran.r-project.org/web/packages/MASS/index.html

[24] R Core Team. A language and environment for statistical computing and graphics. R Foundation for Statistical Computing and Graphics; 2018; [cited 2018 June 28]. Available from: http://www.R-project.org/

[25] Vink G, Frank L, Pannekoek J, et al. Predictive mean matching imputation of semicontinuous variables. Statist Neerlandica. 2014;68(1):61–90.

[26] Enders C. Applied missing data analysis. New York: The Guilford Press; 2010.

2926 R. M. SCHOUTEN ET AL.

[27] Wu W, Jia F, Rhemtulla M, et al. Search for efficient complete and planned missing data designs for analysis of change. Behav Res Methods. 2016;48(3):1047–1061.
[28] Bunting B, Adamson G, Mulhall P. A Monte Carlo examination of an MTMM model with planned incomplete data structures. Struct Equ Modeling. 2009;9(3):369–389.
[29] Rhemtulla M, Little T. Planned missing data designs for research in cognitive development. J Cogn Dev. 2012;13(4):425–438.
[30] Graham J, Taylor B, Olchowski A, et al. Planned missing data designs in psychological research. Psychol Methods. 2006;11(4):323–343.
[31] Boeschoten L, Vink G, Hox J. How to obtain valid inference under unit non-response? J Off Stat. 2017;33(4):963–978.
[32] Little R, Rubin D. Statistical analysis with missing data. New York: John Wiley & Sons; 2002.
[33] Molenberghs G, Kenward M. Missing data in clinical studies. Chicester: John Wiley & Sons; 2007.
[34] Boeschoten L, Oberski D, De Waal A. Estimating classification error under edit restrictions in combined survey-register data using Multiple Imputation Latent Class Modeling (MILC). J Off Stat. 2017;33(4):921–962.
[35] Boeschoten L, Oberski D, De Waal A. Estimating classification error under edit restrictions in combined survey-register data. Statistics Netherlands: CBS Discussion Paper;2016.
[36] Horton N, Kleinman K. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. J Amer Statist Assoc. 2007;61(1):79–90.
[37] Miller I, Miller M. Mathematical expectation. Chapter 4. In: John E. Freund's, editor. Mathematical statistics with applications. New Jersey: Pearson Prentice Hall; 2004. p. 129–163.

Appendices

Appendix 1

Consider two discrete random variables Z_1 and Z_2 such that z_{1i} is the probability of case i to be missing on variable Y_1 and that z_{2i} is the probability of case i to be missing on variable Y_2. We give each case the same probability p and we use this probability for both variables. Hence,

$$z_{1i} = z_{2i} = p \quad \forall i \in \{1, 2, \ldots, n\}.$$

To know the expected proportion of missing values in Y_1, we calculate the expectation of Z_1. By definition [37],

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i \cdot f(x),$$

and therefore

$$\mathbb{E}[Z_1] = \sum_{i=1}^{n} p \cdot \frac{1}{n}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} p$$

$$= \frac{n \cdot p}{n} = p.$$

We desire to choose p such that the total percentage of cases with missing values on both variables Y_1 and Y_2 equals a certain value m. In other words, we desire to let $\mathbb{E}[Z_1 \cdot Z_2] = m$. By definition,

because Z_1 and Z_2 are dependent,

$$\mathbb{E}[Z_1 \cdot Z_2] = \mathbb{C}ov(Z_1, Z_2) + \mathbb{E}[Z_1] \cdot \mathbb{E}[Z_2]. \tag{A1}$$

Since $\mathbb{E}[Z_1] = \mathbb{E}[Z_2]$,

$$\mathbb{C}ov(Z_1, Z_2) = \mathbb{V}ar(Z_1) = \mathbb{E}[(Z_1 - \mathbb{E}[Z_1])^2] = \mathbb{E}[Z_1{}^2] - (\mathbb{E}[Z_1])^2, \tag{A2}$$

which gives

$$\mathbb{E}[Z_1 \cdot Z_2] = \mathbb{E}[Z_1{}^2] - (\mathbb{E}[Z_1])^2 + \mathbb{E}[Z_1] \cdot \mathbb{E}[Z_1] = \mathbb{E}[Z_1{}^2]. \tag{A3}$$

Furthermore, by definition [37],

$$\mathbb{E}[X^2] = \sum_{i=1}^{n} x_i^2 \cdot f(x),$$

and therefore,

$$\begin{aligned}\mathbb{E}[Z_1{}^2] = m &= \sum_{i=1}^{n} p^2 \cdot \frac{1}{n} \\ &= \frac{1}{n} \cdot n \cdot p^2 \\ &= p^2.\end{aligned}$$

Hence,

$$p = \sqrt{m}. \tag{A4}$$

Equation (A4) can be used to calculate the required random fixed probability. For instance, when an overlapping missingness percentage of 50% is desired, probability p should be $\sqrt{0.5} = 0.707$.

Note that, as a consequence, the proportion of totally complete cases equals

$$\begin{aligned}\mathbb{E}[(1 - Z_1)^2] &= \sum_{i=1}^{n} (1 - \sqrt{0.5})^2 \cdot \frac{1}{n} \\ &= \frac{1}{n} \cdot \sum_{i=1}^{n} 0.293^2 \\ &= \frac{n \cdot 0.086}{n} = 0.086.\end{aligned}$$

Appendix 2

Let X be a continuous random variable with as its probability density function,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-1/2((x-\mu)/\sigma)^2}.$$

Furthermore, let Z_1 and Z_2 be two continuous random variables such that z_{1i} is the probability of case i to be missing on a variable Y_1 and z_{1i} is the probability of case i to be missing on a variable Y_2. We relate the values of Z_1 and Z_2 to those of X by the logistic distribution function. Hence,

$$Z_1 = Z_2 = g(X) = \frac{1}{1 + e^{-\alpha X - b}}.$$

By definition [37], the expected value of $g(X)$ is given by

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x) \, dx$$

and therefore,

$$\mathbb{E}[Z_1] = \int_{-\infty}^{\infty} \frac{1}{1 + e^{-\alpha x - b}} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-1/2((x-\mu)/\sigma)^2} \, dx.$$

When we choose $\sigma = 1$ and $\mu = 0$ such that X has a standard normal distribution, and choose $\alpha = 1$ and $b = 0$ such that $g(X)$ is the standard logistic function,

$$\mathbb{E}[Z_1] = \int_{-\infty}^{\infty} \frac{1}{1 + e^x} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-1/2x^2} \, dx$$

$$= 0.5.$$

In other words, the expected percentage of cases with a missing value on Y_1 is 50%. The same holds for variable Y_2. However, the percentage of cases with missing values on both Y_1 and Y_2 will be different. Namely, in accordance with Equations (A1)–(A3), and since

$$\mathbb{E}[g(X)^2] = \int_{-\infty}^{\infty} g(x)^2 \cdot f(x) \, dx,$$

the proportion of cases with missing values on both Y_1 and Y_2 is

$$\mathbb{E}[Z_1 \cdot Z_2] = \mathbb{E}[Z_1{}^2] = \int_{-\infty}^{\infty} \left(\frac{1}{1 + e^x} \right)^2 \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \, dx$$

$$= 0.29. \tag{A5}$$

For $\alpha \neq 1$ and $b \neq 0$, the logistic distribution will look different. Hence, if multiplied with a normally distributed variable X, the obtained missingness percentages will become different as well. The joint missingness percentage for two variables can be determined with:

$$\mathbb{E}[Z_1 \cdot Z_2] = \mathbb{E}[Z_1{}^2] = \int_{-\infty}^{\infty} \left(\frac{1}{1 + e^{-\alpha x - b}} \right)^2 \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-1/2x^2} \, dx. \tag{A6}$$

For instance, to obtain an overlapping missingness percentage of 50 with $\alpha = 1$, b should be shifted with 0.922, as can be calculated with

$$\mathbb{E}[Z_1 \cdot Z_2] = \mathbb{E}[Z_1{}^2] = \int_{-\infty}^{\infty} \left(\frac{1}{1 + e^{-x - 0.922}} \right)^2 \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-1/2x^2} \, dx$$

$$= 0.50.$$

In addition, an extension of Equation (A6) to a certain number of variables r is possible by solving

$$\mathbb{E}[Z_1 \cdot Z_2 \cdot \ldots \cdot Z_r] = \mathbb{E}[Z_1{}^r] = \int_{-\infty}^{\infty} \left(\frac{1}{1 + e^{-\alpha x - b}} \right)^r \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-1/2x^2} \, dx.$$

Appendix 3

Table A1. Evaluation of R-function `ampute`.

mechanism	cor	cca			mi		
		bias	ciw	cov	bias	ciw	cov
	0.1	−0.000	0.139	0.955	−0.001	0.145	0.956
	0.2	0.002	0.139	0.950	0.002	0.143	0.940
	0.3	0.002	0.139	0.948	0.002	0.143	0.941
	0.4	0.002	0.139	0.945	0.001	0.140	0.947
mcar	0.5	0.002	0.139	0.944	0.002	0.138	0.943
	0.6	0.002	0.139	0.941	0.002	0.137	0.946
	0.7	−0.001	0.139	0.949	−0.001	0.133	0.945
	0.8	0.002	0.139	0.935	0.001	0.131	0.942
	0.9	−0.001	0.139	0.955	−0.001	0.127	0.939
	0.1	−0.016	0.139	0.937	0.000	0.149	0.962
	0.2	−0.033	0.138	0.847	0.001	0.145	0.960
	0.3	−0.053	0.138	0.671	−0.001	0.144	0.940
	0.4	−0.071	0.138	0.476	−0.002	0.142	0.946
mar	0.5	−0.084	0.137	0.333	0.001	0.141	0.941
	0.6	−0.103	0.136	0.148	−0.001	0.138	0.937
	0.7	−0.121	0.135	0.063	−0.002	0.134	0.936
	0.8	−0.138	0.134	0.027	−0.002	0.131	0.944
	0.9	−0.155	0.132	0.005	−0.002	0.127	0.946
	0.1	−0.009	0.139	0.949	−0.001	0.145	0.953
	0.2	−0.017	0.139	0.931	0.000	0.144	0.959
	0.3	−0.025	0.139	0.897	0.001	0.142	0.947
	0.4	−0.035	0.138	0.825	−0.000	0.141	0.945
weak mar	0.5	−0.045	0.138	0.753	−0.002	0.139	0.949
	0.6	−0.052	0.138	0.686	−0.001	0.136	0.950
	0.7	−0.060	0.137	0.580	−0.001	0.134	0.951
	0.8	−0.070	0.137	0.490	−0.001	0.131	0.938
	0.9	−0.078	0.136	0.403	0.000	0.127	0.947
	0.1	−0.172	0.131	0.004	−0.171	0.137	0.005
	0.2	−0.173	0.131	0.000	−0.168	0.135	0.000
	0.3	−0.170	0.131	0.002	−0.156	0.134	0.011
	0.4	−0.173	0.131	0.000	−0.149	0.133	0.011
mnar	0.5	−0.172	0.131	0.000	−0.133	0.131	0.025
	0.6	−0.174	0.131	0.001	−0.118	0.130	0.063
	0.7	−0.175	0.131	0.002	−0.096	0.128	0.180
	0.8	−0.173	0.131	0.001	−0.069	0.126	0.432
	0.9	−0.174	0.131	0.001	−0.040	0.125	0.762
	0.1	−0.050	0.138	0.703	−0.038	0.149	0.831
	0.2	−0.066	0.138	0.543	−0.036	0.146	0.827
	0.3	−0.080	0.137	0.367	−0.033	0.143	0.838
	0.4	−0.094	0.137	0.226	−0.029	0.142	0.877
weak mnar	0.5	−0.109	0.136	0.123	−0.028	0.140	0.877
	0.6	−0.124	0.135	0.065	−0.025	0.136	0.883
	0.7	−0.137	0.134	0.032	−0.021	0.134	0.896
	0.8	−0.148	0.133	0.008	−0.014	0.130	0.915
	0.9	−0.160	0.132	0.003	−0.009	0.127	0.938

Notes: Bias, confidence interval width (ciw) and coverage rate (cov) of $\mathbb{E}_{Y_1}$ estimates are shown for complete case analysis (cca) and multiple imputation by predictive mean matching (mi). Data correlations are $\rho \in \{0.1, 0.2, \ldots, 0.9\}$ (cor).

Appendix 4

```
ampute(data = sample, prop = 0.5, mech = 'MCAR', patterns =
c(0, 0, 1))$amp
```

```
ampute(data = sample, prop = 0.5, mech = 'MAR', type =
'RIGHT', patterns = c(0, 0, 1), weights = c(0, 0, 1))$amp
```

```
ampute(data = sample, prop = 0.5, mech = 'MNAR', type =
'RIGHT', patterns = c(0, 0, 1), weights = c(1, 0, 0))$amp
```