Original Article

Multiple Imputation by Predictive Mean Matching When Sample Size Is Small

Kristian Kleinke

Department of Psychology, Bielefeld University, Germany

Abstract: Predictive mean matching (PMM) is a state-of-the-art hot deck multiple imputation (MI) procedure. The quality of its results depends, inter alia, on the availability of suitable donor cases. Applying PMM in small sample scenarios often found in psychological or medical research could be problematic, as there might not be many (or any) suitable donor cases in the data set. So far, there has not been any systematic research that examined the performance of PMM, when sample size is small. The present study evaluated PMM in various multiple regression scenarios, where sample size, missing data percentages, the size of the regression coefficients, and PMM's donor selection strategy were systematically varied. Results show that PMM could be used in most scenarios, however results depended on the donor selection strategy: overall, PMM using either automatic distance-aided selection of donors (Gaffert, Meinfelder, & Bosch, 2016) or using the nearest neighbor produced the best results.

Keywords: missing data, multiple imputation, predictive mean matching, small samples

Introduction and Overview

Since Rubin (1987) laid out the theoretical foundation of multiple imputation (MI) and Schafer (1997a, 1997b) published software to impute incomplete data, the approach has enjoyed ever-increasing popularity and is nowadays one of the standard methods to handle missing data (Schafer & Graham, 2002): MI routines are now implemented in all major statistical packages, supporting a wide range of missing data scenarios and models, including both fully parametric methods (e.g., Schafer, 1997a, 1997b) and also more robust procedures, like predictive mean matching (PMM), which is the default imputation technique for continuous data in the MI software `mice` in R (van Buuren & Groothuis-Oudshoorn, 2011) and also in some other packages. An overview of available MI procedures and packages is given in Horton and Kleinman (2007) and at www.multiple-imputation.com.

While MI had originally been developed for handling missing data in large public-use data files, for example, from surveys and censuses (Rubin, 1987), the practical use of MI has shifted over the years: today, many practitioners also impute much smaller data sets, including data from psychological or medical research, where sample sizes are often quite small. Although, small sample size adjustments have been made to the MI framework (Barnard & Rubin, 1999) and implemented, for example, in SAS PROC MIANALYZE, in Stata, and also in `mice`, systematic research that evaluated MI's performance in such settings is very scarce (e.g., Graham & Schafer, 1999). The present study contributes to fill in this gap and evaluated the performance of multiple imputation in settings where sample sizes ranged from $N = 20$ to $N = 100$.

Additionally, using predictive mean matching (PMM) as default imputation technique might not be unproblematic, when sample size is small: classical PMM approaches impute an observed value, whose value predicted by a linear regression model is among a set of k values (the so-called donor pool), which are closest to the value predicted for the missing one. When sample size is small, the number of suitable donors could also be small. Setting the size of the donor pool k too large might result in the selection of inadequate donors, implausible imputations, and as a consequence biased inferences. Choosing a very small donor pool, on the other hand, might result in one single donor being chosen again and again. This could lead to an increased correlation of the m imputations, a too small between-imputation variance component, underestimated standard errors, and the benefits of creating multiple imputations over using a single imputation might ultimately be

lost (see also the discussion of the bias-variance-tradeoff in Schenker & Taylor, 1996).

The aim of the present study was (a) to evaluate if PMM is able to produce sufficiently accurate parameter estimates and standard errors, when sample size is small, (b) to explore, what size of the donor pool yields the best trade-off between unbiased parameters and adequate standard errors in various small sample size scenarios, and (c) to explore, if more flexible donor selection strategies like automatic distance-based selection of donors, proposed by Gaffert, Meinfelder, and Bosch (2016), work better in that regard.

Theoretical Background

The basic idea of MI is: (a) to fill in each missing value $m > 1$ times by different values, which are equally plausible under the specified imputation model, (b) to analyze the m completed data sets separately by standard complete data procedures (e.g., such as regression analysis) and (c) to combine the m sets of parameter estimates into a single overall set of results using Rubin's (1987) formula. The variability between the m imputations is supposed to reflect the additional uncertainty in parameter estimation due to missing data in an adequate way. If done "properly," MI usually yields both widely unbiased parameter estimates and adequate standard errors (cf. Schafer & Graham, 2002). For a definition of "properness" in that regard (see Rubin, 1987, Chap. 4). One of the standard techniques to create the m imputations is PMM.

Predictive Mean Matching

The idea of matching predicted means in the context of missing data imputation has been first mentioned in Rubin (1986, 1987), Rubin and Schenker (1986), and Little (1988). PMM can be enumerated among the hot deck imputation procedures (Andridge & Little, 2010). The basic principle of hot deck methods is to find one suitable donor value from an observed case that is in some regard "similar" to the missing case. PMM matches potential donors and donees via the closeness of predicted means. For each potential donor case, the fitted value (based on some meaningful regression model) is calculated and compared to the value predicted for the incomplete case. Classical PMM approaches draw one case from a pool of k cases, whose predicted values are closest to the one predicted for the missing case. The observed value of this donor case is then used to fill in the missing one. Further donor selection strategies will be discussed in the following sections. For an overview and discussions of different implementations of PMM in various software packages and their default settings, see, for example, Allison (2015), Morris, White, and Royston (2014), and Gaffert et al. (2016). For a detailed description of the PMM algorithm in `mice`, which I used in this paper, see van Buuren (2012), Algorithm 3.3.

Advantages and Disadvantages of Predictive Mean Matching

One of the major advantages of PMM is its robustness: in comparison to fully parametric procedures like, for example, Schafer's (1997a) NORM approach, PMM is less sensitive to model misspecifications, including nonlinear associations, heteroscedasticity, and deviations from normality (Morris et al., 2014; Schenker & Taylor, 1996; Vink, Frank, Pannekoek, & van Buuren, 2014; Yu, Burton, & Rivero-Arias, 2007). This is because the parametric linear model is only used for matching the incomplete case to a potential donor. By imputing an actual observed value, PMM is usually able to preserve the original distribution of the data quite well, even when the assumptions of the underlying linear resgression model are violated. Furthermore, PMM always imputes a "valid" value, meaning that the imputed value will fit the respective scale of measurement and will always be within the range of the observed values, which makes rounding unnecessary. Rounding following normal model multiple imputation can be problematic (cf. Horton, Lipsitz, & Parzen, 2003). The practice of imputing actual observed values, however, could also have disadvantages. Performance of PMM might be poor in situations where no or only few suitable donor cases could be found. Van Buuren (2012) summarizes the potential drawbacks of PMM as follows: PMM "cannot be used to extrapolate beyond the range of the data, or to interpolate within the range of the data if the data at the interior are sparse. Also, it may not perform well with small datasets" (p. 74).

From a theoretical point of view, it is plausible that PMM might perform poorly, when sample size is small: for a given missing data percentage, a smaller sample implies less suitable donors. In consequence, selecting inadequate observed values could introduce bias. This means that, for a constant sample size N, bias should increase with increasing missing data percentage. Likewise, for a constant missing data percentage p_{mis}, bias should increase with decreasing sample size. Furthermore, in addition to missing data percentage, results are also likely to depend on the donor selection strategy, as will be outlined in the next sections. Currently, there are no evaluation studies that

systematically tested, if or under what conditions PMM can be used, when sample size is small. For practitioners it is important to know, if the PMM technique in general fails, when sample size is small, or if only certain settings lead to biased inferences in certain scenarios.

The Size of the Donor Pool in Classical PMM and Its Effects on Statistical Inferences

The default size of the donor pool varies across the statistical packages. The MI procedure in SAS and the current version of `mice` in R use $k = 5$ as default. Some older versions of `mice` used $k = 3$, while Solas and `ice` in Stata sample from a pool of $k = 10$ donors by default (cf. Allison, 2015; Morris et al., 2014). For sufficiently large samples with $N \geq 100$, Schenker and Taylor (1996) have found only small differences in performance between using $k = 3$ and $k = 10$ donors, meaning that practitioners typically do not have to adjust the size of the donor pool to obtain reasonable results. When sample size is small, however, the size of the donor pool might have a greater effect on the accuracy of both parameter estimates and standard errors. To obtain unbiased parameter estimates, decreasing the size of the donor pool might become increasingly important, the smaller the sample is. On the other hand, as already mentioned, downsizing the donor pool might come at the cost of underestimated standard errors.

Newer Adaptive and Distance-Based Donor Selection Strategies

To solve the problem of determining which number of donors might overall work best in a given scenario, and to overcome some of the shortcomings of classical PMM listed above, Schenker and Taylor (1996) proposed an "adaptive technique that chooses the number of possible donors case-by-case based on the density of complete cases in the neighborhood of the incomplete case in question" (p. 430). Unfortunately, their approach has never been implemented in any of the major statistical packages and has hardly been used in practice.

A more recent solution, where one donor is selected with probability inversely proportional to its distance from the respective incomplete case, has been proposed by Siddique and Belin (2008). The procedure is available as a SAS macro called MIDAS (Multiple Imputation using Distance-aided Selection of donors). It uses an approximate Bayesian bootstrap to introduce between-imputation variability. For a detailed description of the implementation in SAS, see Siddique and Harel (2009). While their solution overcomes the problem of specifying the number of donors, their formula (Siddique & Belin, 2008, Equation 1) also includes a parameter k, which can best be described as a closeness parameter that adjusts the probability of selection. This parameter must be set by the user. Setting $k = 0$ means that each donor has the same selection probability (which comes down to a simple random hot deck). With $k \to \infty$, the nearest neighbor is always chosen. In their simulation study, Siddique and Belin (2008) found that a closeness parameter of $k = 3$ produced reasonable results. They, however, only examined scenarios, where $N \geq 100$. No recommendations are yet available for small sample size scenarios.

Gaffert et al. (2016) published a "touched-up" version of the MIDAS macro for R, which they call `midastouch`. One of the main differences to the solution by Siddique and Belin (2008) is that the user can, but does not necessarily have to specify the closeness parameter. In the default settings, the R function automatically sets the parameter according to the R^2 of the imputation model based on the full set of donors. Additionally, the function uses a correction for the total variance as originally suggested by Parzen, Lipsitz, and Fitzmaurice (2005). They use this correction because the approximate Bayesian bootstrap used by Siddique and Belin (2008) yields unbiased estimates only when the number of observed values $N_{obs} \to \infty$. For a more detailed discussion, and for a full list of differences between MIDAS and `midastouch`, see Gaffert et al. (2016).

While using such automatic distance-based donor selection procedures might be preferrable from a practitioner's point of view, as it overcomes the problem of specifying the size of the donor pool, little is yet known about the quality of these procedures, especially in small sample size scenarios.

Research Questions and Hypotheses

In summary, the present study had several aims: (1) to test, if overall, PMM yields accurate statistical inferences in a variety of small sample size scenarios; (2) to explore, if (and how) the size of the donor pool needs to be adjusted to produce acceptable parameter estimates and standard errors in a given scenario. Bias in parameter estimates was hypothesized to increase with decreasing sample size, increasing missing data percentage, and increasing size of the donor pool, while bias in standard error estimates was believed to increase with decreasing size of the donor pool; and (3) to test, if automatic distance-based donor selection (`midastouch`) produces overall better results than using a donor pool of constant size k.

Method

Design of the Study

Overview

The Monte Carlo study was designed as a $4 \times 5 \times 3 \times 5$ factorial experiment regarding a multiple regression scenario, in which data in the dependent variable were missing. The factors that were manipulated were sample size, the missing data percentage, the regression weights of the predictors and thus the R^2 of the regression model, and finally the way, how PMM identified a donor case. I first describe the data generation process and the factors in the study, then the quality criteria used to evaluate the performance of PMM multiple imputation in these scenarios.

Data Generation

The data generation process was similar to the one by Schenker and Taylor (1996). I used a variant of their baseline model, with, x_1 and x_2 being the predictors in the regression model and y being the dependent variable. x_1 and x_2 were independently distributed as $\mathcal{N}(5, 1)$. The model for y given x_1 and x_2 was linear and homoscedastic. y_1 were obtained by

$$y_i = 10 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \quad (1)$$

where $e_i \sim \mathcal{N}(0, 1)$.

Experimental Conditions

The first factor I manipulated was sample size, with $N \in \{20, 30, 50, 100\}$. Secondly, I varied the missing data percentages with $p_{\text{mis}} \in \{10\%, 20\%, 30\%, 40\%, 50\%\}$. How missing data were introduced will be described in the next paragraph. Thirdly, the β-weights in Equation (1) were varied and were set to $\beta_1 = \beta_2 \in \{0.2, 0.5, 1\}$, which yielded average R^2-values of .13, .36, and .67, respectively, indicating small, medium, and large effect sizes, respectively. Finally, I compared classical PMM, as it is implemented in the function `mice.impute.pmm` from R package `mice` version 2.22 (van Buuren & Groothuis-Oudshoorn, 2011) using a donor pool of constant size k, with $k \in \{1, 3, 5, 10\}$ against the automatic distance-based donor selection variant `midastouch` version 1.3 (Gaffert et al., 2016).[1] This resulted in a total of 300 experimental conditions. Each condition was replicated 1,000 times.

Introduction of Missing Data

Analogous to the study by Schenker and Taylor (1996), values were deleted only from the dependent variable y, while the predictors remained completely observed. While Schenker and Taylor (1996) examined MCAR missingness – that is, missing completely at random (Rubin, 1976), missing data in this study followed a MAR (missing at random) mechanism (Rubin, 1976). Missingness in y depended on predictor x_1. I subsequently refer to x_1 as the "cause of missingness." The probability p_i^{mis} for each y_i to be missing was determined by a logit model:

$$p_i^{\text{mis}} = \text{invlogit}\,(-6 + x_{1i}). \quad (2)$$

Depending on the experimental condition, 10%, 20%, 30%, 40%, or 50% of observations in y were selected with probabilities p_i^{mis} from all y-values and their values were deleted.

The intercept of -6 in Equation (2) was chosen to get a wide range of selection probabilities, with low selection probabilities for the major part of the sample. They were thus mostly cases with large x_1-values that had a high chance of having a missing value in y. The minimum observed missingness probability was 0.33%, the maximum 98.72%. The median was 27.1%.

Data Imputation

Missing data were imputed using the R package `mice` (van Buuren & Groothuis-Oudshoorn, 2011). The number of imputations m was set to be equal to the respective missing data percentage.[2] If, for example, in one condition 50% of the data in y were missing, $m = 50$ imputations of each missing value were created. The complete variables x_1 and x_2 were used to predict missing data in y. Depending on the respective condition, either function `mice.impute.pmm` with $k \in \{1, 3, 5, 10\}$ or function `mice.impute.midastouch` was used to create the imputations.

Data Analysis

The linear model in Equation (1) was fitted to the m completed data sets from each replication in each condition and results were combined using Rubin's formula for MI inference (Rubin, 1987), with degrees of freedom being calculated by the Barnard and Rubin (1999) correction formula.

Quality Criteria

I evaluated PMM's performance in terms of estimation accuracy and estimation consistency regarding the marginal mean, precision of standard error estimates, and in terms of how well the procedure was able to preserve the original distribution of y (cf. Schenker & Taylor, 1996).

[1] R, `mice`, and `midastouch` are available from https://cran.r-project.org.

[2] Research by Bodner (2008) suggests that the quality of statistical inferences could be improved by using more than the formerly standard $m = 5$ imputations. Bodner proposed a rather complex procedure to determine m based on the estimated fraction of missing information λ. I used a simpler proxy and set m equal to the observed missing data percentage in y.

A measure for estimation accuracy is the average distance of the estimate from the true parameter – that is, bias, which is defined as $Q - \widehat{Q}$ where Q is the population quantity and $\widehat{Q}$ is the average parameter estimate across the 1,000 replications. I report relative bias, which is defined as $\frac{Q-\widehat{Q}}{Q} \times 100\%$, and which makes the quantity independent from the respective scale of measurement. Note that there are no commonly agreed on criteria, when bias or relative bias is "significant." I follow Forero and Maydeu-Olivares (2009), who deem absolute values of less than 10% parameter bias as acceptable.

Secondly, consistency in parameter estimation is reflected in the variance of the estimates across the 1,000 replications. A small variance signifies consistently good estimates across the replicated samples. A hybrid measure that reflects both accuracy and consistency in parameter estimation is the root mean square error (RMSE), defined as $\text{RMSE} = \sqrt{\text{bias}^2 + \text{variance}^2}$, which signifies the typical distance between the estimate and the true value. Note that there are no definite criteria as to when either the variance of the parameter estimates or the RMSE is unacceptably large. Obviously, one would like both quantities to be small, signifying a precise and consistent inference (cf. Rubin, 1996).

Thirdly, I report coverage rates (CR), a hybrid measure that reflects both bias in parameter estimates and bias in standard error estimates. CR is defined as the percentage of 95% confidence intervals that include the true parameter. A coverage rate close to 95% indicates that the standard error estimates are large enough, so that the true parameter is inside the interval most of the time. Schafer and Graham (2002) deem rates below 90% – the double of the nominal error rate – as seriously low. Undercoverage may result from large biases (so that the interval is too far to the left or to the right to cover the true parameter), from underestimated standard errors, or from a combination of both factors.

Finally, with regard to how PMM was able to preserve the original distribution of y, I looked at the percentages of the respective sample with observations greater than the "true" 5th, 25th, 50th, 75th, and 95th percentiles of y (cf. Schenker & Taylor, 1996). Values should obviously be close to 95%, 75%, 50%, 25%, and 5%, respectively.

Results

Complete Data Results

Firstly, to test that the simulation of the data worked well, I computed parameter estimates and quality statistics based on the complete data (i.e., before any missing data were introduced). Like Schenker and Taylor (1996), I focused on results regarding the marginal mean and the distribution of y. These values are displayed in Table 1. As can be seen in that table, the simulation worked well and the parameter estimates of the marginal mean were close to the specified population quantity. Furthermore, coverage was around 95% and the estimated quantiles of the simulated data were very close to the theoretical quantiles. Note that the variance estimates (and thus also the RMSE estimates) of the marginal mean were larger on average, the smaller the sample size got. Variance ranged from 0.51 (N = 100) to 3.26 (N = 20). RMSE was approximately 1.8, when N = 20; 1.4, when N = 30; 1.1, when N = 50; and around 0.7, when N = 100. Naturally, statistical inferences become more stable and more precise, when based on larger rather than smaller samples. We need to bear this in mind, when discussing results of the different PMM conditions regarding variance and RMSE later on.

Fractions of Missing Information

Secondly, to convey an idea of the extent of the missing data problems simulated in this study, I list the estimated average fractions of missing information λ (cf. Schafer, 1997a) regarding the marginal mean, which quantify the level of uncertainty about the imputed values, and the impact missing data have on (a) the respective parameter estimate and (b) the performance of statistical tests: the average λ was .09, when p_{mis} = 10%; .17, when p_{mis} = 20%; .25, when p_{mis} = 30%; .32, when p_{mis} = 40%; and .39 when p_{mis} = 50%. λ has a range between 0 and 1. Generally, the larger λ, the more biased results could be.

I now present results regarding how well the respective PMM variants were able to estimate the marginal mean, its standard error, and the quantiles of y in the simulated scenarios.

Relative Bias

Results regarding relative bias in the estimates of the marginal mean are summarized in Figure 1. Generally, biases got larger, the larger the donor pool was, the more data had to be imputed and the larger the β-weights were. Larger β-weights signified a "stronger" missing data mechanism, with stronger referring to the relationships between the cause of missingness x_1, y, and missingness in y (see Equations 1 and 2). Furthermore, biases generally increased with decreasing sample size.

Overall, nearest neighbor PMM produced the best results. When $N \geq 50$, all estimates were sufficiently accurate. When N = 30, bias was found only in the most extreme condition, where β = 1 and 50% of the data had

Table 1. Complete data estimates

N	β	%BIAS	VAR	RMSE	CR	% > P5	% > P25	% > P50	% > P75	% > P95
20	0.2	0.31	3.24	1.80	95.50	94.86	75.38	50.48	25.13	4.84
20	0.5	0.42	3.26	1.81	94.80	94.85	75.31	50.43	25.14	5.14
20	1.0	−0.01	3.09	1.76	95.20	95.03	75.29	49.81	25.15	5.03
30	0.2	0.00	1.83	1.35	95.80	95.01	74.71	49.85	25.11	4.92
30	0.5	0.27	1.93	1.39	95.50	95.10	75.08	49.68	25.40	5.16
30	1.0	0.06	1.93	1.39	95.20	95.09	75.09	50.02	25.23	5.05
50	0.2	0.47	1.15	1.07	94.40	95.07	74.83	49.84	24.91	5.00
50	0.5	−0.05	1.11	1.05	94.50	95.13	75.18	50.06	25.16	4.98
50	1.0	0.14	1.13	1.06	95.50	95.20	75.11	50.21	24.95	5.04
100	0.2	0.07	0.55	0.74	94.70	94.90	75.01	49.94	24.92	4.95
100	0.5	0.17	0.58	0.76	94.50	94.95	75.00	49.83	24.96	5.00
100	1.0	−0.28	0.51	0.72	95.30	95.07	75.22	50.00	24.82	4.93

Notes. *N* = sample size; β = parameters β_1 and β_2 in the regression model, which were set to be equal; %BIAS = relative bias of the marginal mean, defined as BIAS/Q × 100%, where Q is the set population quantity and BIAS is defined as $Q - \widehat{Q}$, where $\widehat{Q}$ is the average estimate across the 1,000 replications; VAR = variance of the estimates across the replicated samples; RMSE = root mean squared error of the marginal mean; CR = coverage rate, that is, the percentage of 95% confidence intervals that include the "true" parameter. The columns "% > P5"–"% > P95" denote the percentage of the sample with *y*-values larger than the respective percentile.

to be filled in. Here, the marginal mean was overestimated by 11.4%. Furthermore, even when $N = 20$, setting $k = 1$ produced widely accurate estimates. Only when β = 1 and $p_{mis} \geq 40\%$ severe biases could be observed.

Secondly, also sampling from a pool of $k = 3$ or $k = 5$ donors yielded low biases in many scenarios. Especially when the model R^2 was small to moderate (i.e., β ∈ {.2, .5}) and less than about 30%–40% of the data were missing, results were generally accurate enough.

Overall largest biases were found in the conditions, where $k = 10$. Especially when the missingness mechanism got stronger, more values in y were missing and sample size was 50 or less, downsizing the donor pool improved estimation accuracy quite noticeably: for example, when $N = 20$, β = 0.5, and $p_{mis} = 50\%$, setting $k = 10$ yielded a bias of −25.6%. Generating the imputations from a donor pool of size $k = 5$ decreased bias to −16.65%. Setting $k = 3$ produced a bias of −13.56%. In comparison, nearest neighbor PMM yielded the lowest bias in this condition of −9.42%.

Finally, also the automatic donor selection procedure `midastouch` yielded sufficiently accurate estimates in most scenarios. Biases in fact were highly similar to those obtained by sampling from a donor pool of size $k = 3$. When $N = 100$, estimates were generally unbiased. When $N = 50$, bias of −10.6% was found in the most extreme condition, where β = 1 and 50% of the data in y were missing. When $N = 30$, large biases were only found in the more extreme scenarios with 50% missing data in y, when β = 0.5, or with $p_{mis} \geq 40\%$, when β = 1.

In summary, classical PMM produced sufficiently accurate estimates most of the time – given that reasonable settings regarding donor selection were applied. Decreasing the size of the donor pool became more important, the smaller the sample size got, especially the more data were missing and the stronger the missing data mechanism became. Furthermore, also automatic distance-aided selection of donors yielded sufficiently accurate estimates in many scenarios. Results were by and large comparable to using a donor pool of size $k = 3$.

Variance of the Parameter Estimates and Root Mean Square Error

RMSE estimates are given in Figure 2, the variance of the parameter estimates across the 1,000 replications is presented in Figure 3. RMSE estimates depended on both the β-weights (with higher values leading to larger RMSE estimates) and missing data percentages (with larger percentages leading to larger RMSE estimates). Furthermore, RMSE estimates increased, when the size of the donor pool increased, but noticeably only in those conditions, where the model R^2 was medium to large and more than about 20%–30% of the data had to be imputed. This was mainly due to the fact, that both bias (see Figure 1) and variance (see Figure 3) increased with increasing missing data percentages and increasing size of β. On the other hand, variance estimates usually were "better" (i.e., lower) using larger donor pools: for example, when $N = 100$, the largest variance estimate was 1.48, when $k = 1$; 1.29, when $k = 3$; 1.21, when $k = 5$; and 1.11, when $k = 10$. In comparison, `midastouch` here yielded a largest estimate of 1.32. In comparison to the complete data estimates (Table 1), PMM produced similar Monte Carlo variance estimates only in some conditions, especially when the missing data

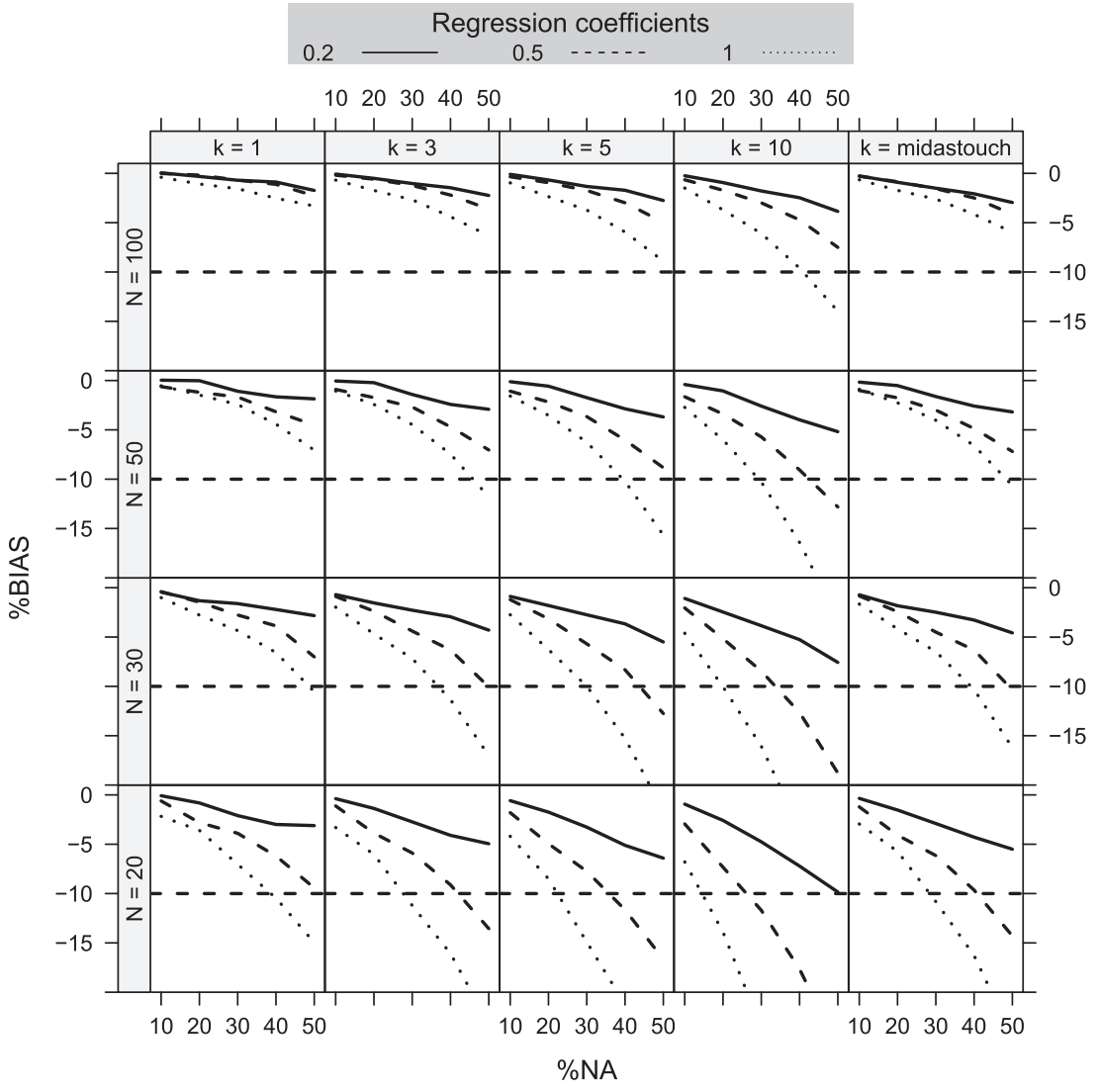

Figure 1. Percent parameter bias of the marginal mean. %BIAS is the relative bias of the marginal mean, defined as BIAS/$Q \times 100\%$, where Q is the set population quantity and BIAS is defined as $Q - \hat{Q}$, where $\hat{Q}$ is the average estimate across the 1,000 replications. N is the sample size; % NA is the missing data percentage; k refers to the donor selection strategy.

percentage was small to moderate and the R^2 of the regression model was small to moderate. Especially when the missingness mechanism was strong, missingness percentages were high, and the donor pool was small, PMM estimates were not as consistently good across the replications as one could have wished for.

Coverage Rates

So far, results have shown that overall, small donor pools produced more accurate estimates than large donor pools. However, as already mentioned, estimation accuracy might come at the cost of underestimated standard errors (cf. Schenker & Taylor, 1996). Coverage rates – as a hybrid measure that reflects both the adequateness of parameter estimates and their standard errors – help to determine, which PMM settings produce an acceptable trade-off between unbiased parameter estimates and unbiased standard errors. These values are summarized in Figure 4.

Firstly, it can be seen that overall, automatic, distance-based donor selection (`midastouch`) produced the best results in terms of coverage. Though coverage was generally somewhat lower, the larger β and thus the strength of the missing data meachanism was, nearly all coverage rates lay above 90%. Suboptimal coverage rates were found only in two conditions, namely when 50% of the data were missing, β = 1, and N = 30 – here coverage was 89.9%. Furthermore, when N = 50, β = 1, and p_{mis} = 50%, coverage rate was 88.7%. The average coverage rate obtained by `midastouch` across all conditions was 94.0%.

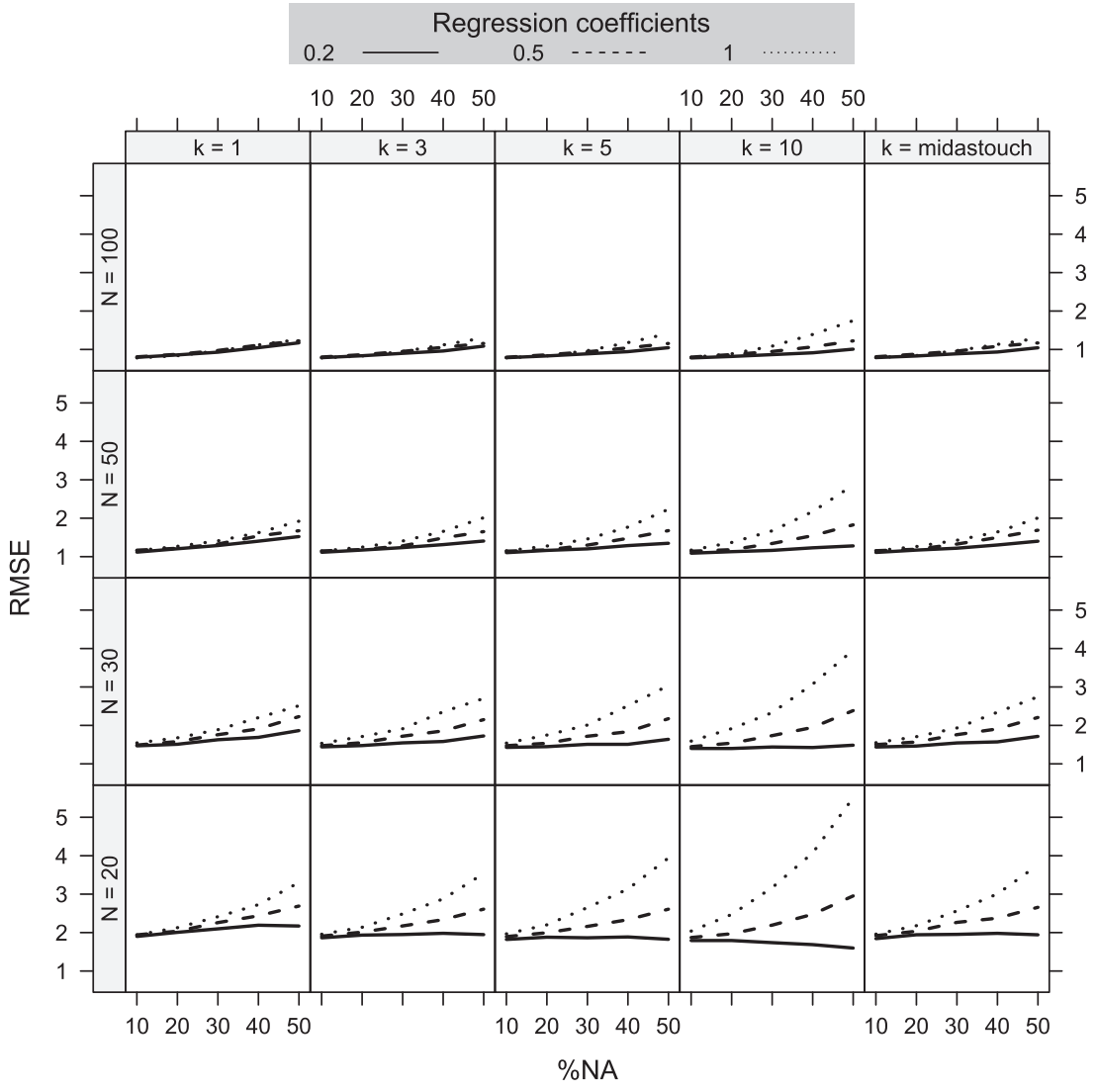

Figure 2. Root mean squared error (RMSE) of the marginal mean. *N* is the sample size; %NA is the missing data percentage; *k* refers to the donor selection strategy.

Secondly, nearest neighbor PMM also produced acceptable results in many scenarios. However, when $k = 1$, the average coverage rate across all conditions of 91.2% was noticeably lower than the one obtained by `midastouch`. Significant undercoverage was found in the more extreme conditions, when 40% or 50% of the data were missing, and β was either 0.5 or 1, regardless of the sample size. It seems that while nearest neighbor PMM produced more accurate parameter estimates in general, `midastouch` produced a better overall trade-off between both accurate parameter estimates and accurate standard errors.

Thirdly, generating the imputations from a pool of three nearest neighbors also produced acceptable results in many scenarios. However, again, coverage rates were on average lower in comparison to using `midastouch`. The average coverage rate across all conditions was 91.9%. Undercoverage was found in some of the more extreme scenarios, where $\beta \in \{0.5, 1\}$ and $p_{\text{mis}} \geq 40\%$.

With an average coverage rate of 91.7% across all scenarios, also `mice`'s default setting of sampling from a fixed pool of five donors yielded mostly acceptable results. Note, however, that the drop in coverage was more pronounced in the extreme scenarios in comparison to using $k = 1$, $k = 3$, or the `midastouch` procedure. This effect got even stronger, when the size of the donor pool was increased to $k = 10$. Here, the average coverage rate across all conditions dropped down to 90.3%.

All in all, results were best using either automatic distance-based donor selection or when imputations were generated from a small donor pool.

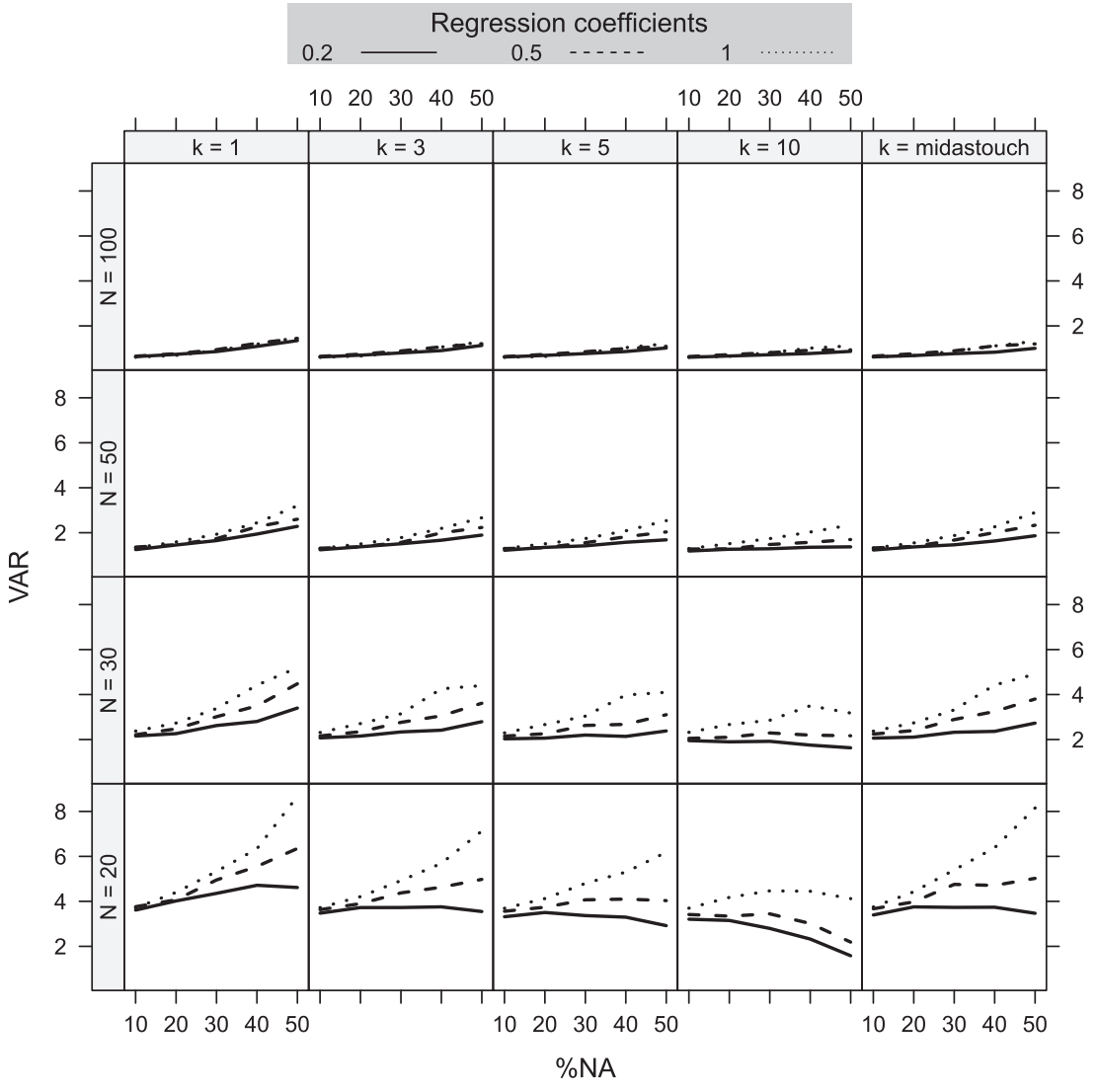

Figure 3. Variance (VAR) of the estimates of the marginal mean across the 1,000 replications. *N* is the sample size; %NA is the missing data percentage; *k* refers to the donor selection strategy.

Estimation of Quantiles

Finally, I examined, how well the different PMM settings fared in preserving the distribution of *y*. Figure 5 displays the percentages of the sample with values larger than the respective percentile separately for the four different sample size conditions. Results were averaged across all β-weight conditions. Each panel in Figure 5 covers the range ±5% of the respective percentile, the gray shaded areas therein denote the range ±2% of the respective percentile. Note that all complete data estimates were within the range ±1% of the respective percentile (cf. Table 1). PMM should not perform noticeably worse. I defined "noticeably worse" as producing estimates outside the interval ±2% of the respective percentile. Note that, again, there are no definite criteria as to when bias in the estimates of percentiles should be regarded as significant.

In general, estimation precision increased with increasing sample size. Most of the time, estimates by classical PMM were reasonable, when $N \geq 50$. When $N = 100$, all estimates were acceptable. When $N = 50$, only sampling from a large pool of 10 donors yielded biased estimates of some percentiles, when 40% or more of the data were missing. However, the smaller the sample became, the smaller the donor pool had to be to obtain accurate estimates: When $N = 30$, nearest neighbor PMM or sampling from a pool of $k = 3$ donors yielded overall acceptable estimates of the respective quantiles. Generating the imputations from larger donor pools yielded biased estimates of some quantiles. Biases here ranged from 2.07% to 4.95%.

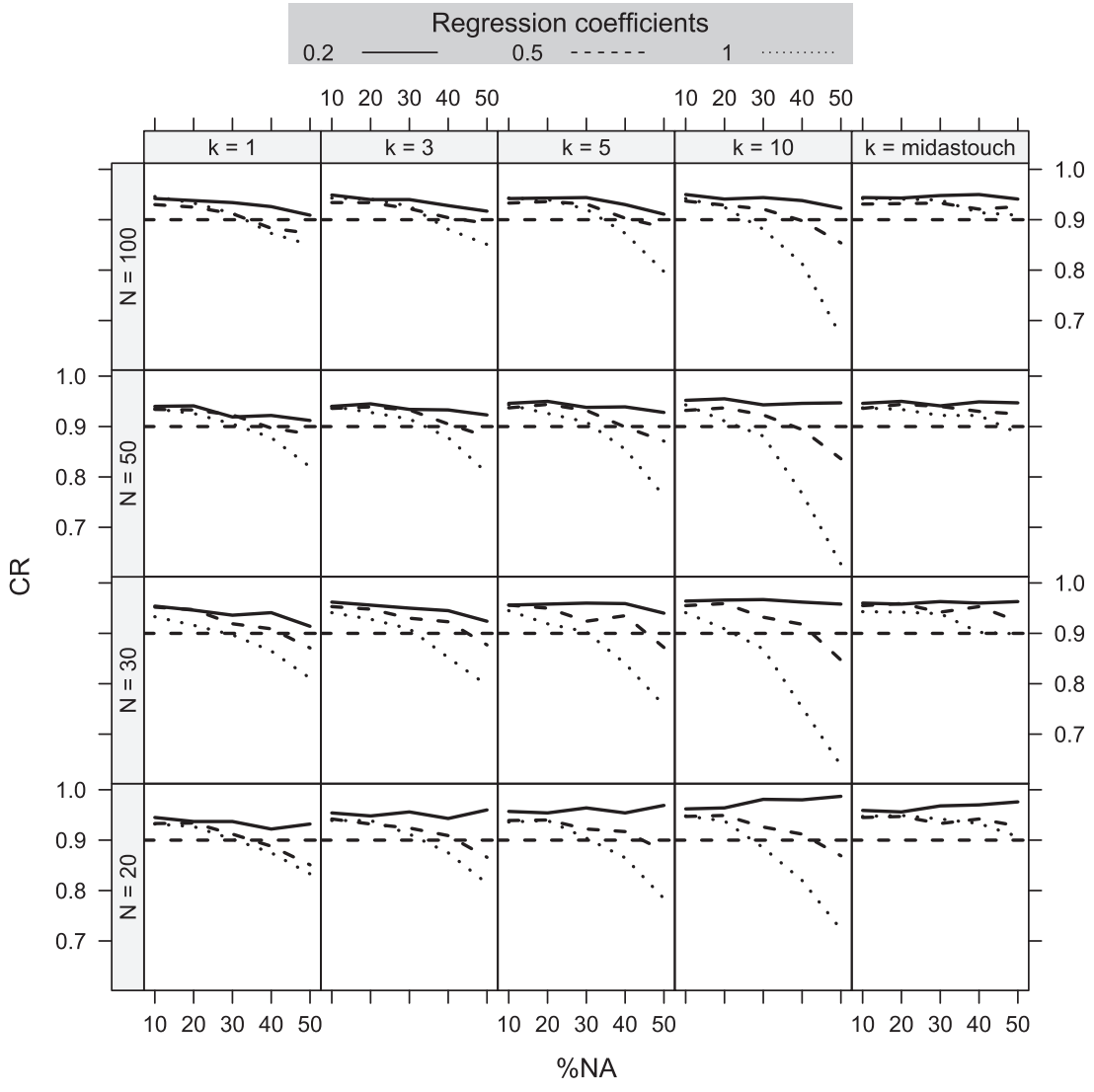

Figure 4. Coverage rates of the marginal mean. CR is the coverage rate of the marginal mean, that is, the fraction of its 95% confidence intervals that include the "true" parameter; *N* is the sample size; %NA is the missing data percentage; *k* refers to the donor selection strategy.

When $N = 20$, again setting $k = 1$ or $k = 3$ produced the most accurate estimates among the fixed donor pool conditions. Generally, biases increased both with increasing missing data percentage and with increasing size of the donor pool. In comparison, `midastouch` worked well, when $N \geq 30$. When $N = 30$, `midastouch` produced only one suboptimal estimate of the 25th percentile, when 50% of the data were missing. The estimate here was 2.47% off.

Separate results for the three β-weight conditions are not shown in Figure 5: estimates, however, were good and results of the different PMM variants were hardly discernible, when up to 30% of the data were missing. When more data had to be imputed, downsizing the donor pool improved the accuracy of results noticeably, the stronger the missingness mechanism became. In this case, best results were obtained by nearest neighbor PMM.

Discussion

The performance of PMM multiple imputation has been evaluated in various scenarios, in which sample size, missing data percentage, the coefficients of the β-weights in the regression model, and the donor selection strategy were systematically varied. Performance has been evaluated in terms of estimation accuracy and consistency regarding estimates of the marginal mean and corresponding standard errors, and in terms of how well the procedure was able to preserve the original distribution of the variable.

Firstly, both classical PMM and the automatic donor selection variant `midastouch` (Gaffert et al., 2016) yielded accurate statistical inferences in many scenarios. Findings of this study thus do not corrobarate the general caveat that PMM might not be an option for small data sets.

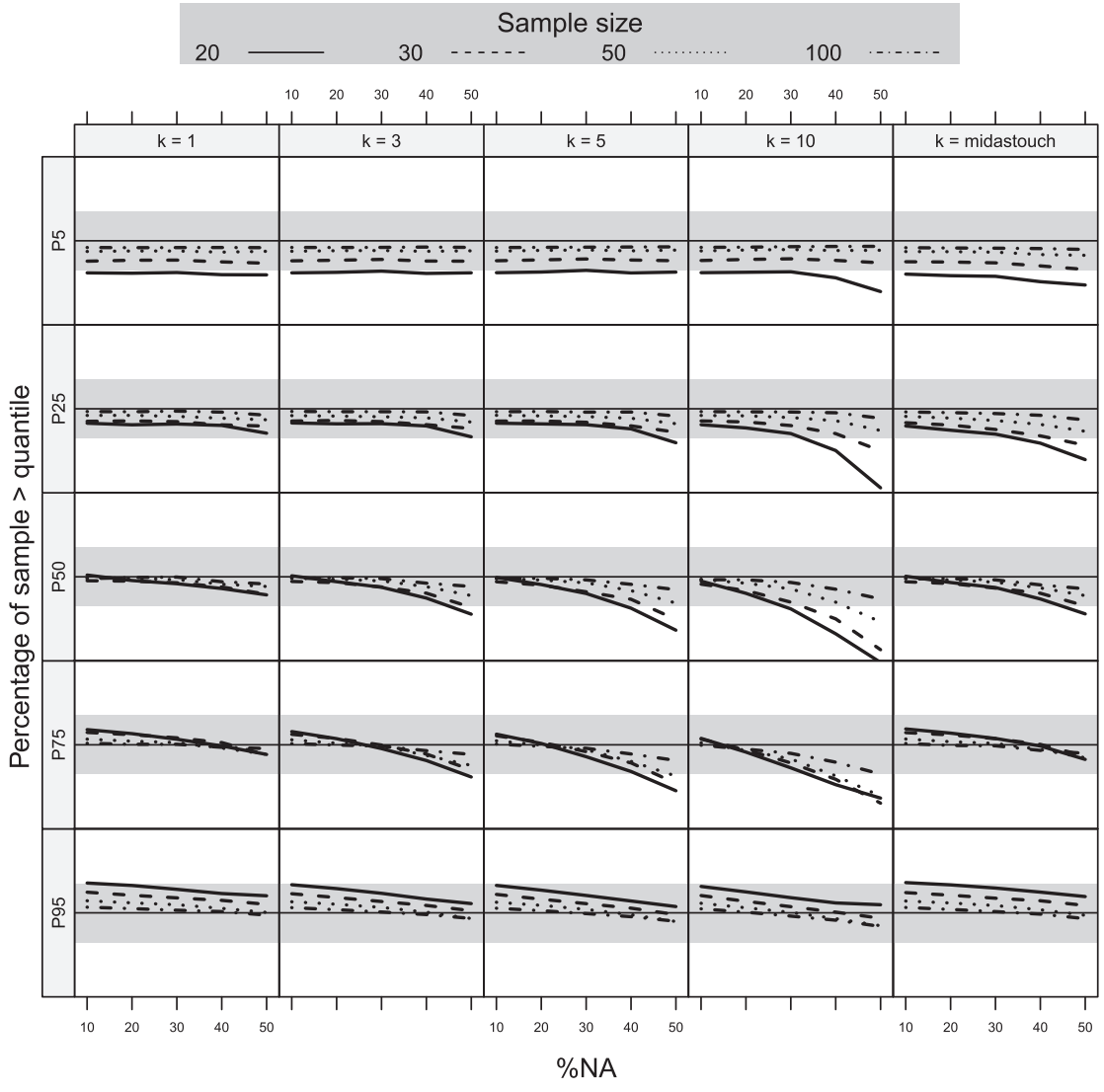

Figure 5. Estimation accuracy of quantiles (averaged over β-weight conditions). The panels display the respective percentage of the sample with values larger then the respective percentile. P5–P95 denote the respective percentiles; %NA is the missing data percentage; *k* refers to the donor selection strategy.

Secondly, as expected, the magnitude of the observed biases depended not only on the interplay of various factors: the size of the donor pool, the missing data percentage, and sample size, but also on the size of the regression coefficients in the data generating model. That the size of the regression coefficients had an effect on statistical inferences is because missingness probabilities in this study depended on one of the predictors, and as a consequence larger β-weights also corresponded to "stronger" MAR mechanisms. Stronger in this context means that the relationships between the predictor, the dependent variable, and missingness in the dependent variable became stronger (see Equations 1 and 2). With increasing size of the regression coefficients, more values from the upper half and the tail of the distribution of y were deleted. This required the missing data procedure to make at least some extrapolations beyond the range predicted by the remaining observed cases – something PMM cannot do (cf. van Buuren, 2012, Chap. 3). As statistical inferences are based on both observed and imputed values, this effect obviously increased with increasing missing data percentage. Secondly, the effect that biases got stronger with increasing size of the β-coefficients was further magnified with an increasing size of the donor pool, and also with decreasing sample size. This is because a decreasing sample size means that fewer suitable donors will be available in the sample. Furthermore, sampling from a large donor pool (in relation to sample size) increases the chance that also rather "dissimilar" cases will be selected, which in turn increases the chance of obtaining biased estimates.

Consequently, nearest neighbor PMM produced the overall lowest biases. Additionally, coverage rates were sufficiently large most of the time. This implies that the standard error estimates were still large enough and that the decrease in between-imputation variance due to a small donor pool did not have an overly huge detrimental effect. Furthermore, setting $k = 3$ or $k = 5$ also produced acceptable coverage rates in many scenarios. Undercoverage was mainly found in the more extreme scenarios, when about 30% or more of the data were missing. Though biases were usually somewhat larger, when $k = 3$ or $k = 5$ in comparison to using the nearest neighbor (see Figure 1), it appears that these biases were buffered by sufficiently large confidence intervals, resulting in adequate coverage, when the missing data problem was not too severe.

As many statistical packages use either $k = 3$ or $k = 5$ as default (cf. Allison, 2015), these are important findings for practitioners, who naturally want to focus on data analysis without having to reflect too much on the issue, what PMM settings would be most appropriate in a given scenario. Worst overall results were found, when $k = 10$.

Thirdly, also the automatic distance-based donor selection procedure `midastouch` (Gaffert et al., 2016) yielded good results. While the overall lowest biases were obtained by using nearest neighbor PMM, `midastouch` produced overall highest coverage rates. It appears that the touched-up version of the MIDAS approach (Siddique & Belin, 2008) yielded more appropriate standard error estimates in comparison to classical nearest neighbor PMM.

Finally, one finding that – on first glance – seemed to be counterintuitive, was that the variance of the estimates across the 1,000 replications tended to decrease with an increasing size of the donor pool. One possible explanation for this finding is that choosing a large k in relation to sample size could have produced a central tendency trend toward a biased marginal mean estimate. Gaffert et al. (2016), for example, stated that "if the distributions of the donors and recipients are roughly comparable then a large k will increase the probability for the donors closer to the center to give their value to the recipients closer to the bounds. That inevitably decreases the variance of y" (p. 6; see also the Appendix in Gaffert et al., 2016). This in consequence could also have decreased the variance of the estimates across the replications. To see, if a central tendency trend might be a plausible explanation here, we need to have a closer look at some of the results. For example, when $N = 20$, $\beta = .2$, $p_{mis} = 30\%$, and $k = 10$, the average estimate of the marginal mean was $\widehat{Q} = 10.48$, the variance of the estimates across the replications was VAR = 2.80, the interquartile range (IQR) was 2.21, with a total range of between 3.90 and 17.67. In comparison, when only $k = 1$ donor was used in the same scenario, the average estimate of 10.21 was closer to the true value of 10, however its variance estimate was larger (4.36), and also the IQR of 2.70 and the total range of between 1.60 and 20.00 were larger. Similar results were also found in other conditions. It appears that using a large donor pool indeed yielded more centered estimates around a biased marginal mean. Future research should explore this effect further.

All in all, results suggest that PMM could be used for missing data imputation in small data sets, when a reasonable donor selection strategy is applied. However, results also imply that practitioners should try to get larger samples. Increasing sample size by even 10 or 20 participants helped to increase the accuracy of statistical inferences quite noticeably.

Limitations

No single simulation study can cover all relevant aspects of interest. Focusing on some aspects makes it necessary to disregard others. The present study, for example, considered only a very basic model with two predictors. Future research could systematically vary the number of predictors in the model, and their relationships with both the dependent variable and how well they predict missingness. Relationships could additionally be more complex, including interactions and higher-order relationships. Furthermore, in this study, only the dependent variable contained missing data, while the predictors remained completely observed. Future research could address more complex missing data scenarios with missingness on both sides of the equation. Also, the model in this study was homoscedastic – and all distributional assumptions were met. While some studies already tested the robustness of PMM toward violations of distributional assumptions (e.g., Kleinke, 2017; Yu et al., 2007), future research should look into this in greater detail in the context of small sample sizes. Finally, Monte Carlo Simulations are naturally artificial. Future simulations could also be based on empirical data sets, therefore being more realistic.

References

Allison, P. D. (2015, March 5). *Imputation by predictive mean matching: Promise & peril*. Retrieved from http://statisticalhorizons.com/predictive-mean-matching

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review, 78*, 40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika, 86*, 948–955. https://doi.org/10.1093/biomet/86.4.948

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling, 15*, 651–675. https://doi.org/10.1080/10705510802339072

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14*, 275–299. https://doi.org/10.1037/a0015825

Gaffert, P., Meinfelder, F., & Bosch, V. (2016). *midastouch: Towards an MI-proper predictive mean matching*. Discussion paper. Retrieved from https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/sowilehrstuehle/statistik/Personen/Dateien_Florian/properPMM.pdf

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician, 61*, 79–90. https://doi.org/10.1198/000313007X172556

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician, 57*, 229–232. https://doi.org/10.1198/0003130032314

Kleinke, K. (2017). Multiple imputation under violated distributional assumptions – A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics, 42*, 371–404. https://doi.org/10.3102/1076998616687084

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*, 287–296. https://doi.org/10.1080/07350015.1988.10509663

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology, 14*, 75–87. https://doi.org/10.1186/1471-2288-14-75

Parzen, M., Lipsitz, S. R., & Fitzmaurice, G. M. (2005). A note on reducing the bias of the approximate Bayesian bootstrap imputation variance estimator. *Biometrika, 92*, 971–974. https://doi.org/10.1093/biomet/92.4.971

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics, 4*, 87–94. https://doi.org/10.1080/07350015.1986.10509497

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473–489. https://doi.org/10.1080/01621459.1996.10476908

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association, 81*, 366–374. https://doi.org/10.1080/01621459.1986.10478280

Schafer, J. L. (1997a). *Analysis of incomplete multivariate data*. London, UK: Chapman & Hall.

Schafer, J. L. (1997b). *Imputation of missing covariates under a general linear mixed model* (Technical Report 97–10). University Park, PA: Pennsylvania State University, The Methodology Center.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schenker, N., & Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis, 22*, 425–446. https://doi.org/10.1016/0167-9473(95)00057-7

Siddique, J., & Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine, 27*, 83–102. https://doi.org/10.1002/sim.3001

Siddique, J., & Harel, O. (2009). MIDAS: A SAS macro for multiple imputation using distance-aided selection of donors. *Journal of Statistical Software, 29*, 1–18. https://doi.org/10.18637/jss.v029.i09

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall/CRC.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Vink, G., Frank, L. E., Pannekoek, J., & van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica, 68*, 61–90. https://doi.org/10.1111/stan.12023

Yu, L. M., Burton, A., & Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research, 16*, 243–258. https://doi.org/10.1177/0962280206074464

Received August 25, 2015
Revision received August 11, 2016
Accepted September 18, 2017
Published online April 23, 2018

Kristian Kleinke
Department of Psychology
Bielefeld University
Postfach 10 01 31
33501 Bielefeld
Germany
kristian.kleinke@uni-bielefeld.de

Kristian Kleinke is a postdoctoral researcher at the University of Bielefeld. His primary research interests are missing data and multiple imputation. He focuses on imputation solutions for complex data structures like panel data, and "non-normal" missing data problems, that is, when convenient distributional assumptions of standard MI procedures are violated.