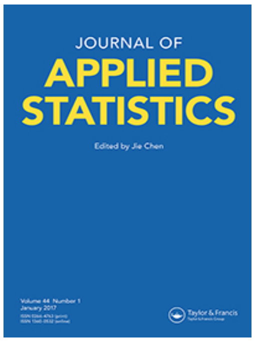

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Missing data methods for arbitrary missingness with small samples

Daniel McNeish

To cite this article: Daniel McNeish (2017) Missing data methods for arbitrary missingness with small samples, Journal of Applied Statistics, 44:1, 24-39, DOI: 10.1080/02664763.2016.1158246

To link to this article: https://doi.org/10.1080/02664763.2016.1158246

Published online: 22 Mar 2016.

JOURNAL OF APPLIED STATISTICS, 2017
VOL. 44, NO. 1, 24–39
http://dx.doi.org/10.1080/02664763.2016.1158246

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Missing data methods for arbitrary missingness with small samples

Daniel McNeish[a,b]

[a]Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD, USA; [b]Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

ABSTRACT

Missing data are a prevalent and widespread data analytic issue and previous studies have performed simulations to compare the performance of missing data methods in various contexts and for various models; however, one such context that has yet to receive much attention in the literature is the handling of missing data with small samples, particularly when the missingness is arbitrary. Prior studies have either compared methods for small samples with monotone missingness commonly found in longitudinal studies or have investigated the performance of a single method to handle arbitrary missingness with small samples but studies have yet to compare the relative performance of commonly implemented missing data methods for small samples with arbitrary missingness. This study conducts a simulation study to compare and assess the small sample performance of maximum likelihood, listwise deletion, joint multiple imputation, and fully conditional specification multiple imputation for a single-level regression model with a continuous outcome. Results showed that, provided assumptions are met, joint multiple imputation unanimously performed best of the methods examined in the conditions under study.

ARTICLE HISTORY
Received 4 January 2015
Accepted 22 February 2016

KEYWORDS
Small sample; missing data; multiple imputation; full information maximum likelihood; incomplete data; finite sample; Monte Carlo simulation

Introduction

Missing data are a pervasive analytic problem across research disciplines and can complicate even the simplest analyses because improper treatment of missing values may potentially bias model estimates and interpretations while conclusions drawn from the model may be faulty as a result [10,26,29,32]. Since the 1970s, a rapidly growing body of literature has addressed the inherent difficulty in both classifying how and why data are missing and how to appropriately handle missing values so that estimates are both trustworthy and representative of population dynamics [10,30].

Although many methodological studies have investigated the properties of the various missing data methods under a variety of conditions [1,2,20,21,23,33,35], very little research has been conducted specifically on the performance of these methods with small samples sizes, which are a common analytic challenge in behavioral sciences as human subjects are expensive and difficult to acquire. Barnes et al. [6] carried out a comprehensive simulation

CONTACT Daniel McNeish ✉ d.m.n.mcneish@uu.nl

study on the performance of multiple imputation (MI) methods for monotone missingness with small samples in clinical trials. Monotone missingness does not require iterative processes and this pattern of missing values is not nearly as common as in behavioral sciences as it is in clinical trials in a biomedical setting. Graham and Schafer [16] also conducted an illustrative simulation study as part of a book chapter that showed that a particular method (joint multiple imputation (JMI)) provided reasonably unbiased estimates of standard errors and point estimates with linear regression with as few as 50 observations. Their study did not include methods other than JMI, however, and is also slightly dated as many advances in both methodology and software have been developed since its publication (e.g. Proc MI and Proc MIANALYZE in SAS, more widespread implementation of alternative types of MI, Barnard & Rubin degrees of freedom with MI and small samples [5], ease of implementing maximum likelihood (ML) in programs like M*plus* and SAS Proc CALIS).

This paper will first briefly overview the classification of missing data and common methods to handle missing values, slanted slightly towards small sample contexts. Then, missing data in the context of small samples will be discussed and the limited existing literature will be reviewed. Lastly, a comparative simulation targeting the small sample performance of listwise deletion (LD), fully conditional specification, ML, and JMI will be provided and its implications considered.

An overview of common missing data terminology and methods

Missing data patterns and missing data mechanisms

A primary concern with missing data is the arrangement of missing and observed values in the data, commonly referred to as the missing data pattern [10,32]. Although there are many different classifications, three general patterns are commonly discussed. The first is a univariate pattern where all individuals have observed data for all variables except a single variable in which all missing values are contained. The second type is monotone missingness which is commonly seen in longitudinal data. With a monotone pattern (also referred to as dropout), when an individual is missing data at one time-point, values for all subsequent time-points are also missing for that individual. The third general pattern of missingness and the pattern of interest in this study is arbitrary missingness which is sometimes informally referred to as a 'Swiss cheese' pattern. With arbitrary missingness, there is no set structure for which variables or participants have missing values. We will focus on the third missing data pattern in this paper.

Based on Rubin's [26] commonly used classification system, the mechanism responsible for missing data is often described as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). With an MCAR mechanism, the probability that a value is missing is not related to any other variables, regardless of whether the variables are included or not included in the data or model (e.g. data are missing from a coding error). With a MAR mechanism, missing values are related to other variables in the data but not to variables not included in the dataset or to the variable itself that is missing values (e.g. missing values for income are related to age but not to the value of income itself). With an MNAR mechanism, the probability that a variable is missing values is related either to variables that are not included in the data or to the hypothetical values on the variable of interest were it not missing (e.g. low-income respondents not reporting their income because it is low). Notationally, for R a binary indicator of whether data are missing, Y_{OBS} data that are observed, Y_{MIS} data that are unobserved, and complete data $Y_{\text{COM}} = (Y_{\text{OBS}}, Y_{\text{MIS}})$, an MCAR mechanism can be written as $P(R|Y_{\text{COM}}) = P(R)$, an MAR mechanism as $P\,(R|Y_{\text{COM}}) = P(R|Y_{\text{OBS}})$, and an MNAR mechanism as $P(R|Y_{\text{COM}}) = P(R|Y_{\text{OBS}}, Y_{\text{MIS}})$, [10,26,29].

Seaman et al. [31] further differentiate between two types of MAR. The definition first proposed by Rubin [26] concerns *realized MAR* which Seaman et al. [31] describe as the probability that a value is missing is not related to any other missing information *for missing data patterns that happened to be realized* (but not for all possible missing data patterns). Seaman et al. [31] also defines a broader type of MAR which subsumes realized MAR–*everywhere MAR*. Seaman et al. [31] describe everywhere MAR as the probability that a value is missing is not related to any other missing information *for all realized and unrealized missing data patterns*. To elucidate the distinction, Seaman et al. [31] provide an example where data are collected for a single variable X and all observations are complete. Missingness here would be realized MAR but not everywhere MAR. For the realized missingness patterns, missing values (which do not exist) are not related to unobserved information. However, for any possible patterns of missing values, the missing values would depend on unobserved information because the data only consist of a single variable. Next, we will briefly overview the different methods (and associated terminology) used in this paper to accommodate missing data.

Listwise deletion

LD (a.k.a. complete case analysis) is straightforward and the most commonly used method to accommodate missing values in behavioral sciences [25]. In LD, any incomplete case is excluded from the analysis. For example, if a case is missing a single value on a single variable, the entire case will be removed from the analysis. When the data are MCAR, LD will produce unbiased estimates because MCAR data are essentially a random subsample of complete data from the overall sample (estimates will also be unbiased in select scenarios where data are not MCAR, see [7,22]). However, a major drawback of LD, especially with the small samples of interest in this paper, is that the sample size is often heavily reduced. For instance, Enders [10] gives the example that with 10 variables and only 2% missingness on each variable, about 18% of the sample will be deleted. With large samples, this may not be an issue. For instance, with a sample size of 5000, losing 40% of observations is not desirable, but it will have a minimal effect on one's power to detect true non-null effects. On the other hand, with a sample size of 50, losing 40% of the observations will be far more drastic.

Maximum likelihood

ML (a.k.a. full information ML) accommodates missing data by allowing the log-likelihood to be built by summing information from each individual in the data [3,9]. To accomplish this, the covariance matrix and mean vector have dimensions specific to each individual. For instance, in a simple three-variable problem, the dimension of the covariance matrix in the log-likelihood would be 3×3 and the mean vector 3×1 for individuals with all complete data. However, if person i has missing data on one of the variables, the dimensions would change to accommodate only observed data such that the covariance matrix would instead be 2 × 2 and the mean vector 2 × 1. In essence, ML does not delete cases but rather makes full use of all the observed information that an individual can contribute; however, it works around missing values rather than imputing or predicting values for the missing observations. ML has been shown to be unbiased provided that data are MAR [2,9,11,12].

Multiple imputation

MI refers to a broad class of methods which share some commonalities. We will first discuss these common features prior to delving into the specifics of the different types of MI. Generally, MI rectangularizes the data by directly imputing values for the missing observations. This occurs in three general phases: the Imputation Phase, the Analysis Phase, and the Pooling Phase [10,29]. The goal of the Imputation Phase is to predict plausible values for the missing values from other observed values (various types of MI differ with respect to how these values are imputed). Inherent in the name, many different datasets are created, each with possibly different imputed values, leaving a researcher with several different versions of the data. The final goal is to obtain a single set of parameter estimates as if the original data were complete, so, as an intermediate step, in the Analysis Phase, the statistical model of interest is applied individually to each of the separate imputed datasets and the parameter estimates are then saved. For instance, if a regression model is of interest and five imputations were performed, then the regression model is applied five times, yielding five sets of regression coefficients. The saved estimates are then passed to the Pooling phase where the m different estimates for each parameter are combined into a single estimate using Rubin's formulas [27] to appropriately calculate the within-imputation and between-imputation variance.

Joint multiple imputation

JMI (a.k.a. joint model multiple imputation or multivariate normal imputation) imputes for the missing values from other observed values in the data through a joint (multivariate normal) posterior predictive distribution. With JMI for arbitrarily missing values, values are imputed using two iterative steps: an Imputation Step (I-Step) and a Posterior Step (P-Step). Estimates are unbiased under the assumption of MAR and multivariate normality although Schafer [29] has suggested that inferences from JMI are still reasonable when multivariate normality is not strictly upheld.

Fully conditional specification multiple imputation

Fully conditional specification (FCS) multiple imputation (a.k.a. multiple imputation with chained equations [MICE], sequential regressions) imputes data on a variable-by-variable basis with a separate imputation model for each variable rather than using a joint multivariate normal posterior predictive distribution for all variables [34,35]. Imputations for continuous variables that will be of interest in this study can be conducted with either a regression method [27] or a predictive mean matching method [17,32] although a main advantage of FCS is realized with categorical or non-normal outcomes [4,34]. With the regression method (FCS-R), imputed values are taken from a model whose coefficients are simulated from a posterior predictive distribution. With predictive means matching (FCS-P), a regression model is also used but for the purpose of identifying potential donor cases. For more detail on FCS and how it compares to JMI, readers are referred to [4,21,34,35].

The missing data problem with small samples

The leading methods for handling missing data in behavioral sciences (i.e. LD, FCS, MI, ML) each have non-trivial shortcomings with small samples that could affect the desirable properties attributed to these methods with larger samples. With LD, discarding information is a major concern with small samples because the amount of information from the onset is limited, making each observations all the more crucial to preserve. With sample sizes below 100 as are common in many behavioral science disciplines, losing 10% or more of the data (as is commonplace with LD with even minimal amounts of missingness) can have drastic effects on statistical power and vastly increase Type-II error rates. Put into more technical statistical language, even though LD can yield unbiased estimates in some instances, the efficiency of the estimates of LD will be quite poor with small samples.

Small sample concerns also exist with ML. Even though ML is often stated to be unbiased so long as the missing values are MAR, this property of ML is more accurately expressed by estimates being *asymptotically* unbiased when data are MAR meaning that finite sample sizes may be a cause for concern. This property is widely recognized in the small sample literature for other methods such as multilevel models – ML is rarely implemented because of the known downward bias of the estimates with as many as 50 or 100 clusters [24] and has, for all intents and purposes, been obviated by the related restricted ML estimator when outcomes are continuous. For missing data applications, although the finite sample bias of ML is a well-documented concern and despite the popularity of ML to handle missing data, no known methodological studies have investigated the performance of ML for missing values under small sample conditions to report where estimates begin to exhibit bias and whether the extent of the bias is detrimental to model interpretation. Savalei [28] did study the small sample performance of model fit criteria in structural equation models with missing data using ML, however.

Although Graham and Schafer [16] showed that JMI may be reasonably unbiased with smaller samples, the underlying concern with small samples lies in the accuracy of the imputation model. That is, both JMI and FCS-R rely on the observed values to impute missing values. As with standard regression analyses with small samples, the accuracy of predictions may be questionable and the final model estimates may potentially be biased as a result. As an additional consideration, JMI for arbitrary missingness relies on Markov Chain Monte Carlo (MCMC). With smaller samples, prior distributions have a much larger impact on the posterior distribution than with larger samples. Therefore, software defaults that intend for the prior distribution to be non-informative may influence the posterior to a greater degree than anticipated. Furthermore, the MCMC chain may have trouble converging to a stationary distribution with diminished sample sizes which would result in higher non-convergence rates. A suggested method to abet convergence is to use a ridge prior which essentially has the effect of adding uncorrelated observations to the data [10,18]. However, with small samples, adding even a small amount of individuals to help convergence may have a non-trivial impact of the resulting model estimates.

Small samples may be particularly problematic when using FCS-P approach as well. Predictive means matching imputes an observed value directly from a different individual in the data that has similar covariate values as the case whose value is missing. With small samples, there may be little overlap in the covariate values among individuals, however, meaning that the donor observations may not be very similar to the observation containing the missing value. Alternatively, one observation may be repeatedly used as a donor on several occasions if few other observations have satisfactory covariate overlap, meaning that many observations will have identical values which may artificially restrict the variance in the data [35].

Simulation study

Data generation model

A comparison of the performance of LD, FCS, ML, and JMI for arbitrary missing data will be evaluated with a simulation study. Given the lack of prior research in this area, the data generation model is fairly simple to ensure that any potential differences are attributable to differential performance of the methods and not to nuances associated with increasingly complex data generation models. The data generation model is formulated by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i, \tag{1}$$

where Y_i is a continuous outcome variable (for the ith observation), X_i are values of continuous predictor variables, β are regression coefficients, and ϵ_i is an independent and identically distributed residual. Predictors variables were generated from standard normal distributions and coefficients were generated such that population effect sizes (in terms of η^2) were as follows: $\eta_{\beta_1} = 0.025$ (a small effect), $\eta_{\beta_2} = 0.100$ (a medium effect), and $\eta_{\beta_3} = 0$ (a null effect). The variance of the residual, ϵ_i, was equal to 16, meaning that the unstandardized population coefficients were 0.72, 1.40, and 0.00 for β_1, β_2, and β_3, respectively. To make the data somewhat less artificial, predictor variables were generated such that each was correlated 0.15 with all others.

Two missing data conditions were included, (everywhere) MAR and MNAR. For the MAR condition, missing data were induced for Y_i, X_{1i}, and X_{3i} such that missingness was attributable solely to other observed variables in the model. Y_i was generated to be missing for individual i if the value of X_{2i} was among the lowest 10%, 20%, 30%, or 50% of the distribution in the sample (depending on the condition for the percentage of missing values, which is discussed in the next section), X_{1i} was generated to be missing for individual i if the value of X_{2i} was among the highest 10%, 20%, 30%, or 50% of the distribution in the sample, and X_{3i} was generated to be missing for individual i if the value of $(X_{1i} + X_{2i})$ was among the highest 10%, 20%, 30%, or 50% of the distribution.[1]

For the MNAR condition, an additional variable (X_{4i}) was generated but not included in the model. Missing values were then induced for Y_i, X_{1i}, and X_{3i} as a function of either the excluded X_{4i} variable or the true values of the particular variable. Specifically, Y_i was generated to be missing for individual i if the value ofX_{4i} was among the lowest 10%, 20%, 30%, or 50% of the distribution in the sample. X_{1i} was generated to be missing for individual i if the value of X_{1i} was generated to be among the highest 10%, 20%, 30%, or 50% of the distribution in the sample, and X_{3i} was generated to be missing for individual i if the value of $(X_{2i} + X_{4i})$ was among the highest 10%, 20%, 30%, or 50% of the distribution. MNAR missingness was generated in this way to represent different types of MNAR mechanisms (e.g. ‘external’ MNAR where missingness is related to a variable not included in the model and ‘internal’ MNAR where missingness is related to the hypothetical value of missing variable itself). For the MNAR condition, the analyses were conducted assuming MAR was upheld in order to investigate the sensitivity of the methods to the violation of the MAR assumption with small samples.

Simulation conditions and outcome measures

Four conditions for sample size were included (20, 50, 100, and 250), the percentage of missing data for each variable had four conditions (10%, 20%, 30%, and 50%), and missing data were accommodated with four methods (LD, FCS, ML, and JMI). For FCS, because the variables with missing values were continuous, two separate methods for imputation were considered – predictive means matching (FCS-P) and regression (FCS-R). For each MI condition (FCS-R, FCS-P, and JMI), three conditions for the number of imputations were utilized (5, 25, and 100), and the imputation model included all variables in the data but without any interactions or higher order terms. As recommended by Enders [10], Barnard–Rubin degrees of freedom [5] were used when performing inferential tests for all MI conditions (using the EDF option in Proc MIANALYZE). A separate JMI condition using a prior ridged (R-JMI)[2] by 10% of $(N-1)$[3] was also included to investigate how convergence and performance was affected by the ridge prior. LD and all MI conditions estimated coefficients with ordinary least squares (OLS). One thousand replications were conducted within each cell of the simulation design and all data were generated and analyzed in SAS 9.3 using Proc IML, Proc CALIS, Proc MI, Proc MIANALYZE, and Proc Reg.

The operating Type-I error rate will be assessed by the proportion of times the null hypothesis is rejected for the null effect of β_3. Statistical power and relative bias of regression coefficients for β_1 and β_2 will also be compared across methods. Results will be compared to those obtained using the complete data as estimated with OLS prior to inducing any missing values.

Results

Type-I error rate

MAR condition

Table 1 shows the operating Type-I error rates across methods, percent missingness, and sample size for the MAR condition. Differences across the number of imputations were small for N-JMI, R-JMI, FCS-R, and FCS-P so results in Table 1 are aggregated across the number of imputation conditions. Following Bradley [8], rejection rates outside the interval [0.025, 0.075] were deemed to be beyond a nominal sampling error rate of 0.05 .

N-JMI performed very near the nominal 0.05 rate at all sample sizes when the percentage of missing data was 30% or less. When half the data were missing, the operating Type-I error rates were highly inflated, even at the largest sample sizes included in the simulation. R-JMI had acceptable rejection rates across sample size conditions with 10% missingness, but, as the percentage of missing data increased, Type-I error rates began to slowly become deflated.

Table 1. Operating Type-I error rate by method for MAR condition.

		Operating Type-I error rate						
% Missing	Sample size	Complete	FCS-R	FCS-P	LD	ML	N-JMI	R-JMI
10	20	0.052	0.059	0.064	0.053	**0.150**	0.052	0.054
	50	0.051	0.052	0.055	0.050	0.081	0.069	0.042
	100	0.053	0.061	0.063	0.039	0.055	0.042	0.049
	250	0.054	0.071	0.049	0.052	0.056	0.056	0.048
20	20	0.052	0.075	0.028	0.055	**0.320**	0.049	0.037
	50	0.051	0.073	0.036	0.055	**0.088**	0.056	0.042
	100	0.053	0.073	0.042	0.045	0.066	0.054	0.048
	250	0.054	0.069	0.039	0.044	0.048	0.040	0.047
30	20	0.052	**0.127**	**0.009**	**0.428**	**0.759**	0.068	**0.024**
	50	0.051	**0.134**	0.032	0.058	**0.184**	0.043	0.034
	100	0.053	**0.127**	0.036	0.048	**0.081**	0.056	0.040
	250	0.054	**0.119**	0.040	0.053	0.056	0.053	0.038
50	20	0.052	**0.516**	**0.000**	NA	**0.985**	**0.313**	**0.003**
	50	0.051	**0.482**	**0.005**	NA	**0.874**	**0.208**	**0.014**
	100	0.053	**0.475**	**0.015**	NA	**0.675**	**0.126**	**0.013**
	250	0.054	**0.404**	**0.020**	NA	**0.375**	**0.086**	0.036

Note: The MI conditions are collapsed over the number of imputations because differences were rather small.
Bold values indicate values that exceeded reasonable values for the nominal 0.05 rate based on criteria in Bradley [8].

Both FCS methods, FCS-R in particular, showed evidence of problematic Type-I error rates at larger percentages of missing data. With missingness of 30% or larger, FCS-R had operating Type-I error rates around 12% but were well behaved with missingness of 20% or less. FCS-P had Type-I error rates that tended towards being deflated as more data were missing and rejection rates were far too small with 50% missingness even with a sample size of 250.

With LD, rejection rates were very near the nominal rate across conditions with 10% and 20% missingness and for samples of 50 or more with 30% missingness. However, in the 50% missingness condition, no replications in any of the conditions could be estimated because no cases had complete data, exhibiting the shortcoming of utilizing LD with small samples and/or many missing values.

ML had acceptable operating Type-I error rates for samples of 50 or more with 10% missingness, but had increasingly inflated operating Type-I as the percentage of missing data increased. The rejection rates became wildly inflated with larger percentages of missing data and were seven times the nominal rate with 50% missingness with a sample size of 250.

MNAR condition

The patterns were largely similar to the results reported in the MAR condition and the values were slightly less well-behaved than in the MAR condition but quite close, so a table of results will not be reported for brevity.

Power

MAR condition

Tables 2 and 3 show power across MAR simulation conditions for β_1 and β_2, which had small and medium effect sizes, respectively. The 50% missingness condition is not shown

Table 2. Power for small effect of β_1 ($d = 0.18$) for MAR condition.

		Power Small Effect (β_1)						
% Missing	Sample size	Complete	FCS-R	FCS-P	LD	ML	N-JMI	R-JMI
10	20	0.064	0.049	0.040	0.056	**0.174**	0.058	0.052
	50	0.089	0.092	0.070	0.080	0.112	0.082	0.079
	100	0.104	0.113	0.095	0.093	0.119	0.111	0.087
	250	0.242	0.218	0.188	0.175	0.193	0.229	0.189
20	20	0.064	0.070	0.024	0.056	**0.306**	0.058	0.045
	50	0.089	0.093	0.042	0.065	**0.116**	0.066	0.056
	100	0.104	0.107	0.064	0.077	0.124	0.118	0.077
	250	0.242	0.182	0.119	0.136	0.174	0.265	0.145
30	20	0.064	**0.129**	0.010	**0.429**	**0.751**	0.091	**0.019**
	50	0.089	**0.141**	0.032	0.065	**0.202**	0.084	0.048
	100	0.104	**0.129**	0.038	0.075	**0.126**	0.129	0.057
	250	0.242	**0.162**	0.071	0.099	0.134	0.252	0.065

Note: The MI conditions are collapsed over the number of imputations because differences were rather small.
Bold values indicate values that the operating Type-I error rate was inflated.

Table 3. Power for medium effect of β_2 ($d = 0.50$) for MAR condition.

		Power Medium Effect (β_2)						
% Missing	Sample size	Complete	FCS-R	FCS-P	LD	ML	N-JMI	R-JMI
10	20	0.145	0.078	0.068	0.065	**0.301**	0.145	0.070
	50	0.325	0.203	0.168	0.168	0.414	0.341	0.276
	100	0.612	0.426	0.329	0.265	0.608	0.605	0.491
	250	0.948	0.807	0.704	0.584	0.912	0.954	0.907
20	20	0.145	0.068	0.052	0.063	**0.454**	0.123	0.034
	50	0.325	0.161	0.101	0.080	**0.398**	0.320	0.160
	100	0.612	0.304	0.214	0.109	0.572	0.595	0.417
	250	0.948	0.688	0.481	0.256	0.868	0.955	0.861
30	20	0.145	**0.097**	0.016	**0.431**	**0.802**	0.110	**0.016**
	50	0.325	**0.105**	0.066	0.068	**0.398**	0.325	0.095
	100	0.612	**0.223**	0.139	0.056	**0.502**	0.614	0.301
	250	0.948	**0.524**	0.282	0.073	0.803	0.959	0.765

Note: The MI conditions are collapsed over the number of imputations because differences were rather small.
Bold values indicate values that the operating Type-I error rate was inflated.

in Tables 2 and 3 because the operating Type-I error rates were so highly inflated that power is essentially uninterpretable across all methods. Conditions where the Type-I error rates were inflated are noted in bold, indicating that the power for these conditions is not trustworthy due to inflated rejection rates.

N-JMI performed exceptionally well and power rates mirrored the power obtained with the complete data regardless of sample size for 30% missingness or less for β_1 and β_2. Both FCS methods had poor power as the percent of missingness increased. Despite rejection rates that were about 2.5 times the nominal rate with 30% missingness, FCS-R failed to maintain power that was comparable to either the complete data or N-JMI. Power with FCS-P was quite low and the performance was closer to LD than to more modern methods, possibly due to the borderline deflated rejection rates shown in Table 1.

Unsurprisingly, power rates for LD were far lower than the complete data because observations are excluded and the sample size is drastically reduced. The reduction in power was increasingly dramatic as the percentage of missingness decreased. Power for ML exceeded power obtained from the complete data in many cases data although this is almost certainly an artifact of the inflated Type-I error rates observed in these conditions. When the Type-I error rates were well-behaved, ML power tended to be less than both the complete data and N-JMI but did exceed the power for both FCS methods. Power for R-JMI was noticeably lower than for N-JMI and became increasingly worse as the percent missingness increased.

Readers may note that power does not necessarily increase monotonically with sample size for all conditions in Tables 2 and 3 as would be expected. This is due to the poorly behaved Type-I error rates that were observed in some conditions that tended to improve as sample size increased. For example, with smaller samples, the rejection rates for ML were exceedingly high so power appeared to be quite high but this is an artifact of the rejection rates. As the rejection rates become more well-behaved, power decreased accordingly and returned to the expected monotonic pattern.

MNAR condition

Tables 4 and 5 show power across MNAR simulation conditions for β_1 and β_2, which had small and medium effect sizes, respectively. Similar to Tables 2 and 3, the 50% missingness condition is not shown because the operating Type-I error rates were so highly inflated that power is essentially uninterpretable.

Similar to the MAR conditions, N-JMI yielded power that was quite close to the values obtained by the complete data, even when 30% of the values were missing. As a general trend, power tended to be worse when data were MNAR with one interesting exception. Both FCS methods actually exhibited slightly higher power in the simulation with MNAR compared to MAR. Although this seems paradoxical based on the underlying assumptions, the unique circumstances of small sample problems are important to note. As will be discussed in the next section, based on the results presented shortly in Tables 6 and 7, the regression coefficient bias was quite large with FCS but the magnitude of the bias was actually less under MNAR compared to MAR. Because the population regression coefficients were positive in sign, the less-biased coefficients under MNAR will have larger magnitudes, on average, and will thus be more likely to be significant.

Table 4. Power for small effect of β_1 ($d = 0.18$) for MNAR condition.

		Power Small Effect (β_1)						
% Missing	Sample size	Complete	FCS-R	FCS-P	LD	ML	N-JMI	R-JMI
10	20	0.064	0.054	0.044	0.045	**0.139**	0.060	0.047
	50	0.089	0.075	0.067	0.064	0.095	0.082	0.066
	100	0.104	0.085	0.058	0.084	0.094	0.117	0.084
	250	0.242	0.182	0.162	0.177	0.183	0.265	0.139
20	20	0.064	0.079	0.036	0.057	**0.244**	0.056	0.052
	50	0.089	0.079	0.058	0.057	**0.120**	0.079	0.052
	100	0.104	0.086	0.068	0.062	**0.106**	0.131	0.069
	250	0.242	0.116	0.095	0.114	0.128	0.250	0.113
30	20	0.064	**0.157**	0.021	**0.291**	**0.534**	**0.090**	**0.035**
	50	0.089	**0.128**	0.050	0.059	**0.164**	0.072	0.043
	100	0.104	**0.124**	0.059	0.056	**0.120**	0.109	0.052
	250	0.242	**0.138**	0.072	0.073	0.115	0.235	0.075

Note: The MI conditions are collapsed over the number of imputations because differences were rather small.
Bold values indicate values that the operating Type-I error rate was inflated.

Table 5. Power for medium effect of β_2 ($d = 0.50$) for MNAR condition.

		Power Medium Effect (β_2)						
% Missing	Sample size	Complete	FCS-R	FCS-P	LD	ML	N-JMI	R-JMI
10	20	0.145	0.121	0.117	0.098	**0.225**	0.136	0.089
	50	0.325	0.294	0.191	0.191	0.353	0.316	0.324
	100	0.612	0.469	0.454	0.413	0.614	0.613	0.609
	250	0.948	0.826	0.922	0.804	0.905	0.944	0.909
20	20	0.145	0.098	0.084	0.074	**0.256**	0.111	0.093
	50	0.325	0.183	0.169	0.156	**0.368**	0.357	0.318
	100	0.612	0.334	0.250	0.263	**0.589**	0.634	0.468
	250	0.948	0.740	0.396	0.585	0.824	0.952	0.851
30	20	0.145	**0.150**	0.046	**0.292**	**0.548**	**0.136**	0.055
	50	0.325	**0.182**	0.107	0.090	**0.324**	0.341	0.228
	100	0.612	**0.325**	0.211	0.131	**0.490**	0.595	0.436
	250	0.948	**0.640**	0.323	0.287	0.784	0.949	0.744

Note: The MI conditions are collapsed over the number of imputations because differences were rather small.
Bold values indicate values that the operating Type-I error rate was inflated.

Table 6. Median relative bias for regression coefficient of β_1 ($d = 0.18$).

			Relative bias of β_1					
% Missing	Sample size		FCS-R	FCS-P	LD	ML	N-JMI	R-JMI
10	MAR	20	**−35**	**−50**	**−12**	−9	6	**−27**
		50	**−13**	**−15**	**−14**	**−20**	−7	**−21**
		100	−9	**−19**	**−17**	**−10**	−8	**−23**
		250	−9	**−18**	**−10**	**−12**	−8	**−17**
	MNAR	20	**−13**	**−24**	**−26**	2	**−13**	**−50**
		50	**−17**	**−13**	**−14**	−5	−3	**−59**
		100	**−14**	**−22**	**−12**	**−13**	−9	**−54**
		250	**−14**	−7	**13**	2	4	**−58**
20	MAR	20	**−24**	**−52**	7	**−25**	**−22**	**−13**
		50	**−17**	**−28**	−2	**−22**	−9	**−26**
		100	**−19**	**−37**	**−14**	**−21**	−9	**−30**
		250	**−13**	**−30**	−7	−8	−4	**−28**
	MNAR	20	**−26**	**−27**	**−31**	**21**	7	**−63**
		50	**−25**	**−41**	**−15**	**−21**	−6	**−50**
		100	**−17**	**−36**	**−23**	**−14**	−3	**−59**
		250	**−13**	**−24**	−5	**−15**	−3	**−54**
30	MAR	20	**−23**	**−62**	**−23**	**172**	−7	**−49**
		50	**−38**	**−63**	**−23**	**−28**	9	**−38**
		100	**−20**	**−51**	**−21**	**−17**	−3	**−35**
		250	**−20**	**−49**	**−21**	**−16**	−6	**−48**
	MNAR	20	**28**	**−25**	**73**	**250**	**16**	**−64**
		50	**−38**	**−52**	**27**	**29**	−8	**−63**
		100	**−19**	**−53**	**−16**	**−16**	−8	**−59**
		250	**−22**	**−29**	**−14**	**−27**	−3	**−63**

Note: Bold values indicate values that the bias exceeded ±10% and was considered severe.

Relative bias of regression coefficients

Tables 6 and 7 show the median relative bias across simulation conditions for β_1 and β_2, respectively. Bolded values indicate severe bias according to criteria in Hoogland and Boomsma [19] and Flora and Curran [14]. X_{2i} was not generated to have any missing values so Table 7 shows the effect that variables with missing values exhibit on a complete

Table 7. Relative bias for regression coefficient of β_2 ($d = 0.50$).

			Relative Bias of β_2					
% Missing	Sample size		FCS-R	FCS-P	LD	ML	N-JMI	R-JMI
10	MAR	20	−1	**−27**	6	−1	3	−7
		50	−2	**−20**	9	3	1	−5
		100	−1	**−17**	−3	1	1	−7
		250	−1	**−12**	2	−1	−2	−7
	MNAR	20	−5	**−14**	−5	−6	−9	**−60**
		50	−2	−8	−3	1	−3	**−54**
		100	−1	−7	−3	1	−3	**−54**
		250	−1	−4	−2	1	1	**−54**
20	MAR	20	**21**	**−34**	**12**	3	−1	**−21**
		50	**12**	**−29**	−4	9	−2	**−14**
		100	**41**	**−25**	7	6	−4	**−11**
		250	4	**−17**	8	1	−2	**−10**
	MNAR	20	−8	**−26**	**−18**	−9	**−11**	**−53**
		50	1	**−17**	3	3	3	**−50**
		100	2	**−16**	−3	1	0	**−53**
		250	3	**−11**	1	3	0	**−52**
30	MAR	20	2	**−54**	**−21**	**24**	−7	**−22**
		50	−5	**−48**	**−17**	1	2	**−21**
		100	2	**−36**	**−11**	1	1	**−15**
		250	2	**−29**	−8	−3	−1	**−14**
	MNAR	20	1	**−26**	**−65**	**78**	−1	**−57**
		50	2	**−13**	**22**	3	3	**−55**
		100	−1	**−10**	2	1	−1	**−53**
		250	−2	**−11**	−3	2	−2	**−52**

Note: Bold values indicate values that the bias exceeded ±10% and was considered severe.

variable. From Table 6, it can be seen that many different methods yield highly biased estimates of regression coefficients when data are missing and the bias tends to increase as the percent of missing data increases. FCS-P, ML, and R-JMI all have particularly poor performance and were quite biased even for relatively small amounts of missing data and samples as large as 250. N-JMI did not exhibit severe bias so long as the sample size was 50 or larger, even under the MNAR condition.

As would be expected, the values in Table 7 were much less extreme than in Table 6 because X_{2i} was complete. However, X_{2i} with FCS-P seemed to be affected by other variables having missing values. The cautions of bias when using a ridged prior were realized as well with noticeably severe bias in R-JMI that increased with MNAR and as the percentage of missingness increased. FCS-R and N-JMI were minimally affected by missing values on other variables and ML performed well when the sample size was at least 50.

Discussion and limitations

Based on the results of the simulation, N-JMI unambiguously performed best for the conditions included in the study. Type-I errors were very nearly at the nominal rate, power was essentially identical to what was obtained with the complete simulated data with up to 30% missingness, and the magnitude of the bias in the regression coefficients was below 10% for nearly all conditions, including the mild MNAR condition. With small sample sizes in general, power is a ubiquitous concern and each observation is crucial to retain. Because ML capitalizes on all observed data but works around missing values and LD removes cases with missing values, these methods handled missing data far less desirably with smaller sample sizes or as the percentage of missing data increased. Because ML and LD do not rectangularize the data, larger percentages of missing data, in essence, reduce the sample size; MI's ability to rectangularize the data by filling in the holes allows the maximal amount of information to be retained in the subsequent analyses.

Despite the fact that the fully conditional specification also retangularizes the data, FCS-R and FCS-P also tended to perform rather poorly with smaller samples. FCS-R exhibited a fair amount of power but with inflated Type-I error rates. FCS-P estimated coefficients with large amounts of bias, presumably because the number of appropriate potential donor observations is reduced with small samples. With FCS-P, a common recommendation is to select the five potential donor cases that most closely match on the relevant covariates of the individual with a missing value. Then, one of the five donor cases is randomly selected and its value used for imputation. With only 20 or 50 individuals, however, the closest five individuals may not be very similar at all and the bias seen in the coefficient estimates seemed to reflect this.

Though not reported previously, convergence with N-JMI was only problematic in the 20 individual conditions in this study with non-convergence rates between 10% and 15%, depending on condition. R-JMI was useful in aiding model convergence in these conditions and had perfect convergence. However, N-JMI otherwise vastly outperformed R-JMI in scenarios in which N-JMI was able to converge. To optimize performance, R-JMI is therefore only recommended with small samples when model convergence is problematic and the percentage of missing values is rather low, otherwise estimates may be subjected to substantial bias and can be improved upon with N-JMI.

Additionally, for the conditions in this study, augmenting the number of imputations did not have a discernable effect of power, which has been similarly observed in small to moderate sample size conditions (though not extending as small as included in the present study) in studies focusing on the number of imputations to perform [15]. However, given the secondary focus of the number of imputations in this study and the more extensive nature of previous studies on the topic, performing 20 or more imputations is recommended for most straightforward models when the increase in computational time is negligible because performing more imputations does not adversely affect estimation provided that the computational time is reasonable.

In applied settings, the differential performance of the N-JMI and other methods for missing data may not be as extensive as observed in the simulation study. Although MI does have the advantage of retaining all cases which is highly desirable with small samples and N-JMI performed best among competing MI methods, the process of imputing values is not always straightforward and slight changes to the imputation model can affect results. JMI assumes multivariate normality which was upheld in this study but may be tenuous in applied research. Imputing for interactions and higher order terms and also properly centering variables can also be somewhat challenging with MI methods, particularly when attempting to specify the imputation model [10,13]. Failing to adequately account for the complexity these situations introduce into the imputation model can attenuate estimates so that estimates are biased towards 0, even if data are MCAR [10]. In methods that do not impute data such as ML, there is no imputation model so these concerns are less relevant (although users of ML may need to consider auxiliary variables to ensure that the mechanism leading to missing values is included in the model). Additionally, as partially evidenced by the R-JMI results, prior distributions have a greater influence on MCMC posterior distributions with smaller sample sizes.

Nonetheless, despite some additional modeling hurdles, for the conditions in this study N-JMI gives researchers the greatest *ability* to maximize power and be confident that Type-I error rates are performing at the nominal rate with small samples because it results in a completed dataset. Granted, properly specifying the imputation model is no small task and modeling small sample data almost always involves tradeoffs. The price paid for carefully considering the imputation model could be rewarded with estimates that are essentially identical to what would have been obtained had all the data been collected. Alternative methods such as LD or ML may be more appealing due to their less-intensive manner for handling missing values, but, with small samples in particular, one pays for convenience with far less desirable performance.

As limitations of this study, first the imputation model included all appropriate variables. In applied data analytic scenarios, this may be quite difficult to accomplish in most scenarios. Second, data were generated such that the multivariate normality assumption inherent with many missing data methods was upheld which is assumed when using ML and JMI. In applied settings with small samples, multivariate normality may be tenuous and the ability to test the tenability of this assumption with small samples is often rather difficult [33]. Second, the data generation model was admittedly and purposefully rather tame. The literature on the performance of methods to handle arbitrary missingness with small samples is virtually non-existent, so the rather simple model allowed for a baseline empirical comparison within a context where the methods have been well-studied and in which few questions of their utility under ideal conditions exist. Even with a fairly simple model, the results remain informative for researchers facing this analytic challenge with real data [see 15 for an oft-cited paper employing an even simpler model] as the results were rather clear-cut. Third, the single-level nature of the model may not be adequate for the common scenarios in disciplines such as behavioral sciences where data are often clustered within higher level units (e.g. suitable for multilevel modeling); future research could extend small sample investigations to the multilevel setting where the options to handle missing data are more diffuse and where performance and recommendations for best practice are less well-defined in the methodological literature.

Notes

1. Missing values for X_{3i} were induced prior to creating missing values for X_{1i}.
2. Hereafter, N-JMI refers to multiple imputation with a non-ridged prior, R-JMI refers to multiple imputation with a ridged prior, and JMI refers to multiple imputation in a broad sense.
3. Honaker *et al.* [19] have recommended 10% as a reasonable upper bound for the ridge prior proportion (p. 20).

References

[1] P.D. Allison, *Missing data techniques for structural equation modeling*, J. Abnor. Psychol. 112 (2003), pp. 545–557.

[2] G. Ambler, R.Z. Omar, and P. Royston, *A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome*, Stat. Methods Med. Res. 16 (2007), pp. 277–298.

[3] J.L. Arbuckle, *Full information estimation in the presence of incomplete data*, in *Advanced Structural Equation Modeling: Issues and Techniques*, G.A. Marcoulides and R.E. Schumacker, eds., Erlbuam, Mahwah, NJ, 1996, pp. 243–277.

[4] M.J. Azur, E.A. Stuart, C. Frangakis, and P.J. Leaf, *Multiple imputation by chained equations: What is it and how does it work?*, Int. J. Methods Psych. Res. 20 (2011), pp. 40–49.

[5] J. Barnard and D.B. Rubin, *Small-sample degrees of freedom with multiple imputation*, Biometrika 86 (1999), pp. 948–955.

[6] S.A. Barnes, S.R. Lindborg, and J.W. Seaman, *Multiple imputation techniques in small sample clinical trials*, Stat. Med. 25 (2006), pp. 233–245.

[7] J.W. Bartlett, J.R. Carpenter, K. Tilling, and S. Vansteelandt, *Improving upon the efficiency of complete case analysis when covariates are MNAR*, Biostatistics 15 (2014), pp. 719–730.

[8] J.V. Bradley, *Robustness?*, Br. J. Math. Stat. Psychol. 31 (1978), pp. 144–152.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Stat. Soc. Ser. B. Stat. Methodol. 39 (1977), pp. 1–38.

[10] C.K. Enders, *Applied Missing Data Analysis*, Guilford Press, New York, 2010.

[11] C.K. Enders, *A primer on maximum likelihood algorithms available for use with missing data*, Struct. Eqn. Model. 8 (2001), pp. 128–141.

[12] C.K. Enders and D.L. Bandalos, *The relative performance of full information maximum likelihood estimation for missing data in structural equation models*, Struct. Eqn. Model. 8 (2001), pp. 430–457.

[13] C.K. Enders and A.C. Gottschall, *Multiple imputation strategies for multiple group structural equation models*, Struct. Eqn. Model. 18 (2011), pp. 35–54.

[14] D.B. Flora and P.J. Curran, *An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data*, Psychol. Methods 9 (2004), pp. 466–491.

[15] J.W. Graham, A.E. Olchowski, and T.D. Gilreath, *How many imputations are really needed? Some practical clarifications of multiple imputation theory*, Prev. Sci. 8 (2007), pp. 206–213.

[16] J.W. Graham and J.L. Schafer, *On the performance of multiple imputation for multivariate data with small sample size*, in *Statistical Strategies for Small Sample Research*, R. Hoyle, ed., Sage, Thousand Oaks, CA, 1999, pp. 1–29.

[17] D.F. Heitjan and R.J. Little, *Multiple imputation for the fatal accident reporting system*, J. R. Stat. Soc. Ser. C. Appl. Stat. 40 (1991), pp. 13–29.

[18] J. Honaker, G. King, and M. Blackwell, *Amelia II: A program for missing data*, J. Stat. Software 45 (2011), pp. 1–47.

[19] J.J. Hoogland and A. Boomsma, *Robustness studies in covariance structure modeling: An overview and a meta-analysis*, Sociol. Methods Res. 26 (1998), pp. 329–367.

[20] V.L. Kristman, M. Manno, and P. Côté, *Methods to account for attrition in longitudinal data: Do they work? A simulation study*, Eur. J. Epidemiol. 20 (2005), pp. 657–662.

[21] K.J. Lee and J.B. Carlin, *Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation*, Amer. J. Epidemiol. 171 (2010), pp. 624–632.

[22] R.J. Little, *Regression with missing X's: A review*, J. Am. Statist. Assoc. 87 (1992), pp. 1227–1237.

[23] G. Liu and A.L. Gould, *Comparison of alternative strategies for analysis of longitudinal trials with dropouts*, J. Biopharm. Statist. 12 (2002), pp. 207–226.

[24] D.M. McNeish and L.M. Stapleton, *The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration*, Education. Psychol. Rev. 2014, doi:10.1007/s10648-014-9287-x

[25] J.L. Peugh and C.K. Enders, *Missing data in educational research: A review of reporting practices and suggestions for improvement*, Rev. Education. Res. 74 (2004), pp. 525–556.

[26] D.B. Rubin, *Inference and missing data*, Biometrika 63 (1976), pp. 581–592.

[27] D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, 1987.

[28] V. Savalei, *Small sample statistics for incomplete nonnormal data: Extensions of complete data formulae and a Monte Carlo comparison*, Struct. Eqn. Model. 17 (2010), pp. 241–264.

[29] J.L. Schafer, *Analysis of Incomplete Multivariate Data*, CRC Press, Boca Raton, FL, 2010.

[30] J.L. Schafer and J.W. Graham, *Missing data: Our view of the state of the art*, Psych. Methods 7 (2002), pp. 147–177.

[31] S. Seaman, J. Galati, D. Jackson, and J. Carlin, *What Is Meant by 'Missing at Random'?*, Statis. Sci. 28 (2013), pp. 257–268.

[32] J.A. Sterne, I.R. White, J.B. Carlin, M. Spratt, P. Royston, M.G. Kenward, A.M. Wood, and J.R. Carpenter, *Multipleimputation for missing data in epidemiological and clinical research: Potential and pitfalls*, BMJ 338 (2009), p. b2393.

[33] M. Tan, H.B. Fang, G.L. Tian, and G. Wei, *Testing multivariate normality in incomplete data of small sample size*, J. Multivariate Anal. 93 (2005), pp. 164–179.

[34] S. Van Buuren, *Multiple imputation of discrete and continuous data by fully conditional specification*, Stat. Methods Med. Res. 16 (2007), pp. 219–242.

[35] I.R. White, P. Royston, and A.M. Wood, *Multiple imputation using chained equations: Issues and guidance for practice*, Stat. Med. 30 (2011), pp. 377–399.