MEDDOCS Open Access Publisher

Journal of Psychiatry and Behavioral Sciences

Open Access | Research Article

A Comparison of the Heckman Selection Model, Ibrahim, and Lipsitz Methods for Dealing with Nonignorable Missing Data

*Corresponding Author(s): Holmes Finch W

Department of Educational Psychology, Ball State University, Indiana.

Email: whfinch@bsu.edu

Received: Dec 01, 2020

Accepted: Jan 19, 2021

Published Online: Jan 25, 2021

Journal: Journal of Psychiatry and Behavioral Sciences

Publisher: MedDocs Publishers LLC

Online edition: http://meddocsonline.org/

Abstract

The presence of missing data is a persistent issue that researchers, data analysts, and statisticians must address in their work. As a result, a number of methods have been developed for dealing with it, from simple single imputation schemes, to complex models based on multiply imputed data. A common assumption among most of these methods is that the information necessary to impute the data under a presumption of randomness is available. Given that this assumption holds, several approaches for imputing missing values, including Multiple Imputation by Chained Equations (MICE) have been shown to be very effective for dealing with missing values. However, when this assumption does not hold, MICE and other imputation methods are not so effective, and may produce imputed data that yields biased results. In order to address this issue, several methods have been developed for the case where data are Missing not at Random (MNAR). This simulation study compared two approaches for handling MNAR data for a dichotomous variable in the context of logistic regression. The results show that an approach based on the Heckman selection model, and adapted for the MICE context, as well as an adapted Ibrahim Lipsitz method performed well, producing data that results in relatively low estimation bias and high confidence interval coverage rates. Implications of these findings are discussed.

Introduction

Missing data are a ubiquitous problem in virtually all areas of research, from the social sciences, to health care, economics, and biology. The presence of missing data can have deleterious impacts on statistical procedures in the form of parameter estimation bias, inflated standard errors, and inaccurate hypothesis test results [1]. Given its ubiquity, as well as the potential negative impacts that it can have on statistical analyses, researchers must make decisions regarding how to deal with it. Such decisions need to account for both the type and amount of missing data, and typically include some form of imputation, or replacement of the missing values with a reasonable approximation of what the value would have been were it not missing. This process is complicated by the fact that the presence of missing data may have a myriad of causes, and that different methods for handling it are more appropriate in different situations. The purpose of this manuscript was to investigate and compare the performance of two approaches for dealing with missing data in perhaps the most difficult situation, where the data are missing not at random. The manuscript is ordered as follows. First is a brief description of the various types of missing data, after which the methods used to deal with the missing data are described. Next, prior research examining the performance of

Cite this article: Holmes FW. A Comparison of the Heckman Selection Model, Ibrahim, and Lipsitz Methods for Dealing with Nonignorable Missing Data. J Psychiatry Behav Sci. 2021; 4(1): 1045.

these methods is discussed, and the goals of the current study are then outlined in light of this prior work. The Monte Carlo simulation methodology is then described, followed by the results of the study, and then a discussion of these results, and their implications for practice.

Types of missing data

Missing data are typically described as coming from one of three sources: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random (MNAR). MCAR occurs when there is no systematic cause to a data value being missing. For example, an MCAR item response was left blank by the respondent completely by accident. With MAR, the variable associated with the missing value has been measured by the researcher. For example, if males are more likely to leave an item on a survey unanswered, and the researcher has collected data on the gender of the respondents, then the missing values would be considered MAR. Finally, MNAR occurs when the missing data are directly linked to the missing value itself. In other words, there is no information in the dataset regarding the missing data mechanism (unlike with MAR), nor are the values missing due solely to randomness (as with MCAR). MNAR data might occur if a respondent taking a survey leaves an item unanswered because they did not want to reveal their actual answer. MNAR data, and methods for dealing with it are the focus of the current study, specifically in the context of logistic regression for a binary response variable. There exist a number of methods for dealing with missing data. These approaches range from only using observations for which all of the relevant variables have values present, to complex model based approaches for estimating what the missing values would have been. One of the most common missing data methods is Listwise Deletion (LD), where all cases that have missing data are removed from the dataset, and then data analyses are conducted on this subset of complete data. Numerous studies have demonstrated that LD is not an optimal approach for dealing with missing data in most situations, as it leads to biased parameter estimates, low power, and inflated standard errors in many situations [1]. In contrast to LD, as well as simple techniques such as mean or regression imputation, more complex approaches based on multiply imputing missing values have been found to work well for MCAR and MAR data [1-3]. Examples of these approaches include full information maximum likelihood [4], multiple imputation using joint models [5], multivariate imputation by chained equations [6], and predictive means matching [7]. Data imputation techniques work by replacing missing values with estimates of what the values might have been based upon information from the non-missing observations of the variable itself, as well as from measured variables in the dataset. For example, missing values to a survey question asking respondent income might be accurately imputed if we know their age, level of education, and employment status. Multiple imputation methods work by creating multiple datasets containing such imputed values with some error added in, to reflect our uncertainty regarding the actual value. Several studies have supported the utility of the chained equations approach to imputation in a wide variety of situations [8], and for that reason it will be discussed in more detail below.

Multivariate Imputation by Chained Equations

Multivariate Imputation by Chained Equations (MICE), also known as fully conditional specification or sequential regression multiple imputation, is a multiple imputation technique that has been shown to be quite useful in practice [6,9]. MICE does not make any assumptions regarding the joint probability distribution of the set of variables for which imputation must be done, but rather works under the assumption that each variable in the dataset has its own unique marginal distribution [6]. The specific distribution to be used for a variable with missing data to be imputed is selected so to be appropriate to its type (i.e., normal, binomial, multinomial, Poisson). For example, if the variable is binary, a logistic regression model will be used, whereas if the variable is continuous the missing data is modeled based on ordinary least squares regression. MICE does work under the assumption that missing data are MAR. It has been shown to yield data resulting in biased parameter estimates when data are MNAR [10]. MICE functions under a six-step process, as described by [10].

1. Replace each missing value in the dataset with a simple imputation, such as a random draw from the sample.
2. These place holder imputations are then set back to missing for a target variable, *y*.
3. The observed values for *y* are regressed on the other variables in the data (or any other variables the researcher would like to use for imputation), using an appropriate model (e.g. binary logistic regression for dichotomous variables.
4. Missing values for *y* are replaced by random draws from the probability distribution implied by the model fit in step 3, using the Gibbs sampler. In other words, imputations for the target variable y_1^{t+1} are drawn from the probability distribution defined by the regression model as $P(y_1 \mid X_2^t, X_3^t, ..., X_k^t)$, where the X^t are the other (nontarget) variables in the dataset used to fit the model, and the imputed value of *y* is drawn from the appropriate (e.g. normal) probability distribution conditioned on the predictors in the regression model.
5. Steps 2-4 are repeated for each variable in the dataset containing missing data. Completing the initial imputations for all variables in the dataset constitutes a single iteration.
6. Steps 2-5 are then repeated for a given number of iterations (e.g. 40) until the imputations have converged over the iterations.
7. The entire set of steps is repeated *m* times, where *m* is the desired number of multiple imputations.

Once the imputation process is completed, the analysis of interest (e.g. regression) is used with each of the *m* imputed datasets and the parameter estimates are combined using [5] rules. Given the research supporting its use in a wide variety of contexts [11-14], particularly when data are MCAR or MAR, MICE will be included in the current study.

Missing data methods designed specifically for MNAR data

Ibrahim and lipsitz method

Described an approach for addressing the case when missing values are present for a dichotomous response variable, and the missing data mechanism is MNAR [15]. This technique, which will be referred to as IL going forward in this manuscript, was developed specifically to address bias that is known to be present when standard imputation approaches are applied to MNAR data [15]. The underlying framework of the IL approach

to handling missing data involves the simultaneous estimation of coefficients for the standard logistic regression model of interest, in conjunction with an additional logistic regression model in which the response variable reflects whether the original dependent variable is missing for an individual. More formally, consider a scenario in which there is a dichotomous response variable, *y*, which can be linked to a matrix of covariates, **x**, through a logistic regression (LR) model of the form:

$$p(y_i = 1 \mid x_i\beta) = \frac{e^{(x_i\beta)}}{1+e^{(x_i\beta)}} \qquad (1)$$

Where

y_i =Response variable for subject i

x_i =Matrix of covariates for subject i

β =Matrix of coefficients linking covariates to response

In addition to this standard LR model, the IL approach also involves the fitting of a second LR model in which the dependent variable is whether subject *i* has a missing value for variable y_i.

$$p(R_i = 1 \mid z_i\alpha) = \frac{e^{(z_i\alpha)}}{1+e^{(z_i\alpha)}} \qquad (2)$$

Where

R_i =if y_i is missing, and 1 otherwise

z_i =matrix of covariates for R_i including both x_i and y_i

α =Coefficients linking z_i to R_i

Together, α and β are defined as Y, and a joint log-likelihood for the data can be written as

$$l(Y \mid x_i y_i z_i r) = \log\left\{\Pi_{i=1}^{n} p(y_i \mid \beta_i x_i)\, p(r_i \mid \alpha_i z_i)\right\} \qquad (3)$$

The parameter set Y can then be estimated using the EM algorithm. Finally, it should be noted that when $\boldsymbol{\alpha}$ is a null vector, the data are MCAR. One issue that was not addressed in the original description and implementation of IL was bias associated with maximum likelihood estimation of the score function when samples are small [16,17], introduced an adjustment to the score function in LR involving the multiplication of the likelihood by a Jeffreys prior. Maity, Pradhan, and Das subsequently applied this Firth correction to the IL method, yielding the following joint log-likelihood

$$l^*(Y \mid x_i y_i z_i r) = l(\beta \mid x_i y) + \frac{1}{2}\log|l(\beta)| + l(\alpha \mid z_i r) + \frac{1}{2}\log|l(\alpha)| \qquad (4)$$

Where

$|l(\beta)|$ =Determinant of the observed information matrix for β

$|l(\alpha)|$ =Determinant of the observed information matrix for α

The likelihood function $l^*(Y \mid x_i y_i z_i r)$ can then be solved using the EM algorithm, with the Firth correction for sample size incorporated into the likelihood function. This approach to dealing with MNAR data will be referred to as FIL henceforth in this manuscript.

Heckman imputation

In addition to IL/FIL, an alternative approach for handling MNAR data is based on work by [18] in the form of a selection model. Initially, the selection model was formulated in the context of continuous data, where a standard regression model was applied to the response variable of interest, and a probit model was applied to the missing data mechanism. Subsequently, [19] extended the Heckman selection model to the case where the dependent variable is dichotomous, and thus ordinary least squares regression is not appropriate. In this context, the dependent variable of interest, *y*, can be modeled using the probit link function as

$$p(y_i = 1 \mid x_i) = \Phi(x_i\beta) \qquad (5)$$

Where

Φ =Standard normal cumulative distribution function.

All other model terms are as defined above for equation (1).

In turn, the missing data mechanism can also be expressed in terms of the probit link as

$$p(Ry_i = 1 \mid x_i^s = \Phi(x_i^s\beta^s) \qquad (6)$$

Where

Ry =Indicator of missingness for y_i (1 if observed, and 0 if missing).

x_i^s =Observed covariates that are potentially associated with the missingness mechanism.

β^s =Coefficients for the probit model linking missing status with x_i^s.

For the models in equations (5) and (6), it is assumed that there is a latent normally distributed variable associated with the observed realization of the dichotomous variable. These relationships take the following forms.

y^t =Latent normally distributed variable associated with *y*, where if $y_i^t > 0$ $y_i = 1$, otherwise $y_i = 0$.

R_{yi}^t =Latent normally distributed variable associated with Ry_i, where if $R_{yi}^t > 0$ $R_{yi} = 1$, otherwise $R_{yi} = 0$.

These latent variables can then be related to the observed covariates in equations (5) and (6) through standard linear regression models, including separate error terms for each (ε_i' and ε_i).

$$R_{yi}^t = (x_i^s\beta^s) + \varepsilon_i^t \qquad (7)$$

$$y_i^t = (x_i\beta) + \varepsilon_i \qquad (8)$$

The error terms in models (7) and (8) are assumed to come from the standard normal distribution, and to have a correlation between them of ρ. When $\rho = 0$, the missing data mechanism is MAR, as the errors for the observed variable and the missingness indicator are independent. On the other hand, when $\rho \neq 0$ the missing data mechanism is MNAR, with larger values of this mechanism being associated with a stronger degree of MNAR. The parameters in models (7) and (8) are $\theta = (\beta, \beta^s, \rho)$, and are instrumental in the imputation of missing values. The selection model algorithm for a binary outcome variable that was proposed by [20] involves the following steps.

Estimate the model parameters θ and ψ, where ψ is the variance-covariance matrix of $\hat{\theta}$.

Draw θ^* from $N(\theta, \psi,)$

Draw the imputation, y_i^* from a Bernoulli dstiribution with

parameter p_i^* from

$$p_i^* = \frac{\Phi_2(X_i\beta^* - X_i^{s*}, \rho^*)}{\Phi(-X_i^s\beta^{s*})} \quad (9).$$

Finally, [20] incorporated their adaptation of the Heckman imputation procedure into the MICE algorithm described above. Specifically, the algorithm outlined above is applied to the dependent variable, which is assumed to be MNAR, and standard MICE imputation is applied to the independent variables in the LR model. Thus, regression models were applied to continuous covariates with missing values, and LR to binary covariates containing missing data. This is the approach that was used in the current study, and which will be referred to as MICE_MNAR henceforth.

Prior research into the performance of MNAR missing data methods

There has been some prior research conducted to examine the performance of these methods for dealing with MNAR data in the context of LR. For example, [16] examined the performance of IL and FIL using a Monte Carlo simulation study. Their focus was primarily on the performance of these approaches with small samples, given that the purpose behind the development of FIL was to address bias in IL when the sample size is small. Their simulation study included sample sizes between 30 and 150, a LR model with 4 normally distributed covariates, each of which had a model coefficient of 1. The MNAR missing data indicator was generated using the model in equation (2), with α for *y* set to a non-zero value, which was manipulated to ensure that approximately 25% of the response variable values were missing for each sample size condition. Results of the simulation study demonstrated that across the sample size conditions, FIL consistently displayed lower estimation bias, smaller standard errors, and higher confidence interval coverage rates than did IL. In nearly all of the simulated conditions, the coverage rates for the model parameters were very close to the nominal 0.95 level for FIL, whereas for IL they were generally below the nominal level. Also conducted a Monte Carlo simulation study to investigate the performance of the MICE_MNAR method with a dichotomous outcome variable [20]. As with [16], the independent variables in the simulation were generated from the standard normal distribution with coefficients of 1. MNAR missing data were generated using equations (7) and (8), with coefficients set to 1, -0.5, and 1 for equation (8). The degree of MNAR data present was determined by the correlation between the error terms in the two equations, with ρ = 0,0.3 or 0.6, representing MAR, light MNAR, and heavy MNAR data. A sample size of 500 was used for all simulation conditions in the study. In addition to simulating the data using the Heckman model, as described above, the authors also simulated data using a LR model in which the missing data indicator was the response, and the original covariates, as well as the dependent variable of interest were predictors. In order to simulate MAR, light, and heavy MNAR data, the coefficient for *y* in this model was set to 0, 1, or 2. This latter set of simulations was designed to assess the performance of MICE_MNAR when the process underlying the missing data did not conform to the Heckman model underlying it. Across study conditions, 30% of the data were made to be missing. Results of this simulation demonstrated that the relative bias for MICE_MNAR was typically less than 2.5% when the missing data mechanism for the response was MNAR. In contrast, the comparison method, Listwise Deletion (LW), exhibited much higher rates of bias for MNAR data. In addition, the empirical standard errors for the MNAR data were lower than those for LW. The authors concluded that for MNAR data, MICE_MNAR outperformed LW, regardless of the way in which it was simulated. When data were MAR, MICE_MNAR had larger standard errors than LW, but its bias was lower.

Study goals

The primary goal of this study was to further research into the performance of imputation methods for cases when data were MNAR in the context of logistic regression with a binary dependent variable. As described above, there exist multiple approaches for handling missing data in these cases, including those from the IL/FIL and selection model/Heckman paradigms. Prior research has outlined situations in which each method has performed well. However, this work has not directly compared these methods with one another using a Monte Carlo simulation methodology. Thus, one of the primary goals of this study was to compare these two promising techniques with one another using the same set of conditions. In addition, prior work examining both of these methods has focused primarily on their performance when the data are MNAR, though some work with MICE_MNAR and MAR data was done, as described above. In practice, researchers will not know the precise mechanism underlying missing values in their dataset. Therefore, it is important to thoroughly investigate how these methods perform when applied to datasets with an underlying MAR mechanism. A third goal of this study was to further prior research into the performance of these methods by including a wider array of simulation conditions than has been used previously, particularly with respect to sample size, the percent of missing values, non-MNAR specific imputation techniques, and the strength of the MAR process when data were generated as MAR. Finally, prior research has demonstrated that using MICE imputation in the standard manner leads to parameter estimation bias when the data are MNAR [10]. Thus, it was anticipated that IL/FIL and MICE_MNAR would perform better than MICE in terms of parameter estimation accuracy, when the missing data mechanism was MNAR.

Methods

In order to address the research goals outlined above, a Monte Carlo simulation study design was used. For every combination of conditions, which are described below, 1000 replications were generated. All simulations were conducted using the R software package, version 4.0 [21]. The data were generated for a binary logistic regression model with three independent variables, each of which was drawn from the standard normal distribution. The dependent variable was generated from a LR model taking the form

$$p(y_i = 1) = \log it(\beta_0^* + \beta_1^* X_1 - \beta_2^* X_2 + \beta_3^* X_3) \quad (10)$$

Where

$$\beta_0^* = 1$$

$$\beta_1^* = 1$$

$$\beta_2^* = 0.5$$

$$\beta_3^* = 0.75$$

Values of $p(y_i = 1)$ that exceeded 0.5 were set to 1, and those of 0.5 or less were set to 0.

Sample size

Data were generated with sample sizes of 50, 100, 200, 400, and 800. These values were selected to represent a range of sample sizes that would be viewed in practice, and which cover sample sizes used in prior research in this area [16,20,22].

Percent missing data

The conditions for the percent of missing data were 10%, 20%, 30%, and 40%. As with the sample size conditions, these values were chosen to represent conditions that might be seen in practice, and because they cover a range of values used in prior missing data research [16,20,22-24].

Type of missing data

Data for the dependent variable were generated from either a MNAR or a MAR process. Prior research examining the performance of the methods that are the focus of this study has emphasized their performance when the outcome variable of interest had missing data that were generated as MNAR. This focus is reasonable, given that they were designed for use in such cases. However, in practice, researchers will not actually know the underlying missing data mechanism when they select the methods to use for dealing with it. Therefore, it is important to know how these methods will perform when the underlying missing data mechanism is not MNAR, but the researcher assumes that it is and applies one of the methods designed for that type of missingness. In order to provide some information about such cases, missing values for the response variable were generated separately for the MNAR and MAR cases. MAR missing data was generated using the ampute function in the R mice library [9]. In order to manipulate the strength of the relationship between the measured variables and the missing values, the coefficients provided to ampute were manipulated to be either 0.5 or 1.0, representing a relatively weaker versus stronger MAR missing data mechanism. As was the case for MAR, MNAR data were generated using the ampute function in R. In keeping with [20], light, and heavy MNAR data were simulated, in this case using ampute weight values for *y* of 1 or 2. This approach to simulating MNAR data was selected to ensure that the results of the simulation did not unrealistically favor MICE_MNAR were the data to have been generated using a selection model.

Missing covariate data

Prior work with the Heckman approach [20] has examined cases in which data are missing both for the response variable as well as the covariates. However, for the IL and FIL methods, more work needs to be done in this regard. In addition, not a great deal of prior work has examined how the strength of the MAR effect is associated with parameter estimation in the context of missing covariate data. Therefore, the strength of the MAR missing data mechanism was manipulated through the weights applied in the missing data generation. The weights used in this study were 0.5 and 1.0, representing relatively weaker and stronger MAR data generation for the missing covariates. The percentages of missing data for the covariates matched those for the response variable, which were listed above.

Methods for handling missing data

The focus of this research was on comparing methods designed to deal with MNAR data. Therefore, the IL, FIL, and MICE_MNAR approaches were included. The former two methods were carried out using the il and fil functions from the brlmr R library [16]. The logit link function was used, given that the dependent variable dichotomous in nature. Otherwise, default settings were applied. The Heckman model approach based on MICE was carried out using the miceMNAR R package [19]. The selection and output equations both included the set of predictor variables. In addition, LW and MICE were also used in this simulation study, to represent common and/or exemplar methods for dealing with missing data that is MAR or MCAR. For MICE and MICE_MNAR, a total of 50 imputed datasets were used.

Outcome variables

Several outcome variables were included in this study. The Absolute Relative Bias (ARB) for each coefficient was calculated as

$$ARB = \frac{|\beta - \beta|}{\beta} \qquad (11)$$

Where

β = Estimated parameter value

β = Data generating parameter value

In addition, the empirical standard error for each estimate was calculated as the standard deviation of β across replications. Finally, for each estimate, the 95% confidence interval for each replication was obtained. The proportion of replications for which the data generating parameter value appeared in the 95% confidence interval was the coverage rate for these intervals, and served as an additional outcome variable for the simulation study. In order to ascertain which of the manipulated study factors, or their interactions, were associated with these outcomes, Analysis of Variance (ANOVA) was used for each. The dependent variables for these ANOVA models were ARB, empirical standard error, and coverage rates, respectively. The independent model terms were the main effects of the manipulated study factors outlined above, and their interactions. In addition to the statistical significance of each, the partial η^2 value was also calculated for each term. Both the hypothesis test results and effect sizes are reported below.

Results

MNAR data

The ANOVA results with respect to ARB indicated that the interaction of missing data method by MNAR level by percent missing by covariate missing status was significantly related to the parameter estimation bias ($F_{12,188}$=3.19, p=0.0229, η^2 =0.17). Parameter estimation bias for the factors in this interaction term appear in (Table 1). Across conditions in (Table 1), coefficient estimates obtained using LW deletion yielded the greatest ARB, followed by those based on standard MICE imputed data. Among the imputation methods designed specifically for the MNAR case, FIL and MICE_MNAR were associated with the least biased estimates. When the covariates had missing data as well as the response variable, estimation bias for all of the methods was slightly higher. This result was particularly notable with standard MICE, with less of an increase in relative bias for MICE_MNAR, IL, or FIL. Finally, when the covariates did not have missing data, estimates using the MICE_MNAR imputed data were the least biased, whereas when the covariates also had missing data, FIL and MICE_MNAR were associated with comparable levels of relative bias, which was the lowest across methods.

With regard to coverage rates, the ANOVA results identified the interactions of missing data method by covariate missing data status by sample size ($F_{16,188}$=3.57, p=0.0058, η^2=0.19), and missing data method by percent of missing data ($F_{12,188}$=20.97, p<0.0001, η^2=0.57) as statistically significant. The coverage rates by method, covariate missing data status, and sample size appear in (Table 2). Across methods, the coverage rates for the LW data were well below the nominal 0.95 rate. The other methods all exhibited very high coverage rates, uniformly 0.95 or higher, except for a few instances involving MICE_MNAR. More specifically, the MICE_MNAR coverage rates were below 0.95 when the covariates had missing data and the sample size was 400 or larger. In addition, FIL also had a coverage rate below 0.95 when the covariates had missing data, and the sample size was 800. The coverage rates by missing data method by percent of missing data in the sample (Figure 1). These results reinforce the finding, discussed above, that estimates based on LW data were had coverage rates of approximately 0.85, well below the nominal 0.95 level. The coverage rates for the other methods were generally around the nominal rate, across percentages of missingness. ANOVA results for the standard error identified the interaction of missing data method and sample size as statistically significantly related to the value of the parameter estimate standard errors ($F_{16,268}$=4.96, p<0.0001, η^2=0.23). For all of the missing data methods, the standard errors declined in value concomitantly with increases in sample size (Figure 2). With regard to methods, the standard errors were smallest for MICE_MNAR and FIL across sample sizes, and were largest for the LW data, followed by standard MICE. Finally, for N=800, the standard errors of the estimates were comparable for IL, MICE, and LW data.

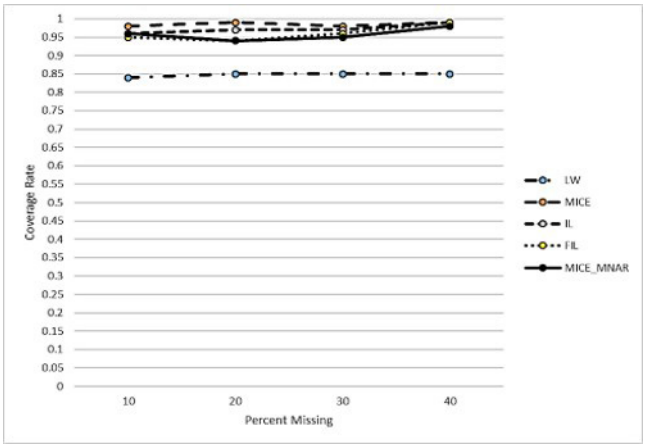

Figure 1: Coverage rates by missing data method and percent missing: MNAR data.

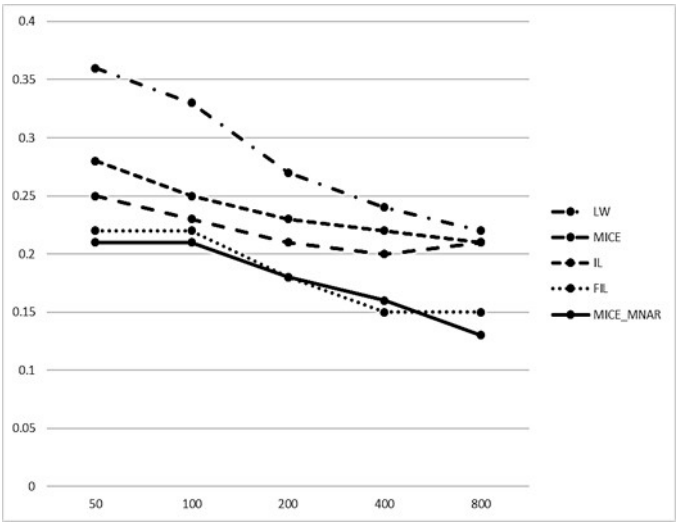

Figure 2: Standard errors by missing data method and sample size: MNAR data.

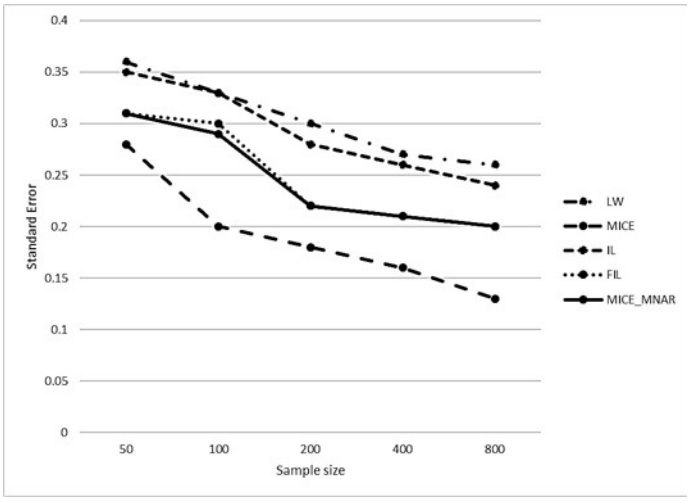

Figure 3: Standard errors by missing data method and sample size: MAR data.

MAR data

When the dependent variable were generated with a MAR process, ANOVA results identified the interaction of missing data method by the MAR weights was significantly related to the relative estimation bias ($F_{4,268}$=54.88, p<0.0001, η^2=0.44). No other terms were found to be statistically significant. The relative estimation bias for the logistic regression by the missing data method and the MAR weights (Table 3). The relative bias results for LW were greater when the MAR weights were larger, whereas for the other methods, there was not a difference in bias across these weights. Additionally, LW had the highest relative bias values across conditions. The lowest estimation bias was associated with the IL missing data approach, with MICE_MNAR and MICE exhibiting similar levels of ARB, and FIL having slightly higher values. For the parameter coverage rates when the missing data were generated as MAR, the interaction between missing data method by MAR weights by percent of missing data was statistically significant ($F_{12,268}$=7.43, p=0.0001, η^2=0.22). LW consistently had the lowest coverage rates across conditions, with values below the nominal rate of 0.95 (Table 4). In contrast, the coverage rates associated with MICE, IL, FIL, and MICE_MNAR were all at or above the 0.95 rate. The ANVOA results indicated that with respect to the standard errors for the MAR data case, the interaction between missing data method and sample size was statistically significant. These standard errors appear in (Figure 3). For all missing data methods, the standard errors declined with increases in sample size. The lowest standard errors were consistently associated with the standard MICE missing data approach, with FIL and MICE_MNAR having comparable standard errors, and those of IL being just below those of LW.

Table 1: Relative estimation bias by missing data method, covariates missing status, MNAR level, and percent missing: MNAR data.

Covariates not missing						
MNAR level	Missing Percent	LW	MICE	IL	FIL	MICE_MNAR
Low	10%	0.66	0.04	0.05	0.02	0.0006
	20%	0.73	0.06	0.05	0.03	0.002
	30%	0.83	0.07	0.04	0.02	0.004
	40%	0.94	0.11	0.07	0.03	0.01
High	10%	0.74	0.07	0.02	0.03	0.01
	20%	0.88	0.09	0.04	0.03	0.01
	30%	0.95	0.10	0.07	0.04	0.02
	40%	1.06	0.13	0.07	0.05	0.03
Covariates missing						
MNAR level	**Missing Percent**	**LW**	**MICE**	**IL**	**FIL**	**MICE_MNAR**
Low	10%	0.69	0.05	0.06	0.03	0.02
	20%	0.85	0.11	0.07	0.02	0.02
	30%	0.93	0.15	0.05	0.04	0.03
	40%	1.09	0.25	0.08	0.04	0.05
High	10%	0.86	0.13	0.05	0.02	0.03
	20%	1.03	0.16	0.05	0.03	0.03
	30%	1.22	0.20	0.06	0.03	0.06
	40%	1.34	0.32	0.08	0.06	0.05

Table 2: Coverage rates by missing data method, covariates missing value status and sample size: MNAR data.

	Sample size	**LW**	**MICE**	**IL**	**FIL**	**MICE_MNAR**
Covariates not missing	50	0.83	0.99	0.99	0.99	0.97
	100	0.85	0.99	0.98	0.99	0.98
	200	0.83	0.98	0.96	0.97	0.97
	400	0.85	0.99	0.97	0.97	0.97
	800	0.85	0.97	0.97	0.98	0.97
Covariates missing	50	0.84	0.96	0.95	0.95	0.96
	100	0.85	0.96	0.97	0.95	0.95
	200	0.85	0.98	0.96	0.95	0.95
	400	0.84	0.97	0.95	0.95	0.93
	800	0.83	0.97	0.95	0.93	0.92

Table 3: Relative estimation bias by missing data method, MAR weight: MAR data.

MAR weight	LW	MICE	IL	FIL	MICE_MNAR
0.5	0.17	0.05	0.02	0.06	0.04
1.0	0.31	0.04	0.01	0.05	0.04

Table 4: Coverage rates by missing data method, MAR weights, and percent missing: MAR data.

MAR weights	Missing Percent	LW	MICE	IL	FIL	MICE_MNAR
0.5	10%	0.84	0.96	0.98	0.98	0.96
	20%	0.85	0.94	0.98	0.97	0.95
	30%	0.85	0.95	0.97	0.96	0.94
	40%	0.85	0.97	0.99	0.95	0.96
1.0	10%	0.83	0.98	0.98	0.98	0.95
	20%	0.84	0.98	0.99	0.98	0.94
	30%	0.84	0.99	0.98	0.98	0.96
	40%	0.83	0.99	0.99	0.98	0.95

Discussion

The purpose of this study was to investigate and compare the performance of two methods designed specifically for use with MNAR data. The dependent variable of interest was dichotomous, so that LR was the modeling strategy of interest, once the missing data issue was addressed. Prior research had demonstrated the potential utility of both IL/FIL and MICE_MNAR when data were MNAR. The current study expanded on this earlier work by increasing the number of sample size conditions that were used, the percent of missing values, and by directly comparing the performance of these two methods with one another. Results of the simulation suggest that when data were MNAR, data from MICE_MNAR yielded the least biased parameter estimates across conditions, with those from FIL being very close behind. These results showing that FIL based data yields less biased estimates than IL is in keeping with earlier work [16]. It is now clear that across larger sample size conditions than had been examined heretofore, this pattern remains. Second, parameter estimate coverage rates for the MNAR data condition were at or above the nominal 0.95 level for all of the methods studied here, suggesting that researchers making use of them can have some confidence in their ability to identify the general region in which the parameters lie. Third, when the data were actually MAR, the methods designed for MNAR data performed comparably to MICE, which is generally considered one of the best methods for use with missing data [8]. Thus, researchers can have some confidence in using either MICE_MNAR or IL/FIL when they are unsure of the underlying missing data mechanism. Finally, although all methods performed relatively better (i.e., lower ARB, smaller standard errors) with larger sample sizes, FIL and MICE_MNAR consistently performed the best even for the largest samples.

Implications for practice

The results of this study present several implications for practice. First, when the dependent variable is dichotomous, thus calling for the use of LR, either FIL or MICE_MNAR appear to be good candidates for researchers to use when dealing with missing data. This conclusion appears to be true even when the underlying missing data mechanism is MAR. It does not appear that researchers need to be concerned that two approaches for dealing with MNAR data will not work in the MAR context. Second, FIL consistently performed as well as, or in some cases better than IL. Thus, although it was designed specifically for use with smaller samples, FIL appears to be a useful tool for relatively large sample cases as well. The use of standard MICE is not recommended when the outcome variable is thought to be MNAR. The results presented here demonstrate that in such cases, the estimates from a model such as LR will be relatively biased. It is important to note, however, that the sensitivity analysis recommended for use with MICE and MNAR data [25] was not used in this study. Such an approach might prove useful, and would likely perform better than standard MICE did here. However, given the positive findings for both MICE_MNAR and FIL, it is not clear that this sensitivity approach is necessary. Finally, even for samples as small as 50, MICE_MNAR yielded data that resulted in relatively low levels of bias, particularly when compared against the other approaches (with the notable exception of FIL) studied here.

Limitations and directions for future research

As with any single study, this research has limitations that need to be acknowledged, and upon which future work can be built. First, the MNAR missing data were generated using one specific approach. This method was selected so as to avoid favoring one method over another. However, it is also true that the selection model underlying the Heckman approach can also be used to generate missing values. Thus, future work should investigate how the methods studied here would perform were the missing data generated using this approach. Second, a wider array of models for both the generation of the original dataset, as well as the missing data, should be investigated. Such models might include a mix of categorical and continuous independent variables, as well as nonlinear terms involving the predictors. Furthermore, additional research should be conducted examining additional levels of MAR and MNAR. The values used here were selected in order to replicate earlier research, which was felt to be the best strategy given that additional sample size and percent missing conditions were already being used, and that two methods were being compared that had hither to not been examined together. However, given that this study has clarified the effects of sample size and percent missing, future work should expand on this work by investigating a wider array of strength of MAR/MNAR conditions. Future work could also include an MCAR condition as well.

References

1. Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods. 2002; 7: 147-177.

2. Enders CK. Applied missing data analysis. New York: The Guilford Press. 2010.

3. Schafer JL, Olsen MK. Multiple imputation for multivariate miss-

ing data problems: A data analyst's perspective. Multivariate Behavioral Research. 1998; 33: 545-571.

4. Enders CK, Bandalos DL. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Structural Equation Modeling. 2001; 8: 430-457.

5. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley. 1987.

6. Van Buuren S, Oudshoorn K. Flexible multivariate imputation by mice. Leiden, The Netherlands, TNO prevention and Health. TNO report PG/VGZ/99.054. 1999.

7. Heitjan DF, Little RJ. Multiple imputation for the fatal accident reporting system. Applied Statistics. 1991; 40: 13-29.

8. Hardt J, Herke M, Brian T, Laubach W. Multiple imputation of missing data: A simulation study on a binary response. Open Journal of Statistics. 2013; 3: 370-378.

9. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011; 45: 1-67.

10. Azur MJ, Suart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? International Journal of Methods in Psychiatric Research. 2011; 20: 40-49.

11. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. American Journal of Epidemiology. 2014; 179: 764-774.

12. Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. American Journal of Epidemiology. 2010; 172: 1070-1076.

13. Giorgi R, Belot A, Gaudart J, Launoy G. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. Statistics in Medicine. 2008; 27: 6310-6331.

14. Schunk D. A Markov chain Monte Carlo algorithm for multiple imputation in large surveys. AStA Advances in Statistical Analysis. 2008; 92: 101-114.

15. Ibrahim JG, Lipsitz SR. Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. Biometrics. 1996; 52: 1071-1078.

16. Maity AK, Pradhan V, Das U. Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. The American Statistician. 2019; 73: 340-349.

17. Firth D. Bias reduction of maximum likelihood estimation. Biometrika. 1993; 80: 27-38.

18. Heckman JJ. The common structure of statistical models of truncation sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Sociological Measurement. 1976; 5: 475-492.

19. Galimard J-E, Resche-Rigon M. miceMNAR: Missing not at random imputation models for multiple imputation by chained equation. R package version. 2018.

20. Galimard J-E, Chevret S, Curis E, Resche-Rigon M. Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. BMC Medical Research Methodology. 2018; 18: 90-103.

21. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020.

22. Kenward MG, Molenberghs G. Parametric models for incomplete continuous and categorical longitudinal studies data. Statistical Methods in Medical Research. 1999; 8: 51-83.

23. Andreis F, Ferrari PA. Missing data and parameters estimates in multidimensional item response model. Electronic Journal of Applied Statistical Analysis. 2012; 5: 431-437.

24. Leite W, Beretvas SN. The performance of multiple imputation for likert-type items with missing data, Journal of Modern Applied Statistical Methods. 2010; 9: 64-74.

25. Albert PS, Follmann D. Shared parameter models. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs. (eds.), Longitudinal Data Analysis, pp. 433-452. Boca Raton, FL: CRC Press. 2009.