A Review of Complete Case Analysis, Multiple Imputation with EM Algorithm and MCMC in Observational Study with an Example of Missing Data in National Childhood Development Study

Jiaqi Bi[a]

[a] *Western University,*
Schulich School of Medicine & Dentistry,
Department of Epidemiology and Biostatistics

Abstract

Numerous publications either fail to identify the validity and the reason for using an imputation method or conduct the imputation without considering the missing mechanism. Some who have done Multiple Imputation (MI) in their studies do not mention the option of the MI, which may lead to potential misinterpretation for the reader. The objective of this paper is to conduct a comprehensive review of the three missing data handling methods including prevalent missing data handling techniques: complete case analysis, MI via Expectation-Maximization (MI-EM), and MI via Markov Chain Monte Carlo (MI-MCMC) to compare their performance and inference within MCAR and MNAR scenarios based on the factors of residual standard errors, sample sizes, and missing portions. The results have shown that in planned missing MCAR, the CCA is valid and its results are accurate. But within the MNAR, the MI-EM or MI-MCMC are suitable choices, while the CCA is biased. The study also identifies that the extreme residual standard error, sample size, and the missing rate will cause a biased results for MI methods. Yet, it is also noted that MI-EM and MI-MCMC can be used interchangeably, as there is no significant difference in performance between these two methods.

Keywords: Missing data, Multiple imputation, complete case analysis, EM algorithm, MCMC

Email address: jbi23@uwo.ca (Jiaqi Bi)

1. Introduction

1.1. Observational Study

An observational study is frequently employed in the fields of Epidemiology, statistical consulting, social sciences, and psychology [1]. The data derived from the observational study is typically utilized for inferring from the sample to the population. The primary distinction between an observational study and a clinical trial lies in the fact that a clinical trial is specifically designed to investigate the treatment effect, with the assignment of treatment and control groups being commonly implemented [1, 2]. Nevertheless, the presence of missing values in an observational study can pose challenges during the analysis process, as the covariate that includes these missing values may have a causal relationship with the outcome. However, due to information being omitted during data collection or sampling, these missing values create further complexities in the analysis [3]. The Multiple Imputation (MI) method is extensively employed by analysts in order to address missing data; however, it is crucial for researchers to comprehend the differences among the various available options. Numerous publications either fail to specify the imputation methods utilized or neglect to justify the rationale behind employing such imputation methods according to Hayati Rezvan et al. [4]. In this paper, I made evaluations based on differences in the residual standard error, the missing percentage, and the sample size, along with an observational study example of the National Childhood Development Study (NCDS) to investigate two methods in multiple imputations.

1.2. Missing Mechanism

The issue of missing data emerged in the publication which consisted of making inferences about missing data, with further elaboration on the missing mechanism introduced by Rubin. The data is regarded as missing completely at random (MCAR) when the missingness exhibits no correlation with either observed or missing values within the data. A less restrictive mechanism is called missing at random (MAR), which allows the missingness depends on the observed components of other variables, i.e., some variables excluding the one that contains missing values may have the capacity to predict the missingness. The occurrence of MAR is more prevalent in practice than MCAR. The third mechanism, referred to as missing

not at random (MNAR), arises when the variable itself is associated with the missingness within the variable [5, 6]. The missing mechanism needs detailed considerations whenever dealing with the missing data, whether the researcher opts for a straightforward approach such as listwise deletion, or the researcher tries to apply a more complex methodology such as multiple imputations [5, 7].

1.3. Multiple Imputation Options

Multiple imputation (MI) is extensively utilized to address missing data in observational studies, offering numerous built-in options within statistical software packages such as R, Stata, SAS, and etc [4, 8, 9, 10]. Additionally, MI encompasses several algorithms, including the multivariate normal algorithm, also commonly known as Expectation-Maximization (EM) algorithm, and the Markov Chain Monte Carlo (MCMC) method, which is often referred to as the chained method. The latter is a widely employed approach in handling non-monotone missing data patterns in clinical trials and in observational studies introduced by Little and Rubin [11] and summarized by Zhang [12]. The MI-EM method iteratively performs E-step and M-step, the E-step computes the expected values of the missing data given the observed data and the current estimate of the model parameters. The M-step computes the maximum likelihood estimate of the parameters given the observed data and the expected values computed in the E-step [13]. The MI-EM strictly requires assumptions on the missing mechanism of MAR of the data. Also, the MI-EM tends to be more "frequentist" flavor when encountering missing data [8]. The MCMC, on the other hand, is a probabilistic method for estimating the distribution of a variable. It uses a random sampling process to generate a series of guesses that eventually converge on the true distribution. It is more computationally intensive than EM, but it can handle more complex models and missing data situations [14]. It does not have to make the assumption of MAR. Indeed, the MI-MCMC is a "Bayesian" methodology [8]. The EM is computationally faster and requires fewer iterations because the MCMC involves integration, which is much more complex in computers. Conversely, the MCMC method generally demonstrates greater robustness when compared to the EM algorithm. In this study, the simulation of the data and the example of the data both satisfy the assumption of the MAR, in order to make the explicit comparison

of these two methods.

2. Methods

2.1. Data Simulation and Example of NCDS Data

This paper involves a simulation study based on two MI methods and an NCDS example. The data is generated and analyzed using Stata software. The setting of the generated data consists of different standard errors σ_ϵ of the residual ϵ, assuming the outcome Y has the following regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

such that $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. The x_i is generated from $N(5, 1.5^2)$, and y_i is generated from setting the intercept $\beta_0 = 10$ and the slope $\beta_1 = 5$. The sample size N where $i = \{0, ..., N\}$ of generated data is considered to be $\{100, 1000, 5000\}$, while $\epsilon = \{1, 2, 5, 10\}$. The outcome Y does not contain missing values, but the covariate $X = \{X_{obs}, X_{mis}\}$ denotes observed values and missing values, and $P(X_{mis}) = \{0, 0.1, 0.2, 0.5, 0.8\}$. Therefore, a total of 60 scenarios are considered, including completely observed cases and partially observed cases. The implementation of linear regression is applied to all scenarios, with the options of a complete case analysis (CCA), MI-EM, and MI-MCMC. The evaluation of imputation methods is based on the examination of fitted coefficients and their statistics. The missing probability remains fixed as mentioned above, while the assignment of missingness to individuals is randomized to simulate the MAR scenario.

The dataset of NCDS is a longitudinal study that commenced in 1958 and focuses on tracking the trajectories of 17,415 individuals who were born in England, Wales, and Scotland during a specific week of the year. This study was originally known as the Perinatal Mortality Survey, which began in 1958 [15]. The principal outcome in the data is the reciprocal birth weight in oz, which is a continuous real number. The interested covariates are the mother's reading comprehensive test score, the infant's sex, and social housing status, the mother's marital status (married or single, indicator denotes the single status), and the mother's age at birth centered around 28 years old [16]. The desired regression is then formulated as

$$y_{birthwt,i} = \beta_0 + \beta_1 x_{readtest,i} + \beta_2 \delta_{sex,i} + \beta_3 \delta_{socialhous,i} + \beta_4 \delta_{single,i} + \beta_5 x_{momage,i} \tag{2}$$

where δ's are indicators corresponding to each category mentioned above for i-th observation. The x's are continuous covariates. The present study examines and compares complete case analysis, MI-EM, and MI-MCMC under the above scenarios. t is important to note that the missing mechanism observed in the NCDS dataset does not strictly adhere to the missing at random (MAR) assumption. Potential reasons for missingness include loss of follow-up or respondents being unwilling or unable to provide responses, which may be the case of MNAR [6]. Therefore, a simple test with a logistic regression based on the probability of missingness and variables is performed to investigate the missing mechanism of this data.

2.2. Multiple Imputation Steps

The multiple imputation steps in all softwares basically follow I-A-P algorithm: Imputation, Analysis, and Pooling. In the I-step, the missing values are imputed m times, creating m complete datasets. Each of these imputations is drawn from the posterior predictive distribution of the missing data given the observed ones. In the A-step, each of the created dataset is analyzed using the statistical methods (linear regression in my study), resulting in m sets of parameter estimates and standard errors. In the P-step, the m sets of parameter estimates are pooled into one set of estimates and standard errors using Rubin's MI rules introduced in 2004 [7]. In the Rubin's rule, the final estimate is the mean of the m estimates, and the standard error is a function of both the within-imputation variance (the average of the m standard errors) and the between-imputation variance (the variance of the m estimates). The pooled estimate is calculated as $\phi = \frac{1}{m}\sum_{i=1}^{m}\phi_i$, the within-imputation variance is calculated as $W = \frac{1}{m}\sum_{i=1}^{m}W_i$, note that W_i is the variance of ϕ_i, and the between-imputation variance is $B = \frac{1}{m-1}\sum_{i=1}^{m}(\phi_i - \phi)^2$. The total variance is simply $T = W + (1 + 1/m)B$ according to Rubin [7] and Marchenko [8], which described in the Stata manual. Either the EM-algorithm, or the MCMC method happens in the I-step, which the details of two algorithms are described further.

2.3. Expectation-Maximization Algorithm

The MI-EM method performs iteratively E-step and M-step, in which the E-step computes the expected values of the missing data given the observed data and the current

estimate of the model parameters, and the M-step computes the maximum likelihood estimate of the parameters given the observed data and the expected values in the E-step. Generally, in the E-step, we denote Y_{obs} and Y_{mis}, and current parameter estimate of θ^t, which means the initial value of the parameter θ. This θ depends on the model, and I used this notation for simplification without loss of generality. The complete data $D = (Y_{obs}, Y_{mis})$ has a log-likelihood, the log-likelihood in our study can be derived from the model,

$$\ell(\theta) = \log L(\theta; Y_{obs}, Y_{mis}) = \log f(Y_{obs}, Y_{mis}). \tag{3}$$

The example of NCDS regression of Equation 2 has the log-likelihood written as (since a linear regression assumes the outcome has a normal distribution):

$$\ell(\boldsymbol{\beta}, \sigma^2; y_i, \mathbf{x}_i) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{birthwt,i} - \boldsymbol{\beta}\mathbf{x}_i) \tag{4}$$

Now back to the general equation of the log-likelihood in Equation 3, the E-step computes the expectations of the complete data log-likelihood, which denotes

$$E(\theta|\theta^{(t)}) = E_{Y_{mis}|Y_{obs},\theta^{(t)}}[\log L(\theta|Y_{obs}, Y_{mis})] \tag{5}$$

where $L(\theta|Y_{obs}, Y_{mis})$ is the complete data likelihood. The M-step finds the parameter that maximizes the expectation of $E(\theta|\theta^t)$:

$$\theta^{(t+1)} = \arg\max_{\theta} E(\theta|\theta^{(t)}) \tag{6}$$

Then these steps are iterated until convergence, i.e., when the updates in the estimates are smaller than the pre-defined threshold. In the NCDS example, one should treat any missing values in any variable as missing data and calculate the expectation of the log-likelihood in Equation 4 under the current parameter estimates.

2.4. Markov Chain Monte Carlo

The MCMC is a Bayesian point of view, and contains a probabilistic method for estimating the distribution of the variable. It uses a random sampling process to generate a

series of guesses that eventually converge on the true distribution. Using the same formula in the last section defined for the observed and missing data, and the parameter, the MCMC method first chooses an arbitrary point to be the first sample and set the current position as the initial point. Then the MCMC starts in iterations. In each iteration, a candidate for the next sample is generated from a proposal distribution which depends on the current position, as known as the "Markov Chain". Then the candidate is probabilistically either accepted as the next sample or the current position is used as the next sample instead, as known as "Monte Carlo" [14]. When we sample from a target distribution $p(\theta|y)$ but cannot do so directly, we construct a Markov Chain instead whose stationary distribution is $p(\theta|y)$. The output of MCMC is generally a sequence of samples that can be used to approximate the target distribution and to compute integrals (such as expected values) with respect to this distribution. In the context of MI, MCMC would be used to draw from the posterior predictive distribution of the missing data given the observed data and the current imputed values of the missing data. The specifics of how to set this up would depend on the model for the complete data (observed and missing) [8, 17, 18]. It's important to note that while the EM algorithm's convergence is to a point estimate, the MCMC's convergence is to a distribution. This reflects the fact that the EM algorithm is a method for maximum likelihood estimation, which aims to find a single "best" estimate, while MCMC is a method for Bayesian inference, which aims to characterize the full posterior distribution of the parameters. Some articles argue the MI-MCMC is more robust due to its ability to handle complex models because its algorithm handles hierarchical models with interaction effects and non-linear models, which leads to a better performance with non-normal data. The EM algorithm requires assumptions about the specific data distribution, which is quite sensitive to violations. Also mathematically, the MCMC avoids finding the local maximum or local minimum, which the MCMC does not have a derivative procedure inside the algorithm. It uses Gibbs sampling to avoid complex mathematical computation [3, 19, 20].

3. Results

3.1. Simulation Study

Tables 2, 3 and 4 are details on the simulation study of the three missing data analysis methods. The data missing mechanism in the simulation as mentioned in the Methods section is MAR, and it can also be seen as MCAR explicitly as the missing rate is planned, which simulates the scenario of drawing m samples from n population. Therefore, all three methods are valid in this scenario. The CCA has better accuracy, but unstable confidence intervals in the MCAR simulation study.

Within the same sample, the desired results considered a good performance should be as close as possible to the regression coefficients when there are no missing values. As shown in Table 2, 3 and 4, when the sample size increases, along with a fixed missing rate, the imputation yields a more strengthened certainty with a narrow range of 95% confidence interval. This can also be seen in Figure 2 and 3. However, the CCA in this situation yields a larger range of confidence intervals compared to the imputation methods, as shown in Figure 1. This is expected since the CCA does not impute values, it performs "listwise deletion" and the analyzed sample size is smaller than the full data, i.e., only Y_{obs} are involved with the statistical analysis. The p-value is not in favor when the sample size is small, but the residual standard error is high for the imputation methods as shown in Table 2 at $\sigma_\epsilon = 5$ and $\sigma_\epsilon = 10$. However, when the missing rate increases, the CCA performs a closer result to the no-missing fitted coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. The MI-EM and MI-MCMC have shown a similar result with inflated $\hat{\beta}_0$'s, and deflated $\hat{\beta}_1$'s. Figures 4, 5 and 6 also illustrate that the confidence interval in CCA has a wider range than the imputation methods, while the accuracy tends to be better than MI-EM and MI-MCMC. Fixing the sample size and the missing rate, the residual standard error affects the imputation accuracy as well referred in Figure 7, 8 and 9. When the residual standard error increases, the CCA does not have a robust performance as the imputation methods. The confidence intervals for the three methods have shown that the CCA has a relatively large range for the regression coefficients, and the confidence intervals for MI-EM and MI-MCMC seem to be stabilized.

3.2. NCDS Example

In order to validate the regression, the density of the response variable must be normally distributed. The infant's birth weight, as shown in Figure 10, is approximately normal which satisfies the assumption of a linear regression. The NCDS data is non-monotonic, and a logistic regression based on the missingness with covariates is performed, several covariates including the mother's age and reading test score are correlated with their missingness, i.e., the variable contains missing values itself predicts the missingness, which is showing the data is MNAR. However, there are publications that fail to identify the missing mechanisms and perform the listwise deletion, which may cause biased or even false results [4]. In this example, the CCA is not valid for MNAR cases, but illustrating the comparison between the three methods is crucial to identify the reason that the CCA is biased.

Table 1 shows the three missing data handling methods results, which consist of the fitted coefficients and their confidence intervals. In the analysis using CCA, the reading score, gender, marital status, and mother's age are correlated with the infant's birth weight. while there are only 9739 observations with no missing values included in the regression. The CCA shows a higher portion of significant results do not explain the validity of the missing data, because the missing data can cause a biased result if the missing values are causal, and it is fairly common in large sample size. However, the MI-EM and MI-MCMC impute the missing values and conduct the statistical analysis, where the mother's age is not significant anymore. The performance between the MI-EM and MI-MCMC is similar, but there are arguments that the MI-EM should only be conducted when the MAR mechanism exists. It is clearly not our case, but the imputation results do not vary much. During the imputation process, the MI-MCMC took much longer than the MI-EM to complete the analysis. This is due to the nature of the two different methods, that the model contains multiple variables that can lead the distribution to be complex for the MI-MCMC to deal with. The result also shows a slight inflation in the $\hat{\beta}_0$ for MI-EM and MI-MCMC, and this happened in the simulation process as well.

4. Conclusion

4.1. Summary

The issue of missing data is an unavoidable challenge in data analysis and demands a systematic resolution. The most crucial step in any analytical process should be the identification of the missing data mechanism, as missing values can potentially be causal and influential. Even within the context of MAR, it is important to note that CCA may still introduce bias. In the present study, we delve into two significant cases. The first involves a simulation study focusing on the MCAR mechanism, and the second considers an example of MNAR. Certain scholarly discourses suggest that the MI-EM should be restricted to MAR scenarios, positing that MI-MCMC offers greater robustness concerning different missing data mechanisms. Moreover, this study identifies specific conditions under which imputation might be highly biased. Notably, when the residual standard error is elevated, the imputation methods may produce inaccurate results, albeit with stable confidence intervals. The rate of missing data is a crucial factor to ascertain, as the findings of this study indicate that a higher missing rate leads to reduced accuracy and an increased Type-II error rate.

Considering the methodology, in the MCAR case, the CCA may be a sufficient choice when dealing with the missing data, because the missing values are "ignorable". The difference in performance and validity of MI-EM and MI-MCMC are not too significant. In the MNAR case, the CCA is biased and will result in an inaccurate result as well as the Type-I error for some covariates. The MI-EM and MI-MCMC both have a more valid performance, while the MI-EM does not seem to be unsuitable for this case. In conclusion, whenever dealing with MCAR or MAR missing data, one should always clarify if the missing is planned before the implementation of the imputation. The MI-EM and MI-MCMC can be both applied to the case of the MNAR, but the CCA should be avoided.

4.2. Discussions on Further Works

Indeed, there is no best imputation method or solution for the missing data. But the statistical inference in the missing data requires lots of details, particularly when the data analysts are confronted with an unknown missing data mechanism. Initial, or preliminary,

tests might offer a beneficial tool for navigating these circumstances. It could be advantageous to implement these preliminary tests on data missingness prior to applying any specific methodology. Subsequently, to evaluate the differences between direct imputation and proper imputation post-preliminary tests, a simulation study could be conducted. This further step would aid in the comprehensive comparison and analysis of these imputation methods.

5. Appendix

5.1. Stata Code

Please see attached do file.

5.2. Tables and Figure

Starts from the next page.

Table 1: The NCDS Example, *** indicates $p < 0.001$, ** indicates $p < 0.05$, * indicates $p < 0.01$.

	Birthweight (CCA)		Birthweight (MI-EM)		Birthweight (MI-MCMC)	
	Coefficient	95%CI	Coefficient	95%CI	Coefficient	95%CI
Reading Score	-0.00264***	-.0030939,-.0021829	-0.00845***	-.0101407,-.0067667	-0.00849***	-.0097772,-.0071963
Sex	0.0383***	.0322473,.0444158	0.0388***	.0305988,.0469918	0.0386***	.0302436,.0468603
Social Housing	0.000865	-.0055073,.007238	0.0110	-.0045974,.0265986	0.00848	-.0081401,.0250934
Marital Status	0.0301***	.0125116,.0477867	0.0425***	.0204553,.0645347	0.0427***	.0205882,.0647902
Mother Age	-0.00137***	-.0019182,-.0008317	0.000542	-.0001921,.0012753	0.000557	-.0001684,.0012832
β_0	0.921***	.9074225,.9338612	1.088***	1.041895,1.134528	1.090***	1.054736,1.126154
N	9739		17631		17631	

Table 2: The Simulation Results for $N = 100$, the coefficients before the semicolon is $\hat{\beta}_1$ and after is $\hat{\beta}_0$. 95% Confidence interval inside the brackets. *** indicates $p < 0.001$, ** indicates $p < 0.05$, * indicates $p < 0.01$.

$N = 100$	Regression Coefficient	CCA	MI-EM	MI-MCMC
$\sigma_\epsilon = 1$				
No Missing	5.029*** (4.895, 5.167); 9.941*** (9.249, 10.063)			
10% Missing		5.031*** (4.890, 5.173); 9.955*** (9.228, 10.680)	4.459*** (3.814, 5.103); 12.820*** (9.402, 16.250)	4.454*** (3.577, 5.331); 12.83*** (8.341, 17.330)
20% Missing		5.011*** (4.861, 5.160); 9.980*** (9.206, 10.750)	3.638*** (2.576, 4.700); 16.910*** (11.810, 22.020)	3.905*** (2.800, 5.010); 15.500*** (9.796, 21.200)
50% Missing		4.994*** (4.803, 5.184); 10.060*** (9.088, 11.030)	2.138** (0.615, 3.661); 24.460*** (16.540, 32.390)	2.356** (0.936, 3.776); 23.310*** (16.360, 30.260)
80% Missing		4.913*** (4.666, 5.160); 10.320*** (9.080, 11.560)	0.359 (-0.614, 1.333); 32.920*** (28.110, 37.730)	0.655 (-0.585, 1.894); 31.700*** (25.670, 37.730)
$\sigma_\epsilon = 2$				
No Missing	5.059*** (4.789, 5.328); 9.883*** (8.497, 11.270)			
10% Missing		5.063*** (4.781, 5.345); 9.909*** (8.456, 11.360)	4.362*** (3.314, 5.409); 13.370*** (7.755, 18.980)	4.446*** (3.742, 5.150); 13.060*** (9.593, 16.520)
20% Missing		5.021*** (4.722, 5.320); 9.960*** (8.413, 11.510)	3.874*** (2.831, 4.917); 15.580*** (9.949, 21.210)	3.685*** (2.369, 5.001); 16.670*** (9.725, 23.620)
50% Missing		4.988*** (4.607, 5.369); 10.110*** (8.176, 12.050)	2.289*** (1.076, 3.502); 23.900*** (18.280, 29.520)	2.414* (0.373, 4.455); 23.440*** (13.440, 33.440)
80% Missing		4.825*** (4.331, 5.320); 10.630*** (8.159, 13.110)	0.627 (-0.439, 1.692); 31.960*** (27.170, 36.750)	0.376 (0.754, 1.507); 32.870*** (27.020, 38.710)
$\sigma_\epsilon = 5$				
No Missing	5.147*** (4.473, 5.821); 9.707*** (6.243, 13.170)			
10% Missing		5.157*** (4.452, 5.863); 9.773*** (6.139, 13.410)	4.464*** (3.521, 5.407); 12.930*** (8.158, 17.710)	4.659*** (3.811, 5.508); 12.190*** (7.871, 16.520)
20% Missing		5.053*** (4.306, 5.800); 0.901*** (6.031, 13.770)	3.932*** (2.888, 4.976); 15.780*** (10.580, 20.970)	3.821*** (2.624, 5.018); 16.150*** (9.841, 22.470)
50% Missing		4.969*** (4.017, 5.922); 10.280*** (5.441, 15.130)	2.155* (0.329, 3.982); 24.570*** (15.940, 33.210)	2.493* (0.681, 4.305); 23.050*** (14.190, 31.920)
80% Missing		4.564*** (3.328, 5.799); 11.590*** (5.398, 17.780)	0.499 (-0.696, 1.693); 32.780*** (27.180, 38.380)	0.467 (-0.759, 1.694); 23.920*** (27.300, 38.530)
$\sigma_\epsilon = 10$				
No Missing	5.294*** (3.946, 6.642); 9.415** (2.487, 16.340)			
10% Missing		5.315*** (3.903, 6.726); 9.545* (2.279, 16.810)	4.542*** (2.745, 6.339); 13.200** (4.242, 22.150)	4.770*** (3.324, 6.216); 12.150** (4.710, 19.590)
20% Missing		5.106*** (3.612, 6.599); 9.801* (2.063, 17.540)	3.742*** (2.017, 5.467); 17.130*** (8.401, 25.870)	3.707*** (1.874, 5.540); 17.060*** (7.545, 26.580)
50% Missing		4.939*** (3.033, 6.845); 10.570* (0.882, 20.250)	1.961 (-0.232, 4.154); 26.040*** (15.050, 37.040)	2.322** (0.653, 3.991); 24.310*** (16.130, 32.490)
80% Missing		4.127** (1.656, 6.598); 13.170* (0.796, 25.550)	0.617 (-1.029, 2.264); 32.37*** (24.140, 40.610)	0.301 (-1.350, 1.952); 33.980*** (25.850, 42.100)

Table 3: The Simulation Results for $N = 1000$, the coefficients before the semicolon is $\hat{\beta}_1$ and after is $\hat{\beta}_0$. 95% Confidence interval inside the brackets. *** indicates $p < 0.001$, ** indicates $p < 0.05$, * indicates $p < 0.01$.

$N = 1000$	Regression Coefficient	CCA	MI-EM	MI-MCMC
$\sigma_\epsilon = 1$				
No Missing	4.995*** (4.955, 5.035)			
10% Missing		4.981*** (4.939, 5.024); 10.070*** (9.845, 10.290)	4.513*** (4.260, 4.766); 12.400*** (11.130, 13.670)	4.485*** (4.315, 4.656); 12.540*** (11.680, 13.410)
20% Missing		4.978*** (4.933, 5.023); 10.100*** (9.870, 10.330)	4.009*** (3.675, 4.344); 14.930*** (13.160, 16.700)	4.022*** (3.691, 4.352); 14.920*** (13.280, 16.570)
50% Missing		4.973*** (4.919, 5.027); 10.160*** (9.884, 10.440)	2.568*** (2.166, 2.970); 22.120*** (20.000, 24.230)	2.669*** (2.240, 3.099); 21.610*** (19.340, 23.880)
80% Missing		5.016*** (4.930, 5.101); 10.010*** (9.557, 10.470)	1.070*** (0.515, 1.626); 29.310*** (26.480, 32.140)	1.068*** (0.573, 1.564); 29.390*** (26.920, 31.870)
$\sigma_\epsilon = 2$				
No Missing	4.990*** (4.910, 5.071); 9.983*** (9.566, 10.400)			
10% Missing		4.962*** (4.877, 5.047); 10.130*** (9.690, 10.570)	4.468*** (4.202, 4.734); 12.570*** (11.240, 13.900)	4.442*** (4.196, 4.688); 12.730*** (11.470, 14.000)
20% Missing		4.956*** (4.867, 5.046); 10.200*** (9.739, 10.670)	3.960*** (3.668, 4.253); 15.140*** (13.660, 16.610)	3.937*** (3.684, 4.190); 15.250*** (13.870, 16.640)
50% Missing		4.947*** (4.839, 5.054); 10.330*** (9.768, 10.880)	2.621*** (2.147, 3.095); 21.790*** (19.460, 24.120)	2.578*** (2.048, 3.109); 21.950*** (19.150, 24.740)
80% Missing		5.032*** (4.861, 5.203); 10.02*** (9.114, 10.930)	1.111*** (0.682, 1.540); 21.790*** (19.460, 24.120)	1.026*** (0.602, 1.450); 29.510*** (27.270, 31.750)
$\sigma_\epsilon = 5$				
No Missing	4.975*** (4.774, 5.177); 9.958*** (8.914, 11.000)			
10% Missing		4.906*** (4.693, 5.118); 10.330*** (9.224, 11.430)	4.390*** (4.078, 4.703); 12.870*** (11.230, 14.510)	4.401*** (4.071, 4.730); 12.850*** (11.170, 14.530)
20% Missing		4.891*** (4.667, 5.114); 10.510*** (9.348, 11.670)	3.956*** (3.629, 4.283); 15.010*** (13.410, 16.610)	3.893*** (3.546, 4.240); 15.330*** (13.540, 17.110)
50% Missing		4.866*** (4.597, 5.136); 10.810*** (9.421, 12.200)	2.621*** (2.058, 3.185); 21.720*** (18.910, 24.530)	2.632*** (2.137, 3.126); 21.640*** (19.090, 24.190)
80% Missing		5.079*** (4.651, 5.507); 10.050*** (7.785, 12.330)	1.063*** (0.507, 1.619); 29.210*** (26.320, 32.090)	1.030*** (0.515, 1.546); 29.390*** (26.690, 32.090)
$\sigma_\epsilon = 10$				
No Missing	4.951*** (4.548, 5.353); 9.916*** (7.828, 12.000)			
10% Missing		4.811*** (4.387, 5.236); 10.650*** (8.448, 12.850)	4.329*** (3.873, 4.784); 13.040*** (10.710, 15.371)	4.340*** (3.870, 4.810); 12.950*** (10.530, 15.380)
20% Missing		4.781*** (4.334, 5.228); 11.010*** (8.696, 13.330)	3.922*** (3.440, 4.404); 15.070*** (12.580, 17.570)	3.840*** (3.337, 4.343); 15.430*** (12.830, 10.020)
50% Missing		4.733*** (4.194, 5.271); 11.630*** (8.842, 14.410)	2.442*** (1.894, 2.989); 22.440*** (19.650, 25.220)	2.486*** (1.819, 3.153); 22.220*** (18.840, 25.600)
80% Missing		5.158*** (4.303, 6.014); 10.110*** (5.569, 14.650)	1.031* (0.255, 1.807); 29.270*** (25.320, 33.210)	1.133** (0.342, 1.923); 28.760*** (24.750, 32.770)

Table 4: The Simulation Results for $N = 5000$, the coefficients before the semicolon is $\hat{\beta}_1$ and after is $\hat{\beta}_0$. 95% Confidence interval inside the brackets. *** indicates $p < 0.001$, ** indicates $p < 0.05$, * indicates $p < 0.01$.

$N = 5000$	Regression Coefficient	CCA	MI-EM	MI-MCMC
$\sigma_\epsilon = 1$				
No Missing	5.007*** (4.989, 5.026); 9.964*** (9.869, 10.060)			
10% Missing		5.009*** (4.989, 5.028); 9.956*** (9.855, 10.060)	4.510*** (4.412, 4.608); 12.440*** (11.940, 12.930)	4.513*** (4.418, 4.609); 12.440*** (11.970, 12.910)
20% Missing		5.015*** (4.994, 5.035); 9.925*** (9.819, 10.030)	3.991*** (3.821, 4.161); 15.010*** (14.160, 15.860)	4.011*** (3.883, 4.139); 14.940*** (14.290, 15.590)
50% Missing		5.018*** (4.992, 5.044); 9.887*** (9.753, 10.020)	2.540*** (2.331, 2.749); 22.230*** (21.120, 23.350)	2.557*** (2.388, 2.726); 22.170*** (21.300, 23.040)
80% Missing		5.009*** (4.969, 5.049); 9.922*** (9.716, 10.130)	1.025*** (0.804, 1.246); 29.730*** (28.660, 30.880)	1.056*** (0.806, 1.307); 29.610*** (28.370, 30.860)
$\sigma_\epsilon = 2$				
No Missing	5.015*** (4.978, 5.052); 9.929*** (9.737, 10.120)			
10% Missing		5.017*** (4.978, 5.056); 9.912*** (9.710, 10.110)	4.499*** (4.383, 4.614); 12.510*** (11.940, 13.070)	4.501*** (4.399, 4.602); 12.490*** (11.970, 13.020)
20% Missing		5.029*** (4.988, 5.070); 9.851*** (9.637, 10.060)	4.025*** (3.907, 4.142); 14.890*** (14.270, 15.500)	4.020*** (3.860, 4.179); 14.900*** (14.080, 15.720)
50% Missing		5.037*** (4.985, 5.088); 9.773*** (9.506, 10.040)	2.516*** (2.258, 2.774); 22.350*** (21.040, 23.660)	2.532*** (2.361, 2.703); 22.260*** (21.290, 23.230)
80% Missing		5.018*** (4.938, 5.098); 9.844*** (9.432, 10.260)	1.047*** (0.848, 1.245); 29.670*** (28.630, 30.700)	1.013*** (0.780, 1.246); 29.830*** (28.650, 31.020)
$\sigma_\epsilon = 5$				
No Missing	5.037*** (4.944, 5.129); 9.822*** (9.343, 10.030)			
10% Missing		5.043*** (4.945, 5.140); 9.780*** (9.275, 10.280)	4.514*** (4.391, 4.637); 12.430*** (11.790, 13.080)	4.533*** (4.406, 4.660); 12.330*** (11.690, 12.980)
20% Missing		5.073*** (4.969, 5.176); 9.627*** (9.093, 10.160)	4.048*** (3.825, 4.271); 14.750*** (13.550, 15.940)	4.070*** (3.870, 4.269); 14.640*** (13.660, 15.620)
50% Missing		5.092*** (4.962, 5.221); 9.433*** (8.766, 10.100)	2.602*** (2.391, 2.814); 21.920*** (20.840, 23.010)	2.585*** (2.323, 2.847); 21.980*** (20.660, 23.290)
80% Missing		5.045*** (4.845, 5.245); 9.610*** (8.580, 10.640)	1.074*** (0.757, 1.391); 29.530*** (27.920, 31.150)	1.081*** (0.874, 1.287); 29.500*** (28.450, 30.550)
$\sigma_\epsilon = 10$				
No Missing	5.073*** (4.889, 5.258); 9.644*** (8.687, 10.600)			
10% Missing		5.085*** (4.890, 5.280); 9.560*** (8.550, 10.570)	4.591*** (4.382, 4.799); 12.040*** (10.980, 13.110)	4.593*** (4.359, 4.826); 12.050*** (10.830, 13.270)
20% Missing		5.145*** (4.939, 5.352); 9.254*** (8.186, 10.320)	4.107*** (3.856, 4.359); 14.470*** (13.180, 15.750)	4.101*** (3.826, 4.377); 14.480*** (13.080, 15.890)
50% Missing		5.183*** (4.925, 5.441); 8.867*** (7.532, 10.200)	2.618*** (2.318, 2.919); 21.870*** (20.350; 23.400)	2.635*** (2.360, 2.909); 21.790*** (20.320, 23.250)
80% Missing		5.090*** (4.690, 5.491); 9.219*** (7.161, 11.280)	1.048*** (0.751, 1.346); 29.660*** (28.160, 31.170)	1.049*** (0.671, 1.426); 29.670*** (27.760, 31.570)

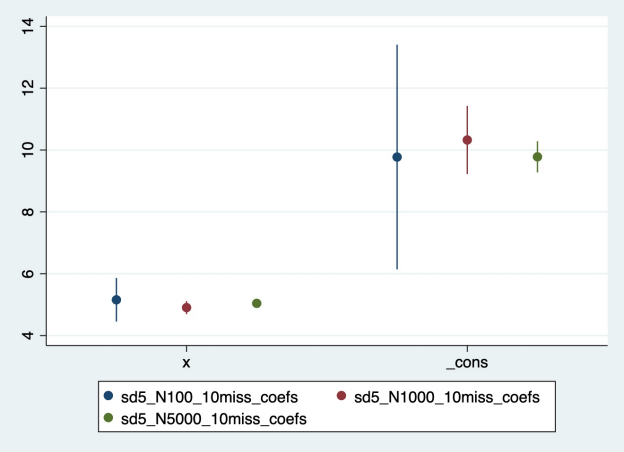

Figure 1: CCA Different Sample Sizes: $N = 100, 1000, 5000$

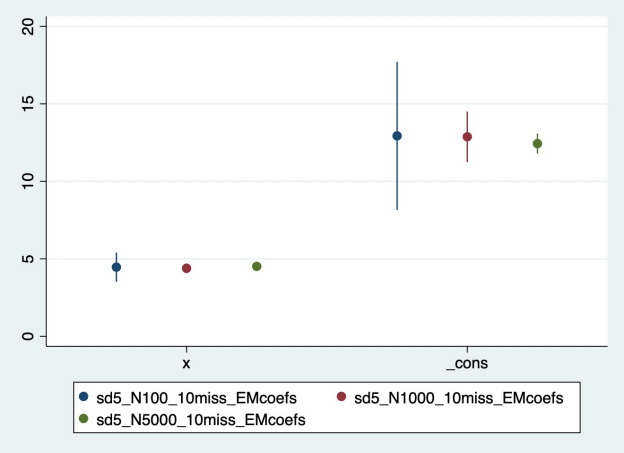

Figure 2: MI-EM Different Sample Sizes: $N = 100, 1000, 5000$

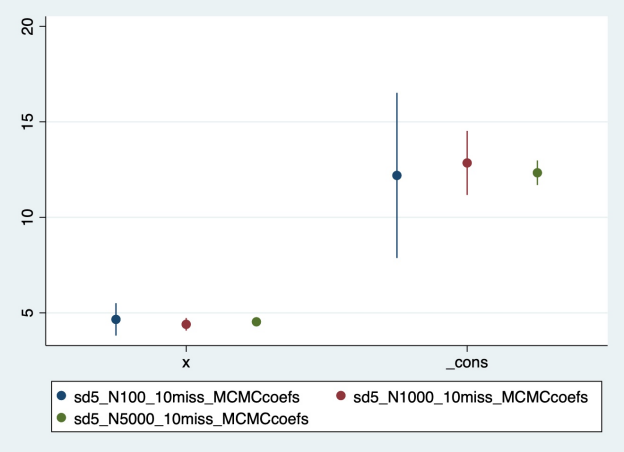

Figure 3: MI-MCMC Different Sample Sizes: $N = 100, 1000, 5000$

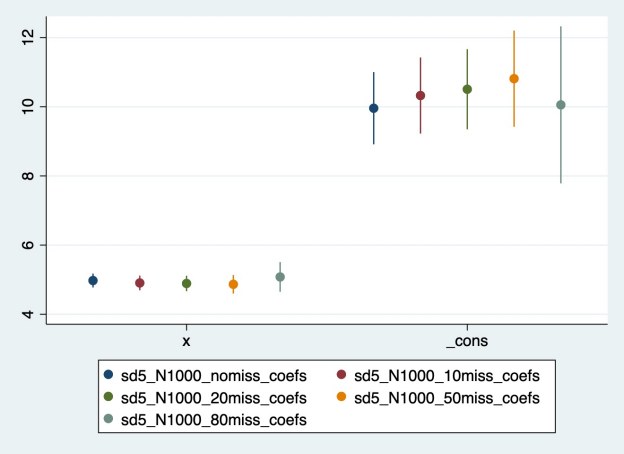

Figure 4: CCA Missing Rate Increases for $\sigma_\epsilon = 5$, $N = 1000$

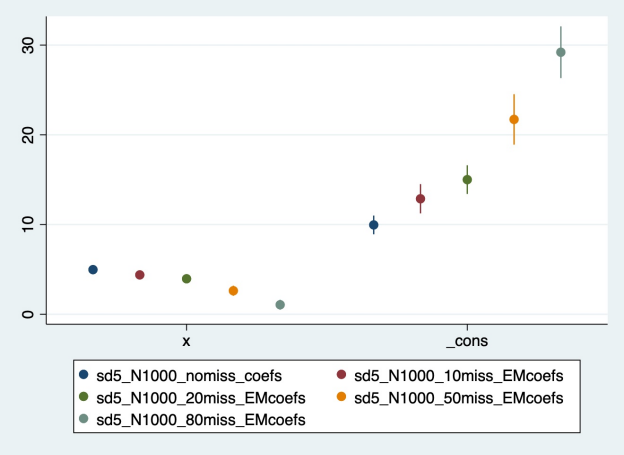

Figure 5: MI-EM Missing Rate Increases for $\sigma_\epsilon = 5$, $N = 1000$

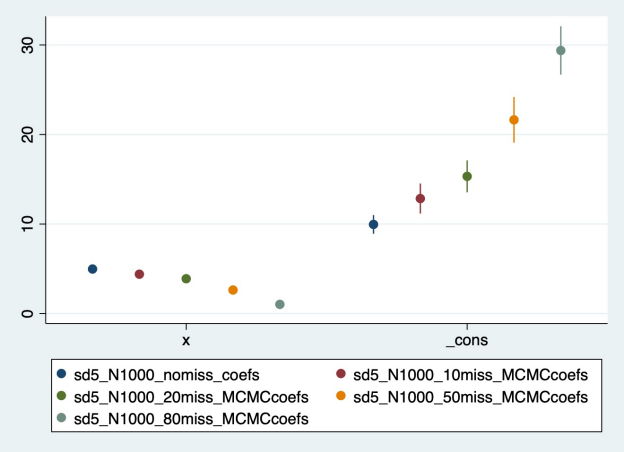

Figure 6: MI-MCMC Missing Rate Increases for $\sigma_\epsilon = 5$, $N = 1000$

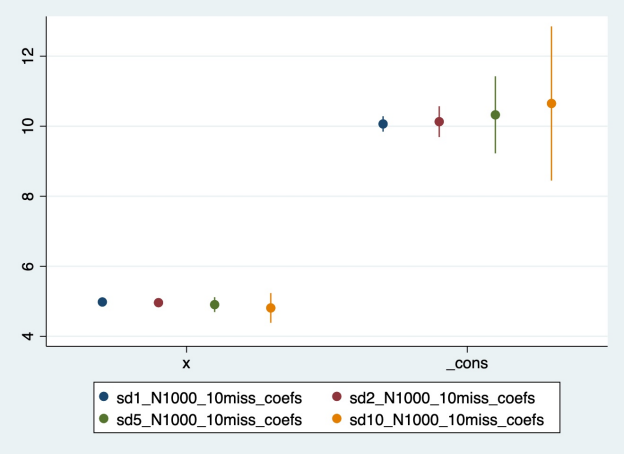

Figure 7: CCA Standard Error Increases for 10% Missing Rate, $N = 1000$

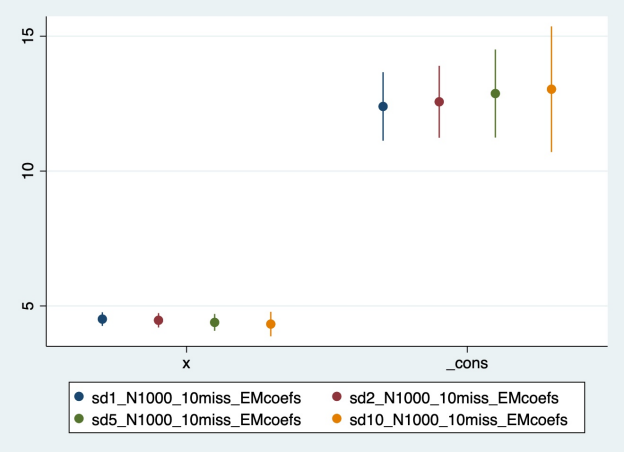

Figure 8: MI-EM Standard Error Increases for 10% Missing Rate, $N = 1000$

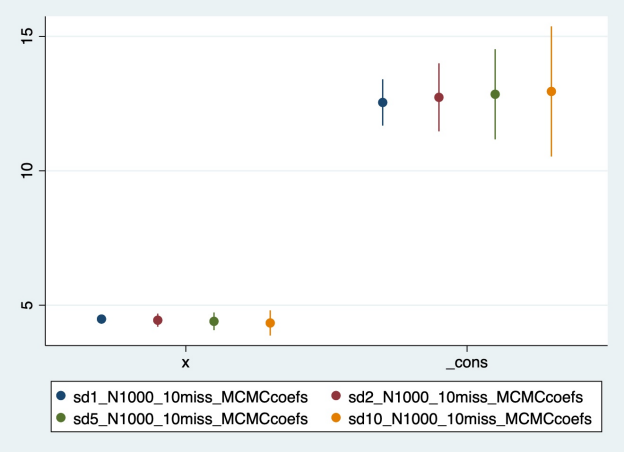

Figure 9: MI-MCMC Standard Error Increases for 10% Missing Rate, $N = 1000$

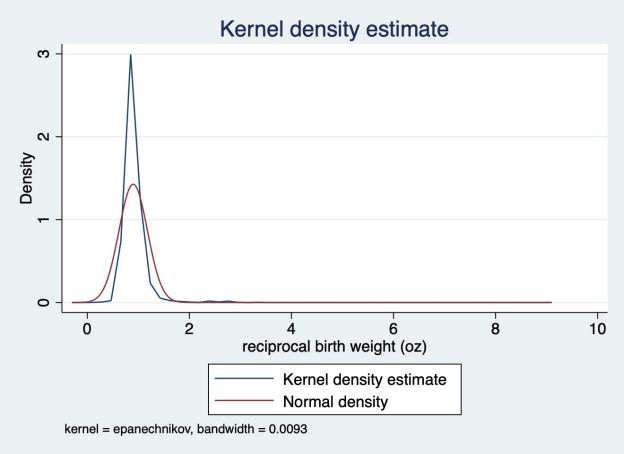

Figure 10: Density Plot for Birth Weight

6. References

References

[1] J. M. Samet, H. Wipfli, E. A. Platz, N. Bhavsar, A Dictionary of Epidemiology, Fifth Edition: Edited by Miquel Porta, American Journal of Epidemiology 170 (2009) 1449–1451.

[2] K. Benson, A. J. Hartz, A comparison of observational studies and randomized, controlled trials, New England Journal of Medicine 342 (2000) 1878–1886. PMID: 10861324.

[3] J. L. Schafer, Analysis of incomplete multivariate data, CRC press, 1997.

[4] P. Hayati Rezvan, K. J. Lee, J. A. Simpson, The rise of multiple imputation: a review of the reporting and implementation of the method in medical research, BMC Medical Research Methodology 15 (2015) 30.

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

[6] R. J. Little, D. B. Rubin, Statistical analysis with missing data, volume 793, John Wiley & Sons, 2019.

[7] D. B. Rubin, Multiple imputation for nonresponse in surveys, volume 81, John Wiley & Sons, 2004.

[8] Y. Marchenko, Multiple-imputation analysis using stata's mi command, in: Presentation given to the 2009 UK Stata users group meeting, London, UK, on September, volume 10, p. 2009.

[9] Y. Yuan, Multiple imputation using sas software, Journal of Statistical Software 45 (2011) 1–25.

[10] Y.-S. Su, A. Gelman, J. Hill, M. Yajima, Multiple imputation with diagnostics (mi) in r: Opening windows into the black box, Journal of Statistical Software 45 (2011) 1–31.

[11] R. J. Little, D. B. Rubin, The analysis of social science data with missing values, Sociological methods & research 18 (1989) 292–326.

[12] P. Zhang, Multiple imputation: theory and method, International Statistical Review/Revue Internationale de Statistique (2003) 581–592.

[13] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1977) 1–38.

[14] S. Brooks, Markov chain monte carlo method and its application, Journal of the royal statistical society: series D (the Statistician) 47 (1998) 69–100.

[15] C. Power, J. Elliott, Cohort profile: 1958 british birth cohort (national child development study), International journal of epidemiology 35 (2006) 34–41.

[16] University College London, UCL Institute of Education, Centre for Longitudinal Studies [data series]. National Child Development Study. 13th Release. UK Data Service, 2023.

[17] W. R. Gilks, S. Richardson, D. Spiegelhalter, Markov chain Monte Carlo in practice, CRC press, 1995.

[18] S. Brooks, A. Gelman, G. Jones, X.-L. Meng, Handbook of markov chain monte carlo, CRC press, 2011.

[19] J. L. Schafer, J. W. Graham, Missing data: our view of the state of the art., Psychological methods 7 (2002) 147.

[20] X.-L. Meng, Multiple-imputation inferences with uncongenial sources of input, Statistical science (1994) 538–558.