Missing Data

An Introduction

Marius Johan Franco Dioli

June 2020

Table of Contents

What is missing data?

Missing data occurs when observations are missing in a data set. When we say an observation is missing, we mean that we have not observed it. The missing observation still has a value, but this value is hidden from us for some reason.

Some basic notation

- Let the complete data matrix $Y = (y_{ij})$ be an $n \times k$ matrix, with the i^{th} row $y_i = (y_{i1}, ..., y_{ik})$ where y_{ij} is the value of variable Y_j for observation i.

Some basic notation

- Let the complete data matrix $Y = (y_{ij})$ be an $n \times k$ matrix, with the i^{th} row $y_i = (y_{i1}, ..., y_{ik})$ where y_{ij} is the value of variable Y_j for observation i.
- Let the missingness indicator matrix $M = (m_{ij})$ be an $n \times k$ matrix such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed.

Some basic notation

- Let the complete data matrix $Y = (y_{ij})$ be an $n \times k$ matrix, with the i^{th} row $y_i = (y_{i1}, ..., y_{ik})$ where y_{ij} is the value of variable Y_j for observation i.
- Let the missingness indicator matrix $M = (m_{ij})$ be an $n \times k$ matrix such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed.
- Now, given m_i, we can partition y_i into y_i^o and y_i^m, corresponding to the components that are observed and missing respectively.

Missingness mechanisms

The missingness mechanism is characterized by the conditional distribution of m_i given y_i, i.e. $f_{M|Y}(m_i|y_i, \theta)$, where θ denotes unknown parameters. We distinguish between three different missingness mechanisms by considering how M is related to Y in order of decreasing strictness of assumptions:

Missing Completely At Random

Missing Completely At Random (MCAR) occurs when, for all and any distinct values y_i and y_i^* in the sample space of Y,

$$f_{M|Y}(m_i|y_i, \theta) = f_{M|Y}(m_i|y_i^*, \phi), \quad \exists \theta, \phi$$

where θ and ϕ denote unknown parameters.

Missing At Random

Missing At Random (MAR) occurs when, for all i and any distinct values y_i^m, y_i^{m*} in the sample space of Y_i^m, we have that the following equation holds:

$$f_{M|Y}(m_i|y_i^o, y_i^m, \theta) = f_{M|Y}(m_i|y_i^o, y_i^{m*}, \phi), \qquad \exists\theta, \phi$$

Missing Not At Random

If $f_{M|Y}(m_i|y_i^o, y_i^m, \theta) = f_{M|Y}(m_i|y_i^o, y_i^{m*}, \phi)$ does not hold for some i and some y_i^m, y_i^{m*}, then the missingness mechanism is called Missing Not At Random (MNAR).

2018	2019 (Complete)	2019 (MCAR)	2019 (MAR)	2019 (MNAR
8.9	7.4	7.4	7.4	NA
9.4	7.3	7.3	7.3	NA
8.4	7.9	NA	7.9	NA
5.3	6	NA	NA	NA
11.1	7.4	7.4	7.4	NA
8.4	10.7	NA	10.7	10.7
6.4	8.8	NA	NA	8.8
5.6	3.5	NA	NA	NA
5.6	7.1	7.1	NA	NA
8.2	6.9	6.9	6.9	NA

2018	2019 (Complete)	2019 (MCAR)	2019 (MAR)	2019 (MNAR
8.9	7.4	7.4	7.4	NA
9.4	7.3	7.3	7.3	NA
8.4	7.9	NA	7.9	NA
5.3	6	NA	NA	NA
11.1	7.4	7.4	7.4	NA
8.4	10.7	NA	10.7	10.7
6.4	8.8	NA	NA	8.8
5.6	3.5	NA	NA	NA
5.6	7.1	7.1	NA	NA
8.2	6.9	6.9	6.9	NA

2018	2019 (Complete)	2019 (MCAR)	2019 (MAR)	2019 (MNAR
8.9	7.4	7.4	7.4	NA
9.4	7.3	7.3	7.3	NA
8.4	7.9	NA	7.9	NA
5.3	6	NA	NA	NA
11.1	7.4	7.4	7.4	NA
8.4	10.7	NA	10.7	10.7
6.4	8.8	NA	NA	8.8
5.6	3.5	NA	NA	NA
5.6	7.1	7.1	NA	NA
8.2	6.9	6.9	6.9	NA

2018	2019 (Complete)	2019 (MCAR)	2019 (MAR)	2019 (MNAR
8.9	7.4	7.4	7.4	NA
9.4	7.3	7.3	7.3	NA
8.4	7.9	NA	7.9	NA
5.3	6	NA	NA	NA
11.1	7.4	7.4	7.4	NA
8.4	10.7	NA	10.7	10.7
6.4	8.8	NA	NA	8.8
5.6	3.5	NA	NA	NA
5.6	7.1	7.1	NA	NA
8.2	6.9	6.9	6.9	NA

What is imputation?

Imputation refers to the substituting, or "filling in", of missing values with plausible estimates. This is done with goal of improving the subsequent analysis. Most analysis techniques take as input complete data, and so imputation complements this by creating a complete (though not *the* complete) data set.

Congeniality

When the analysis model differs substantially from the imputation model we say that they are uncongenial. Such differences might cause the subsequent analysis to be biased, or in the worst case, invalid.

Ad hoc methods

Although ad hoc is pejorative in certain circumstances, it is used here to indicate that the methods below are motivated by convenience rather than any methodological concern. Below is a simplified version of the earlier example.

	T_0	T_1	T_2	T_3	T_4	T_5
Person A	8.1	8.6	7	9	8.2	8.0
Person B	6	5.2	NA	NA	NA	NA

Complete Case Analysis

Complete Case Analysis (CCA) involves restricting analysis to entries with complete data. CCA holds an intuitive appeal, as it is both exceedingly simple to implement, and with the only seeming downside being a loss in accuracy. However, if the missingness mechanism is not MCAR, then CCA will lead to distortions in parameter estimates and their variances[Schafer and Graham, 2002].

	T_0	T_1	T_2	T_3	T_4	T_5
Person A	8.1	8.6	7	9	8.2	8.0

Mean imputation

Mean imputation involves substituting the missing values with the mean of the observed values for that variable. As stated in Schafer and Graham [Schafer and Graham, 2002], this has the effect of preserving the mean of the distribution one wishes to draw inference about, but distorts other parameters such as variance and covariance.

	T_0	T_1	T_2	T_3	T_4	T_5
Person A	8.1	8.6	7	9	8.2	8.0
Person B	6	5.2	5.6	5.6	5.6	5.6

Last Observation Carried Forward

Last Observation Carried Forward (LOCF) involves imputing the missing values for an individual with the last observed value for that individual. By replicating already existing values, LOCF has a dampening impact on the variance estimate.

	T_0	T_1	T_2	T_3	T_4	T_5
Person A	8.1	8.6	7	9	8.2	8.0
Person B	6	5.2	10	10	10	10

Regression Imputation

Regression imputation functions by regressing the variable(s) with missing entries on the fully observed variables, or subsets that are fully observed. This is the best of the ad hoc methods, as it gives us much more plausible imputations that vary more like the true data [Molenberghs et al., 2014, P. 37]. However, even in this simple case it yields imputations that tend to be less variable than the observed data. Therefore, using these imputations in inference will most often lead to underestimating standard errors and p-values. If there are non-linear relationships in the data, this method's performance breaks down.

Multiple Imputation

Multiple Imputation (MI) is the process of creating multiple distinct imputed data sets, running our analysis, and then combining the results. It functions in the following way:

1. M imputed datasets are generated through repeated random draws from the predictive distribution of the missing values, $f(Y^m|Y^o, M)$, under a particular model for missingness.

Multiple Imputation

Multiple Imputation (MI) is the process of creating multiple distinct imputed data sets, running our analysis, and then combining the results. It functions in the following way:

1. M imputed datasets are generated through repeated random draws from the predictive distribution of the missing values, $f(Y^m|Y^o, M)$, under a particular model for missingness.
2. The intended analysis is performed separately on all M datasets

Multiple Imputation

Multiple Imputation (MI) is the process of creating multiple distinct imputed data sets, running our analysis, and then combining the results. It functions in the following way:

1. M imputed datasets are generated through repeated random draws from the predictive distribution of the missing values, $f(Y^m|Y^o, M)$, under a particular model for missingness.
2. The intended analysis is performed separately on all M datasets
3. The results of the analyses are combined using Rubin's or Meng's rules to create a single MI result.

Combining rules

At step 2 of MI, we have M sets of parameter estimates and their variances: $\hat{\theta}_i, \hat{Var}_i, \quad i = 0, ..., M$. The parameter estimate is simply the average

$$\bar{\theta} = \frac{1}{M} \sum_{i=0}^{M} \hat{\theta}_i$$

and the variance associated with $\bar{\theta}$ is given by

$$T = \overline{Var} + (1 + \frac{1}{M} B) \qquad (*)$$

$$\overline{Var} = \frac{1}{M}\sum_{i=0}^{M} \hat{Var}_i$$

measures the within imputation variability and

$$B = \frac{1}{M-1}\sum_{i=0}^{M}(\hat{\theta}_i - \bar{\theta})(\hat{\theta}_i - \bar{\theta})^T$$

measures the between imputation variability. The adjustment $(1 + \frac{1}{M}B)$ in equation $(*)$ is due to the finite number of imputations. If we let $M \to \infty$ then we achieve asymptotic equality to the true variance estimate.

Strengths and weaknesses of MI

Strengths:

1. By modeling the missing variables and combining multiple imputed datasets, MI achieves better variance estimates.

Weaknesses:

1. Computationally expensive.

Strengths and weaknesses of MI

Strengths:

1. By modeling the missing variables and combining multiple imputed datasets, MI achieves better variance estimates.
2. Can model different missingness mechanisms.

Weaknesses:

1. Computationally expensive.
2. Specifying the imputation model can be difficult, for example if the there is a mix of variable types and a hierarchical structure in the data.

Strengths and weaknesses of MI

Strengths:

1. By modeling the missing variables and combining multiple imputed datasets, MI achieves better variance estimates.
2. Can model different missingness mechanisms.
3. Can capture complex relationships between the observed and the missing data.

Weaknesses:

1. Computationally expensive.
2. Specifying the imputation model can be difficult, for example if the there is a mix of variable types and a hierarchical structure in the data.
3. Improper specification of the imputation model can cause uncongeniality.

Experimental setup

We looked at a simple linear regression with a single covariate defined by the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon, \qquad \epsilon \sim N(0,1) \qquad i \in (0, ..., n)$$

. Missingness was be simulated in both X and Y. Data was generated by drawing from a bivariate normal distribution

$$[YX] \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix})$$

The true regression coefficient is then $\beta_1 = 0.8$. In our experiment we drew 1000 data points from this distribution. MCAR was simulated by removing the first p values in the data set. MAR and MNAR were simulated using the logit function. The number of imputations M was set to 20.

Experimental setup

We will look at the following metrics to evaluate the performance of MI:

1. Size of the confidence interval of β_1

We will compare the performance of MI to an equivalent analysis performed with CCA for reference. Our experiment was then conducted in the following manner:

Experimental setup

We will look at the following metrics to evaluate the performance of MI:

1. Size of the confidence interval of β_1
2. Coverage of the confidence interval

We will compare the performance of MI to an equivalent analysis performed with CCA for reference. Our experiment was then conducted in the following manner:

Experimental setup

We will look at the following metrics to evaluate the performance of MI:

1. Size of the confidence interval of β_1
2. Coverage of the confidence interval
3. Bias

We will compare the performance of MI to an equivalent analysis performed with CCA for reference. Our experiment was then conducted in the following manner:

Experimental setup

We will look at the following metrics to evaluate the performance of MI:

1. Size of the confidence interval of β_1
2. Coverage of the confidence interval
3. Bias

We will compare the performance of MI to an equivalent analysis performed with CCA for reference. Our experiment was then conducted in the following manner:

1. Draw 1000 observations from the distributions of $[YX]$

Experimental setup

We will look at the following metrics to evaluate the performance of MI:

1. Size of the confidence interval of β_1
2. Coverage of the confidence interval
3. Bias

We will compare the performance of MI to an equivalent analysis performed with CCA for reference. Our experiment was then conducted in the following manner:

1. Draw 1000 observations from the distributions of $[YX]$
2. Simulate 5, 10, 20, 40, 60, 80, and 90 percent missingness under different missingness mechanisms.

Experimental setup

We will look at the following metrics to evaluate the performance of MI:

1. Size of the confidence interval of β_1
2. Coverage of the confidence interval
3. Bias

We will compare the performance of MI to an equivalent analysis performed with CCA for reference. Our experiment was then conducted in the following manner:

1. Draw 1000 observations from the distributions of $[YX]$
2. Simulate 5, 10, 20, 40, 60, 80, and 90 percent missingness under different missingness mechanisms.
3. Create linear models for each of these 7 datasets based on CCA and MI

Experimental setup

We will look at the following metrics to evaluate the performance of MI:

1. Size of the confidence interval of β_1
2. Coverage of the confidence interval
3. Bias

We will compare the performance of MI to an equivalent analysis performed with CCA for reference. Our experiment was then conducted in the following manner:

1. Draw 1000 observations from the distributions of $[YX]$
2. Simulate 5, 10, 20, 40, 60, 80, and 90 percent missingness under different missingness mechanisms.
3. Create linear models for each of these 7 datasets based on CCA and MI
4. Repeat 1000 times

Experimental setup

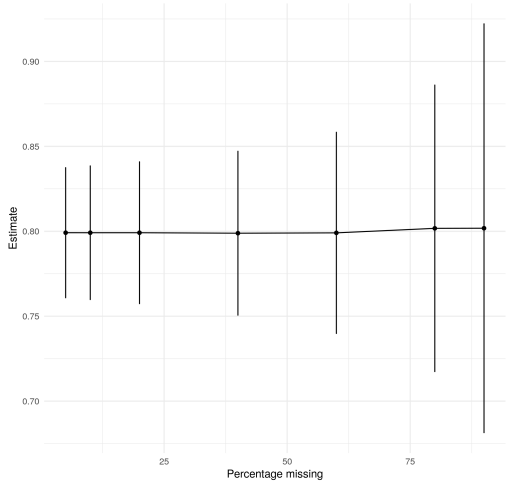
We will look at the following metrics to evaluate the performance of MI:

1. Size of the confidence interval of β_1
2. Coverage of the confidence interval
3. Bias

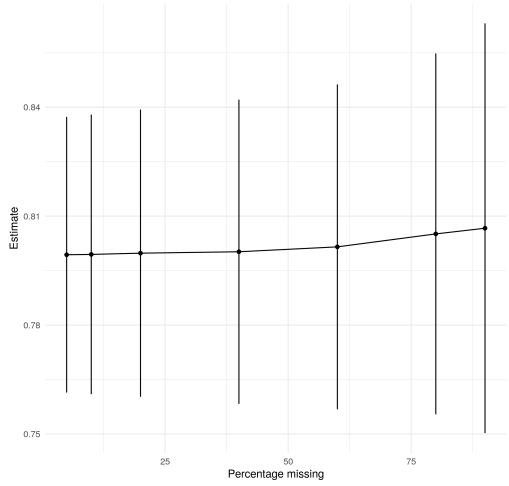
We will compare the performance of MI to an equivalent analysis performed with CCA for reference. Our experiment was then conducted in the following manner:

1. Draw 1000 observations from the distributions of $[YX]$
2. Simulate 5, 10, 20, 40, 60, 80, and 90 percent missingness under different missingness mechanisms.
3. Create linear models for each of these 7 datasets based on CCA and MI
4. Repeat 1000 times
5. Evaluate the results based on the above metrics.

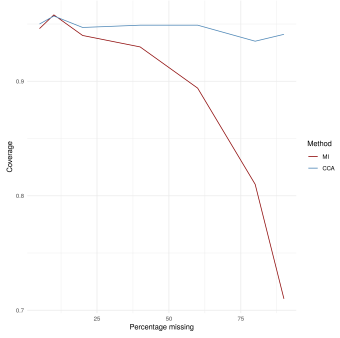
Results



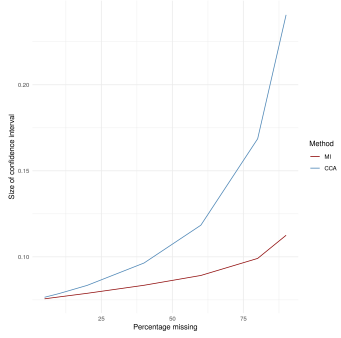
(a) *β_1 estimate with CCA*



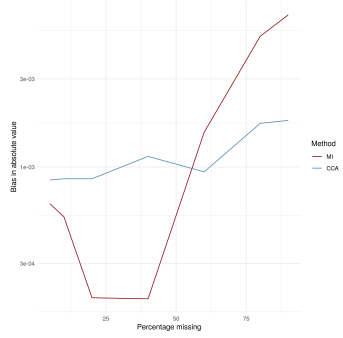
(b) *β_1 estimate with MI*



(c) *Coverage*

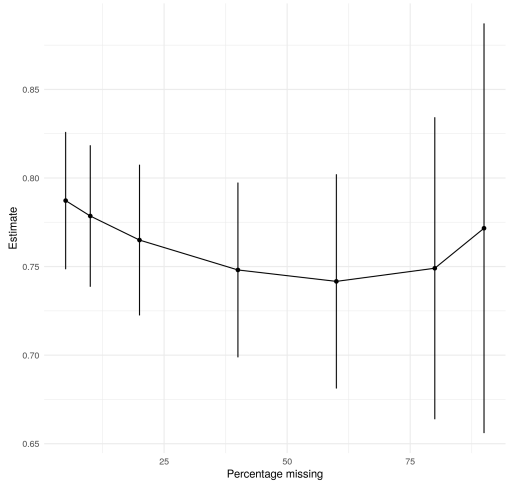


(d) *Average CI size*

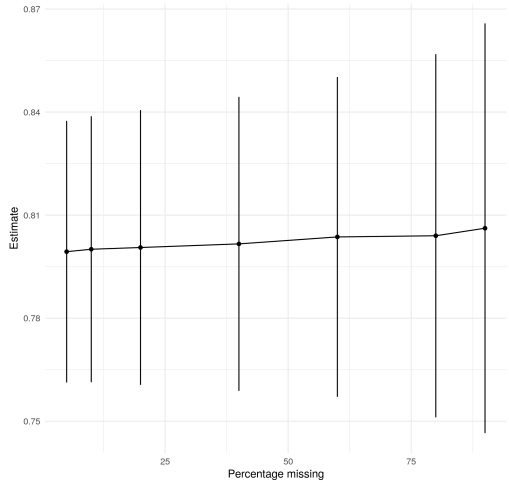


(e) *Average bias*

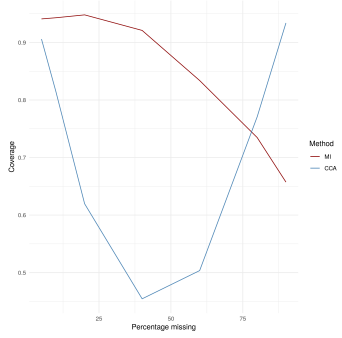
MAR in X



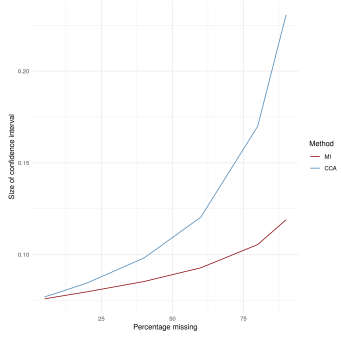
(f) *β_1 estimate with CCA*



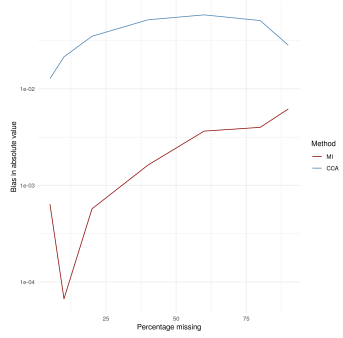
(g) *β_1 estimate with MI*



(h) *Coverage*

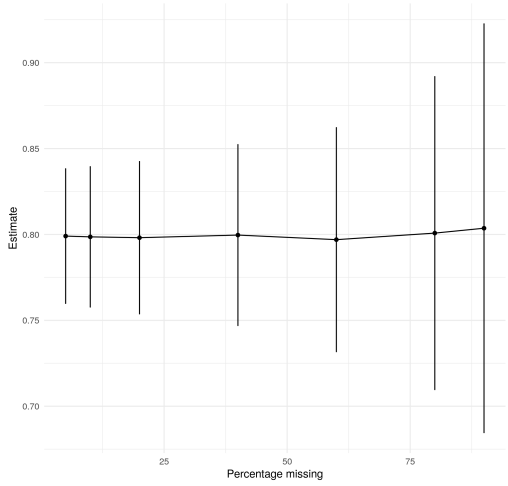


(i) *Average CI size*

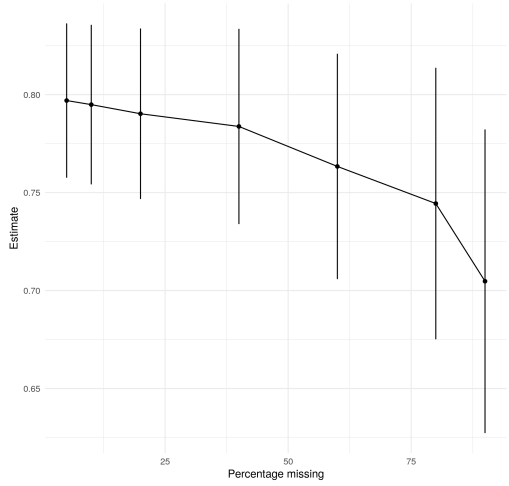


(j) *Average bias*

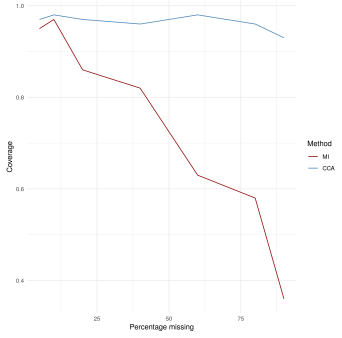
MAR in Y



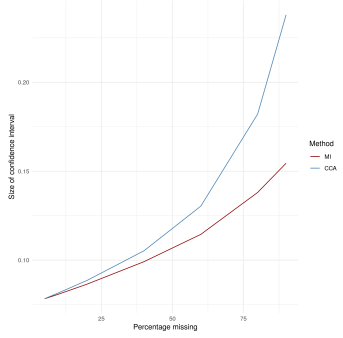
(k) *β_1 estimate with CCA*



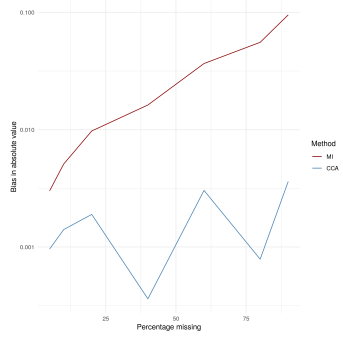
(l) *β_1 estimate with MI*



(m) *Coverage*

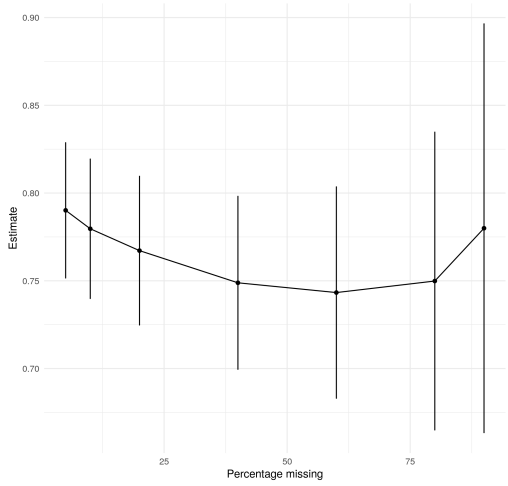


(n) *Average CI size*

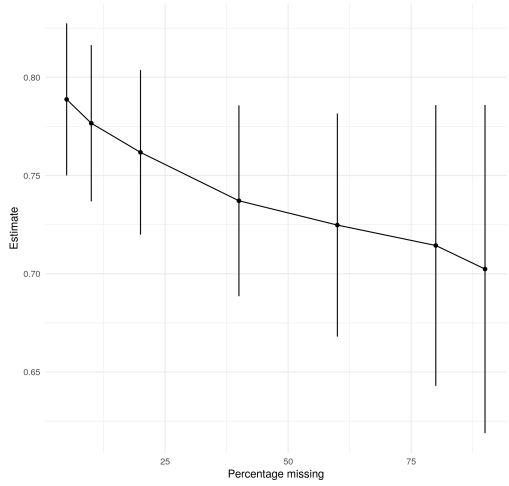


(o) *Average bias*

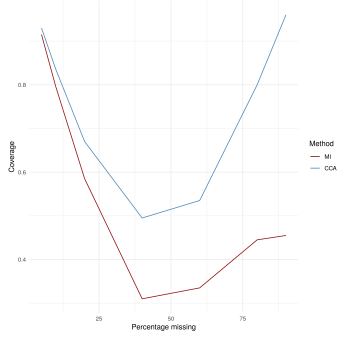
MNAR



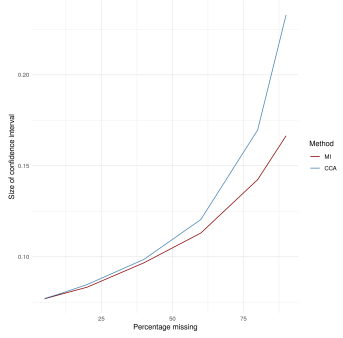
(p) *β_1 estimate with CCA*



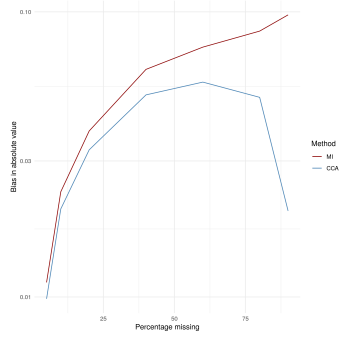
(q) *β_1 estimate with MI*



(r) *Coverage*



(s) *Average CI size*



(t) *Average bias*

Conclusion and further work

We conclude that:

1. Missing data is an important consideration in data analysis.

Conclusion and further work

We conclude that:

1. Missing data is an important consideration in data analysis.
2. Imputation can in certain circumstances help with missing data.

Conclusion and further work

We conclude that:

1. Missing data is an important consideration in data analysis.
2. Imputation can in certain circumstances help with missing data.
3. Theory and literature disagree with experimental results. Further study will be needed to verify these results.

Conclusion and further work

We conclude that:

1. Missing data is an important consideration in data analysis.
2. Imputation can in certain circumstances help with missing data.
3. Theory and literature disagree with experimental results. Further study will be needed to verify these results.

Possible future investigations could be made into the effect of MI on prediction and classification, especially in the context of machine learning models. Another interesting avenue of exploration is replacing the conditional distributions with other modeling tools such as trees or neural networks. Perhaps the messy nature of much of modern data will make the underlying assumptions of conditional distributions untenable.


Kenward, M. G. and Molenberghs, G. (2009).
Last observation carried forward: A crystal ball?
Journal of Biopharmaceutical Statistics, 19:872–888.


Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, B., Verbeke, G., Carpenter, J. R., Rizopolous, D., Davidian, M., Rotnitzky, A., Vansteelandt, S., Carlin, J. B., van Buuren, S., Goldstein, H., Hogan, J. W., Daniels, M. J., Hu, L., White, I. R., Mallinckrodt, C., Belin, T. R., and Song, J. (2014).
Handbook of missing data methodology.
Chapman and Hall/CRC, Boca Raton, FL.


Rubin, D. B. (1996).
Multiple {Imputation} after 18+ {Years}.
Journal of the American Statistical Association, 91(434):473–489.


Schafer, J. L. and Graham, J. W. (2002).
Missing data: Our view of the state of the art.
Psychological Methods, 7(2):147–177.