Incomplete Data: The Raghunathan Project

Andrew Stranberg, David Schischke and Zachary Hougardy
s4anstra@uni-trier.de, s1ddschi@uni-trier.de, s4zahoug@uni-trier.de

Objective [1]

The goal is to infer about the regression coefficients of the model $Y = \beta_0 + \beta_1 X + \epsilon, \epsilon \sim N(0, \sigma^2)$ generated from a sample of $n = 1000$ (X, Y, W) drawn from a trivariate normal distribution with

$$\text{Mean} \begin{bmatrix} 0 & 1 & 2 \end{bmatrix} \qquad \text{Covariance Matrix} \begin{bmatrix} 1 & 1 & 1.7 \\ & 2 & 1.5 \\ & & 4 \end{bmatrix}$$

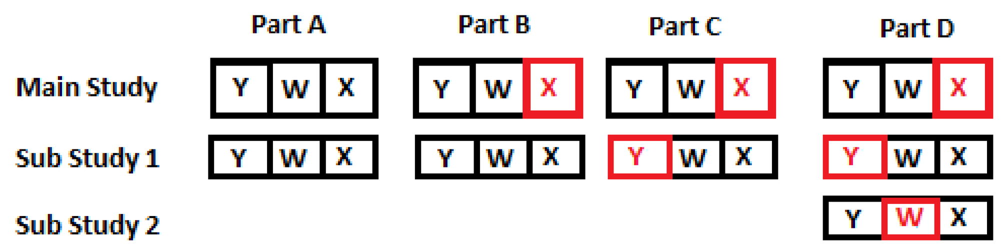
By splitting the observations into a main study ($n = 900$) and a sub study ($n = 100$), three missing data scenarios are created.

1. Delete X from in the main study
2. Delete X from the main and Y from the substudy
3. Delete X from the main, Y from the first 50 and W from the last 50 subjects of the substudy

The missing data is imputed for each of the three scenarios and β_0, β_1, σ^2 are estimated. The estimates are then compared against the base case, created using the complete data, for bias, mean squared properties, coverage rate and length of the Confidence Intervals (CIs) of the β_1 estimate. The process is simulated 250 times.

Methodology

All scenarios are tested using Predictive Mean Matching (PMM), Classification and Regression Trees (CART), Bayesian linear regression (NORM) and Weighted Predictive Mean matching (wPMM) using the MIDAStouch algorithm [2]. All methods are implemented using the mice package [3]. Even though Y is the Dependent Variable (DV), it is also used for the imputation of X.



Part A

Part A uses the full data to compute the true values of β_0, β_1 and σ^2, which are subsequently used for comparing the bias, mean square properties and actual coverage rates of β_1 for Parts B, C and D in each iteration.

Part B

Even though we have $n_{\text{mis}}(X) = 900$, we expect good performance for all imputation strategies [4]. We use 5 parallel imputations and 20 iterations.

Part C

Since there are no joint observations of (X, Y), we expect that it will be difficult to estimate the correlation between both variables correctly and hence, we expect worse estimates for β_1 [4]. We expect a sequential regression approach to work best in this case. Literature indicates that CART is well suited for this task [5]. We use 25 parallel imputations and 10 iterations.

Part D

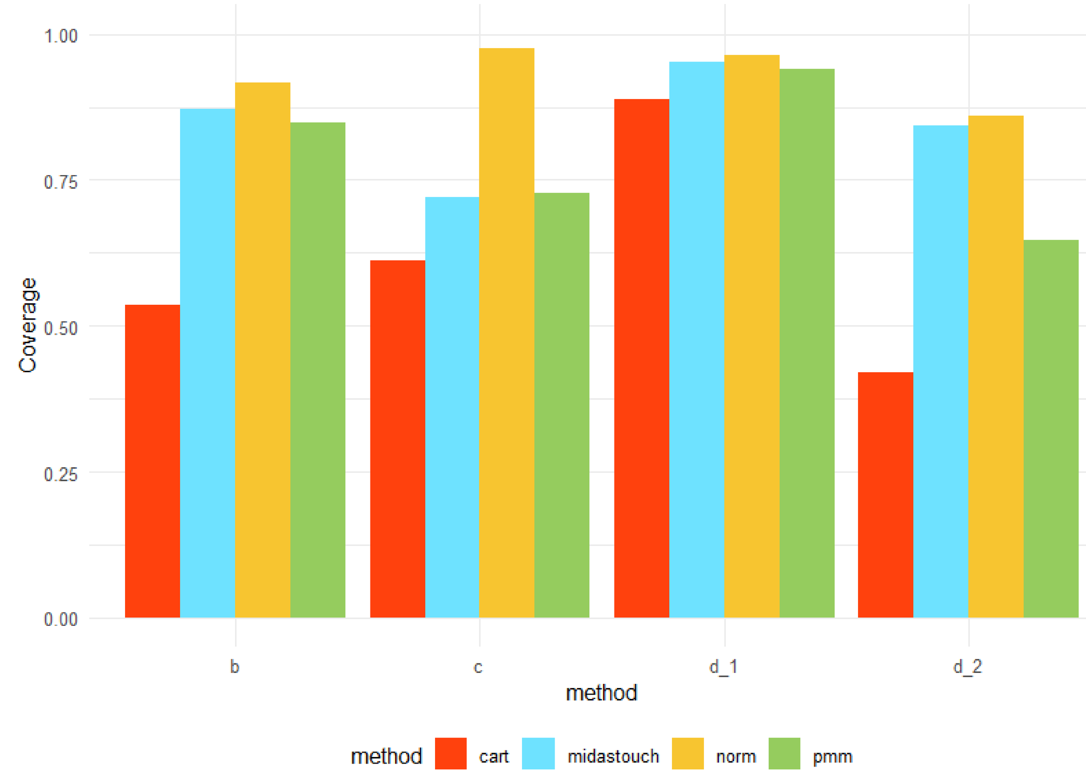
The final scenario leaves a small amount of cases for both (X, Y) as well as (X, W). Since the donor pool is quite small in this scenario, we expect wPMM to deliver the best imputation results [2]. We present both a regular (d_1) as well as a sequential approach, completing the substudy first (d_2). We again use 5 parallel imputations and 20 iterations.

Estimates and Coverage

Average CI length

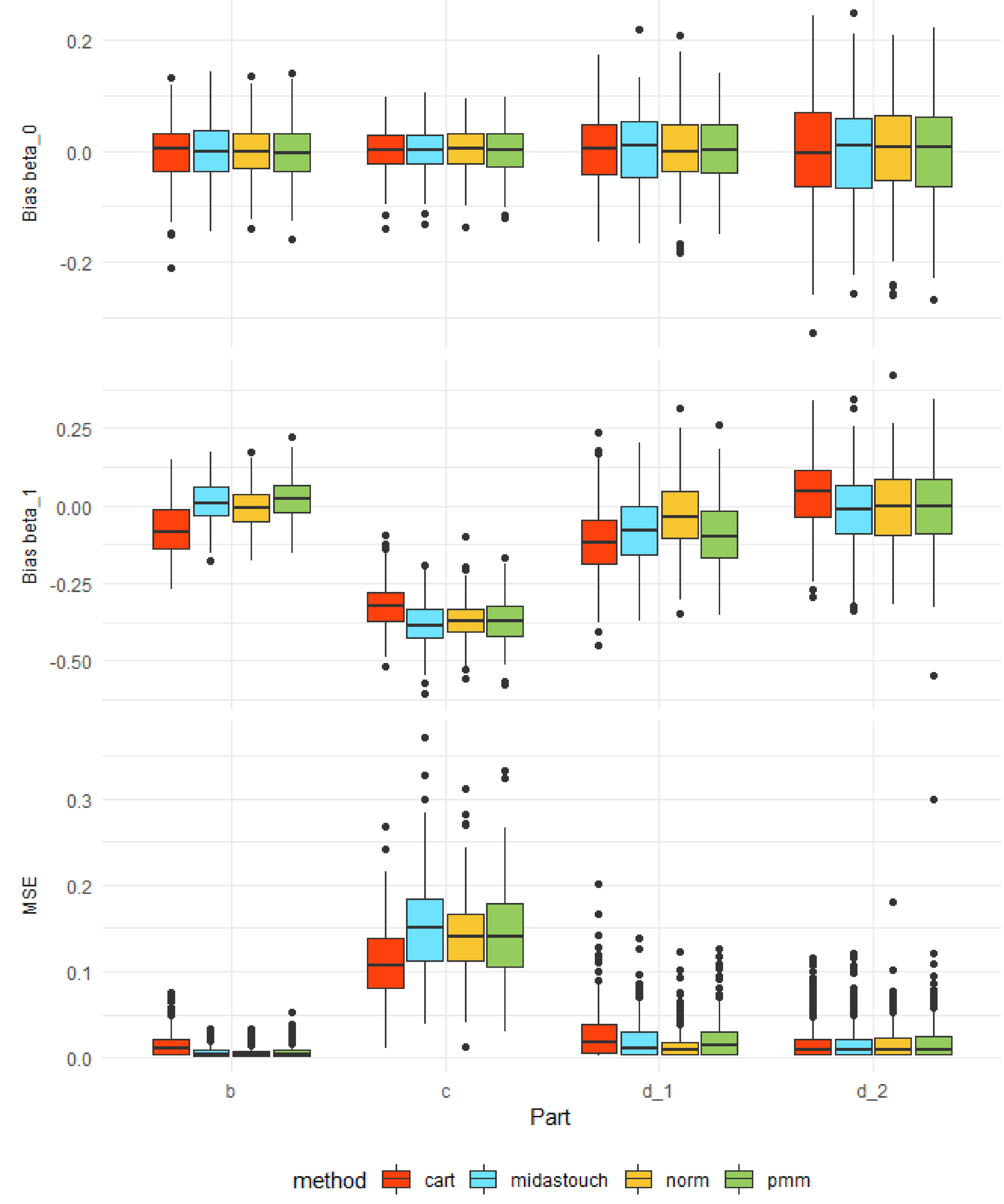
	CART	wPMM	NORM	PMM
b	0.2733	0.3120	0.3140	0.1621
c	0.8155	1.1079	0.8484	0.6440
d_1	0.6246	0.5724	0.6667	0.5338
d_2	0.2678	0.4213	0.4259	0.1553

Coverage



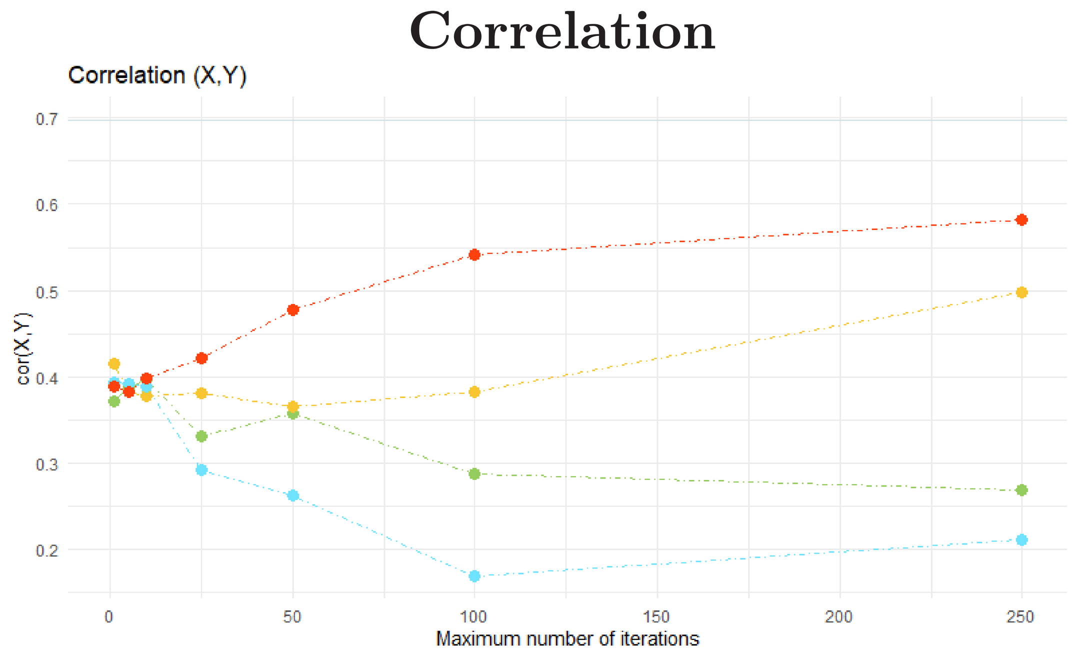
Bias and MSE Estimates

Bias and MSE

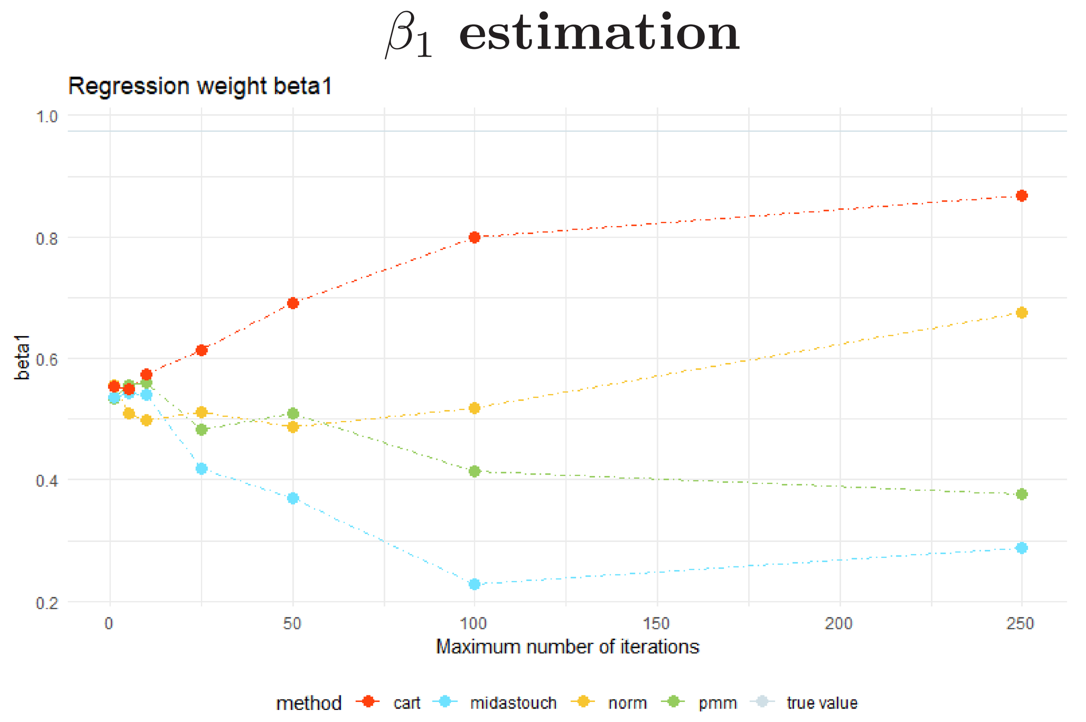


Correlation vs β_1 estimation

Correlation



β_1 estimation



Conclusion

Part B: The imputations yielded low estimation bias, MSE and good coverage for all methods excluding CART. Norm proved to be the best of the 4 methods.

Part C: CART provided the best estimates for β_1, whereas all other methods failed to estimate the correlation between (X, Y) and thus also underestimated β_1.

Part D: While the coverage of d_1 was higher than d_2, we expect this is partially due to large CIs rather than better estimation. While NORM obtained better average β_1 estimates in d_2 its MSE was actually higher (on average) than d_1.

Additional Comments:

- Good performance might be due to overestimation of CIs: Rubin's rules need to be adapted when working with fully simulated data [6]. (We have unreasonably high coverage for d_1)
- CART works better when some variables are not jointly-observed, otherwise it exhibits worse performance
- Contrary to our expectations, NORM outperforms wPMM in all scenarios, even in Part D where we have small donor pools

References

[1] Trivellore Raghunathan. Other applications. In *Missing data analysis in practice*, chapter 8, pages 155–173. CRC press, 2015.

[2] Philipp Gaffert, Florian Meinfelder, and Volker Bosch. Towards an mi-proper predictive mean matching. 2016.

[3] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[4] Stef Van Buuren. Multivariate missing data: Fully conditional specification. In *Flexible imputation of missing data*, chapter 4.5. CRC press, 2018.

[5] Lane F. Burgette and Jerome P. Reiter. Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, 172(9):1070–1076, 2010.

[6] Jerome P Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188, 2003.