Additional Supplementary Materials for “Logistic Regression with Missing Covariates – Parameter Estimation, Model Selection and Prediction within a Joint-Modeling Framework”

Wei Jiang[a,*], Julie Josse[a], Marc Lavielle[a], TraumaBase Group[b]

[a] *Inria XPOP and CMAP, École Polytechnique, France*
[b] *Hôpital Beaujon, APHP, France*

Abstract

This document presents some supplementary simulation results for the paper “Logistic Regression with Missing Covariates – Parameter Estimation, Model Selection and Prediction within a Joint-Modeling Framework” [1].

1. Simulation results varying percentage of missingness

We varied the percentage of missingness from 10% to 30% and results of bias are shown in Figure 1.

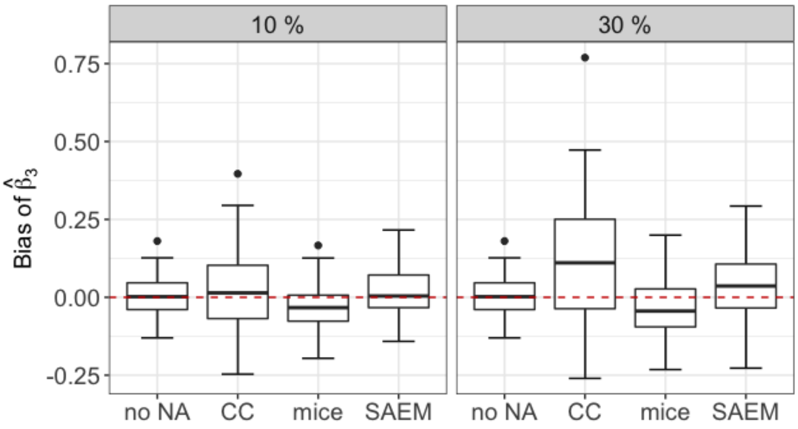
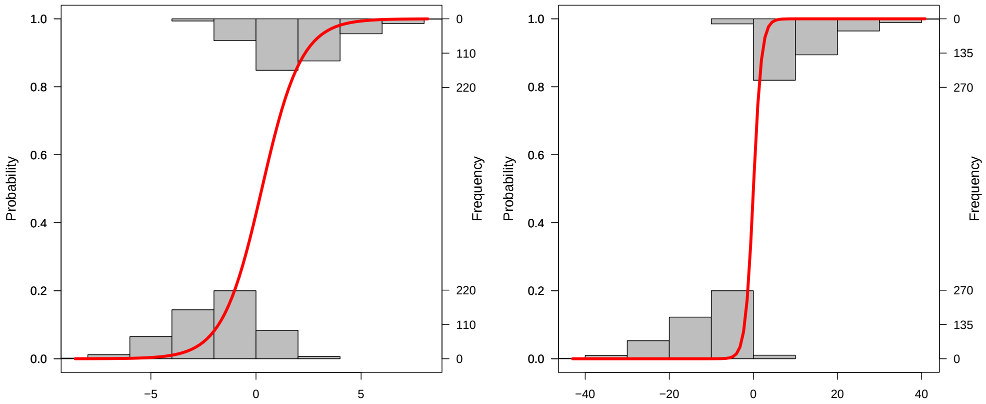


Figure 1: Empirical distribution of the bias of $\hat{\beta}_3$ obtained over 1000 simulations, varying the percentage of missingness (left: 10%; right: 30%) under MCAR, with $n = 1000$ with methods no NA, CC, mice and SAEM.

[*] Corresponding author
Email address: wei.jiang@polytechnique.edu (Wei Jiang)

2. Simulation results varying the separability of classes

We varied the separability of classes by augmenting the value of design matrix $X' = 2X$ or $X' = 5X$ to influence the link function $X'\beta$, where X is the design matrix used in the previous simulation setting in Subsection 6.1. We present the data $(y, X'\beta)$ in Figure 2 and the results of bias are shown in Figure 3. The left plots represents a case with medium level of separability, where the proposed methodology had a good performance of estimation; while the right plots shows a nearly perfect linear separability, where the performance of mice was strongly affected but the proposed method is still acceptable in comparison to the case without missing values.



(a) $X' = 2X$

(b) $X' = 5X$

Figure 2: Logistic regression $(y, X'\beta)$ plot varying the value of link function $X'\beta$.

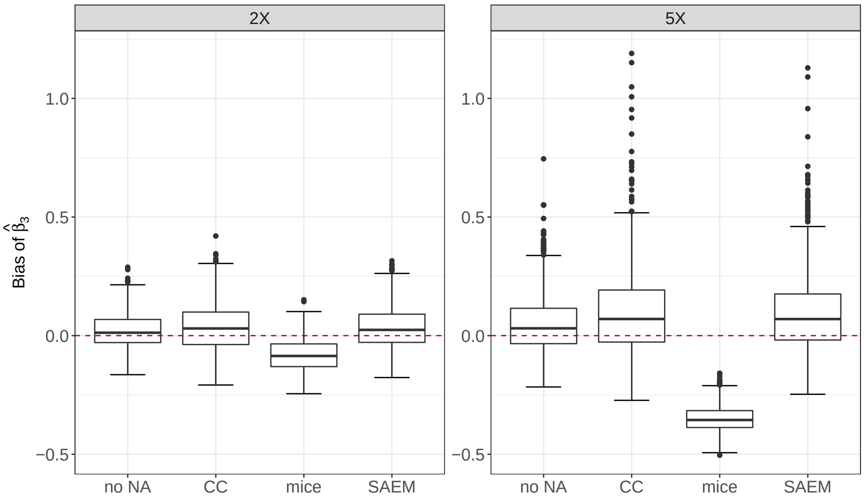


Figure 3: Empirical distribution of the bias of $\hat{\beta}_3$ obtained over 1000 simulations, varying the link function (left: $X' = 2X$; right: $X' = 5X$) under MCAR, with $n = 1000$ with methods no NA, CC, mice and SAEM.

3. Simulation results of comparison with MCEM

We generated a small sample with $n = 200$ in order to illustrate the performance of MCEM, which is computationally intensive. The bias and standard error of estimates over 100 simulations are shown in Figure 4.

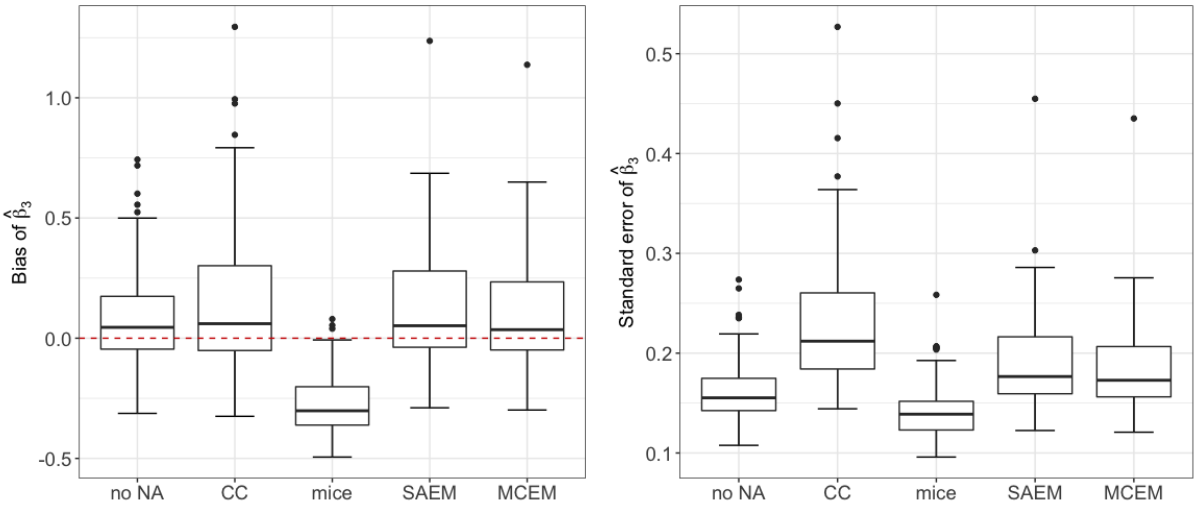


Figure 4: Empirical distribution of the bias and standard error of $\hat{\beta}_3$ obtained over 100 simulations, under MCAR, with $n = 200$ and 10% of missing values, with methods no NA, CC, mice, SAEM and MCEM.

Table 1: Coverage (%) for $n = 200$, correlation C and 10% MCAR, calculated over 100 simulations. Bold indicates under coverage. Inside the parentheses is the average length of corresponding confidence interval over 100 simulations.

parameter	no NA	CC	mice	SAEM	MCEM
β_0	96 (1.61)	96 (2.20)	97 (1.50)	96 (1.73)	96 (1.71)
β_1	98 (1.44)	95 (1.98)	97 (1.40)	97 (1.70)	99 (1.67)
β_2	97 (0.72)	96 (0.98)	96 (0.69)	97 (0.84)	96 (0.82)
β_3	92 (0.63)	90 (0.90)	**46** (0.56)	89 (0.74)	89 (0.72)
β_4	92 (0.30)	96 (0.41)	95 (0.30)	93 (0.34)	92 (0.34)
β_5	94 (0.43)	94 (0.60)	**54** (0.38)	92 (0.50)	92 (0.49)

Table 1 presents the coverage if the confidence interval for all parameters over 100 simulations and inside the parentheses is the average length of corresponding confidence interval over 100 simulations.

References

[1] W. Jiang, J. Josse, M. Lavielle, T. Group, Logistic Regression with Missing Covariates–Parameter Estimation, Model Selection and Prediction within a Joint-Modeling Framework, arXiv preprint arXiv:1805.04602 .