Simulation-based comparative performance of multiple imputation methods for incomplete longitudinal ordinal datasets

AF. Donneau, Medical Informatics and Biostatistics, School of Public Health, University of Liège
afdonneau@ulg.ac.be

Introduction

Multiple imputation (MI) is now a reference solution for handling missing data [1]. The idea is to replace each missing value not only once but by a set of M ($M > 1$) plausible values, thus reflecting the uncertainty about the prediction of the unknown missing values. The default method for MI is the data augmentation process, a Markov Chain Monte Carlo method [2], which assumes multivariate normality. For longitudinal studies with missing ordinal data, where the Gaussian assumption is no longer valid, application of the data augmentation method is questionable. In the following, consider a sample of N subjects and let Y be an ordinal outcome variable with K levels assessed on T occasions on each subject. Denote by Y_{ij} the assessment of Y on the ith subject ($i = 1, \cdots, N$) on the jth occasion ($j = 1, \cdots, T$). Associated with each subject, there is a $p \times 1$ vector of covariates, say $\mathbf{x}_{ij}$ measured at time j.

Statistical methods

- The Generalized estimating equations (**GEE**) method [3] was applied to analyze complete longitudinal ordinal data.
- Imputation mechanisms

Multivariate normal imputation (MNI):
Assuming normality, iterate between
I-step: Given an estimate for the mean and the covariance matrix, missing values are imputed by randomly drawing from a multivariate normal distribution.
P-step: New values for the mean and the covariance are simulated by drawing from a posterior distribution.
Both steps are iterated long enough to obtain a stationary Markov chain. Then, the last element of that chain is used to impute Y_{ij}^{mis}.

Ordinal imputation model (OIM):
Consider the proportional odds model $logit[Pr(Y_{ij} \leq k)|\mathbf{x}^*_{ij}] = \gamma_{0k} + \mathbf{x}'^*_{ij}\gamma$, where vector $\mathbf{x}^*_{ij}$ includes $\mathbf{x}_{ij}$, possible auxiliary covariates and the previous outcomes $(Y_{i1}, ..., Y_{i,j-1})$.
Iterate the following steps:
1. Draw new values for $\hat{\mathbf{\Gamma}} = (\gamma_0', \gamma')'$ from $\mathbf{\Gamma}^* = \hat{\mathbf{\Gamma}} + \mathbf{V}'_{hi}\mathbf{Z}$ where $\mathbf{V}_{hi}$ is the upper triangular matrix of the Cholesky decomposition of $V(\hat{\mathbf{\Gamma}})$ and $\mathbf{Z}$ is a $[(K-1)+q]-$vector of independent random Normal variates.
2. For each missing value, Y_{ij}^{mis}, compute $P[Y_{ij}^{missing} = k|\mathbf{x}^*_{ij}]$ $(k = 1, ..., K)$.
3. Impute each missing value, Y_{ij}^{mis}, by randomly drawing from a multinomial distribution with probabilities derived in step 2.

Simulation plan

1. Longitudinal ordinal data-generating model :
$logit[Pr(Y_{ij} \leq k|x_i, t_j)] = \beta_{0k} + \beta_x x_i + \beta_t t_j + \beta_{tx} x_i t_j$
with a binary group effect ($x = 0$ or 1), an assessment time (t) and an interaction term between group and time [4].
2. MAR Missing data generating mechanisms:
$logit[Pr(D_i = j|x_i, Y_{i,(j-1)})] = \psi_0 + \psi_x x_i + \psi_p Y_{i,(j-1)}$
3. Simulation patterns:
$K = 2, 3, 4, 5 , 7$ $T = 3, 5$
$N = 100, 300, 500$ Missing= 10, 30, 50%
↪ 90 different combination patterns. For each pattern, S = 500 random samples were generated.

Results - Well balanced data

Relative bias (RB %, Mean ± SD)

	MNI	OIM
β_x	89.4 ± 13.1	99.5 ± 15.5
β_t	84.6 ± 10.4	100.9 ± 8.95
β_{tx}	90.6 ± 5.73	99.7 ± 5.37

Effect of the simulation parameter on RB

		K	N	T	Missingness
β_x	MNI			↑	
	OIM	↑	↓	↑	
β_t	MNI	↑			↑
	OIM	↑	↓	↑	
β_{tx}	MNI	↑		↑	↑
	OIM	↑	↓	↑	

↑ Absolute bias increases
↓ Absolute bias decreases

- **MNI**

Binary covariate, β_x: RB was lower in long term than in short term studies (92.3 ± 12.0 % vs 86.5 ± 13.5 %; $p = 0.034$). For the **time effect**, β_t, RB decreased significantly with K ($p < 0.0001$) and with the percentage of missingness ($p < 0.0001$) but was unaffected by N and T. It decreased from 96.4 ± 5.31 % for K=2 to 76.6 ± 9.07 % for K=7 and from 90.9 ± 4.08 % for 10% of missingness to 80.2 ± 14.0 % for 50% of missingness. Similar finding were obtained for the **interaction** term, β_{tx}, except that a significant effect was also noted for T (91.7 ± 5.82 % vs 89.4 ± 5.47 %; $p = 0.007$).

- **OIM**

RB behaved similarly for each regression parameter. RB decreased significantly with K ($p < 0.0001$), as well as with T ($p < 0.05$) but increased with the sample size N ($p < 0.05$). As opposed to the MNI method, no effect was observed for the rate of missingness.

Results - Skewed data (Only short study (T=3))

- **MNI**: Except for the time effect, RB increased significantly with K (β_x: $p < 0.0001$, β_t: $p = 0.068$, β_{tx}: $p = 0.0002$) and with the percentage of missingness (β_x: $p < 0.0001$, β_t: $p = 0.57$, β_{tx}: $p = 0.0005$).
- **OIM**: No relationship was observed between RB and the modeling parameters.

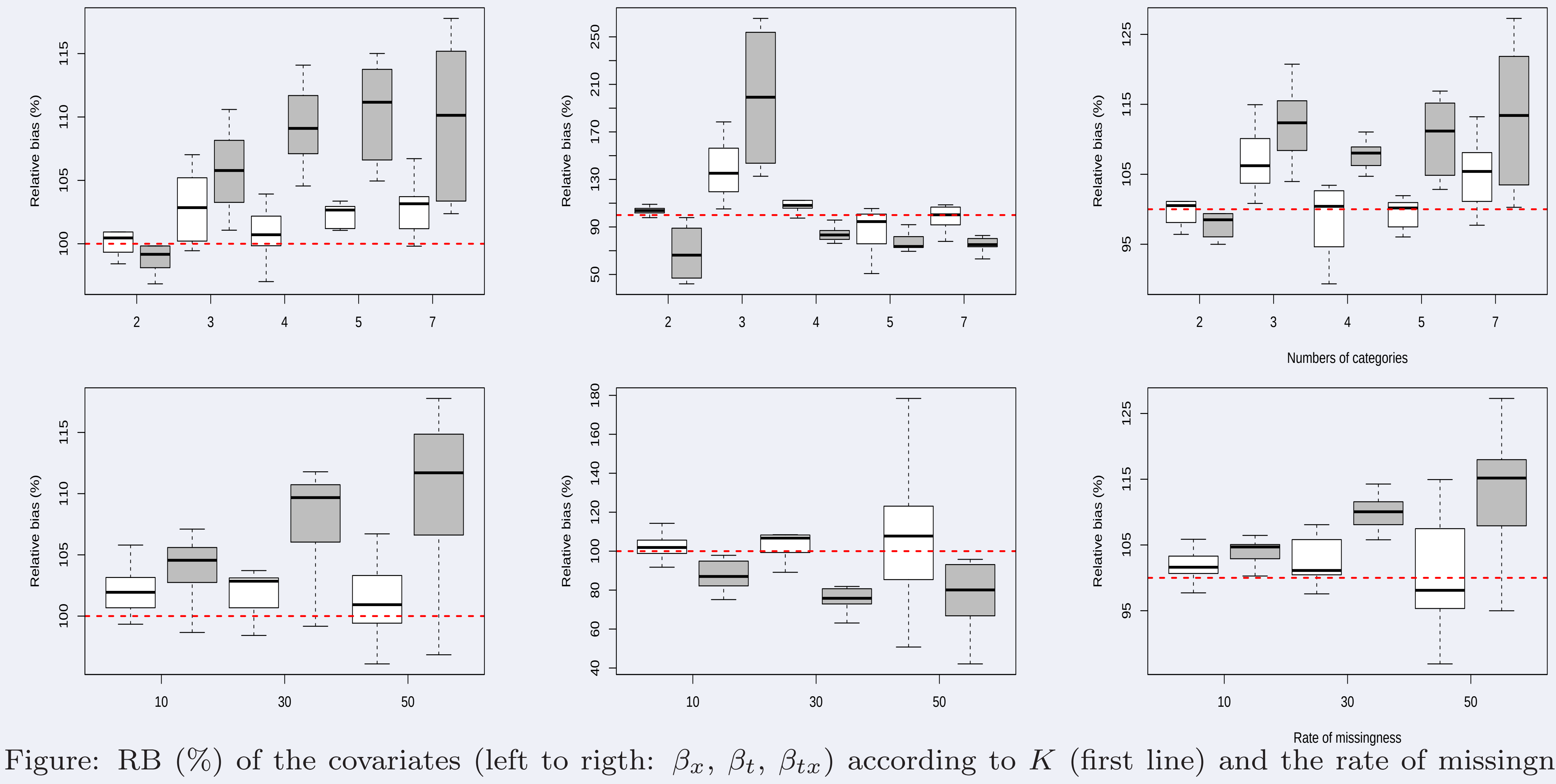

Figure: RB (%) of the covariates (left to rigth: β_x, β_t, β_{tx}) according to K (first line) and the rate of missingness (second line) (MNI= shaded boxplot - OIM=empty boxplot)

Conclusions

- Clearly, the MNI algorithm yields highly biased model parameters estimates while those derived under the OIM method are almost unbiased.
- It is suggested to impute missing longitudinal ordinal data using an appropriate method.

References

[1] Rubin, D. B. *Multiple imputation for nonresponse in surveys.* Wiley: New York, 1987.
[2] Schafer, J.L. *Analysis of incomplete multivariate data.* Chapman & Hall, 1997
[3] Lipsitz, SR., Kim, K., Zhao, L. *Analysis of repeated categorical data using generalized estimating equations. Statistics in Medicine* 1994; **13**(11):1149–1163.
[4] Ibrahim, N., Suliadi, S. *Generating correlated discrete ordinal data using R and SAS IML.* Computer Methods and Programs in Biomedicine 2011; **104**(3):122–132.