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

Anne-Françoise Donneau

Medical Informatics and Biostatistics, University of Liège

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 (MCMC) 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 For longitudinal studies with missing ordinal data, where the Gaussian assumption is no longer valid, application of the data augmentation method is questionable.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 *i*th subject ($i = 1, \dots, N$) on the *j*th occasion. Associated with each subject, there is a $p \times 1$ vector of covariates, say $\boldsymbol{x}_{ij}$ measured at time *j*.

Objective: Compare the performance of the data augmentation (MCMC) and an ordinal imputation regression model (OIM) for incomplete longitudinal ordinal data for situations frequently encountered in practice.

STATISTICAL METHODS

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

  - Data augmentation algorithm (MCMC): Assuming normality, iterate

  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 regression model (OIM): Consider the proportional odds model $logit\left[\Pr\left(Y_{ij} \le k \middle| x_{ij}^{*}\right)\right] = \gamma_{ok} + \gamma x_{ij}^{*}$, where vector x_{ij}^{*} includes x_{ij}, possible auxiliary covariates and the previous outcomes $(Y_{i1}, \dots, Y_{(i,j-1)})$.

  1. Draw new values for $\Gamma = (\gamma_{0k}, \gamma)$ from $\Gamma^{*} = \hat{\Gamma} + V_{hi}' Z$ where V_{hi}' is the upper triangular matrix of the Cholesky decomposition of $V(\hat{\Gamma})$ and Z a vector of independent random Normal variates.
  2. For each missing value, Y_{ij}^{mis}, compute $\Pr\left(Y_{ij}^{mis} = k \middle| x_{ij}^{*}\right), k = 1, \dots, K$.
  3. Impute each missing value, Y_{ij}^{mis}, by randomly drawing from a multinomial distribution with probabilities derived in step 2.

SIMULATION PLAN

- **Longitudinal data generating model** [4]: $(i = 1, \dots, N; j = 1, \dots, T; k = 1, \dots, K)$

  $$logit\left[\Pr\left(Y_{ij} \le k \middle| x_{ij}, t_j\right)\right] = \beta_{ok} + \beta_x x_{ij} + \beta_t t_j + \beta_{tx} t_j x_{ij}$$

  which incorporates a binary group effect ($x = 0, 1$), an assessment time (t) and an interaction term between group and time.
- **Monotone missing at random data mechanism**: $(i = 1, \dots, N; j = 1, \dots, T)$

  $$logit\left[\Pr\left(Drop\ out\ at\ time\ j \middle| x_{ij}, Y_{(i,j-1)}\right)\right] = \psi_o + \psi_x x_i + \psi_{prev} Y_{(i,j-1)}$$

- **Different simulation patterns**
  - Number of levels: $K = 2, 3, 4, 5, 7$
  - Number of time points: $T = 3, 5$
  - Sample size: $N = 100, 300, 500$
  - Rate of missingness: $10\%, 30\%, 50\%$

➢ In total, 90 different combination patterns, from each of which 500 random samples were generated.

SAS PROCEDURES (SAS Version 9.2)

1. **Simulation of longitudinal ordinal data:** SAS IML macro [4]
2. **Imputation mechanisms** (*M* = 20): PROC MI with
   - MCMC statement for use of the data augmentation process
   - MONOTONE LOGISTIC statement for application of the ordinal imputation regression model
3. **Analysis of the *M* completed datasets:** SAS IML macro [3]
4. **Combination of the *M* derived results:** PROC MIANALYZE

RESULTS

Relative bias (RB %, Mean ± SD)

	MCMC	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

		K	N	T	Missing
β_x	MCMC			↑	
	OIM	↑	↓	↑	
β_t	MCMC	↑			↑
	OIM	↑	↓	↑	
β_{tx}	MCMC	↑		↑	↑
	OIM	↑	↓	↑	

↑ Absolute bias significantly increased
↓ Absolute bias significantly decreased

- **MCMC**

Binary covariate, β_x, RB was lower in long than in short 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 findings were obtained for the **interaction term, β_{tx},** except that a significant effect was also noted for T (short: 91.7 ± 5.82% vs. long: 89.4 ± 5.47%; *p* = 0.0007).

- **OIM**

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

CONCLUSION

- Clearly, the MCMC data augmentation 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 imputations for nonresponse in survey*. 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 13 (11):1149--1163.1994.

[4] Ibrahim, N., Suliadi, S. *Generating correlated discrete ordinal data using R and SAS IML* Computer Methods and Programs in Biomedicine104(3):122—132, 2011.

THE POWER TO KNOW.

UNIVERSITE de Liège