Introduction:

This document ties together the three different files that make up the GRASP example for performing different multiple imputation methods to impute data with mixed types of incomplete longitudinal variables. The sample data used in this simulation study was extracted from the National Health and Aging Trend Study (NHATS), which includes four waves of observations on 5309 adults aged 65 years or older. The simulated datasets contain four time-varying incomplete variables: one binary variable, one continuous variable, and two count variables. In the simulation, we assume that the missing data mechanism is missing at random for the monotone missing data pattern and not missing at random for the intermittent missing data pattern. To simulate a monotone missing pattern, we generate the drop-out indicators for each subject, and the proportion of participants who start to drop out from the study at rounds 2, 3, and 4 are set to 20%, 15%, and 10%. To simulate an intermittent missing data pattern, we generate the missing indicators for each of the incomplete variables, and the average missing proportions of each of the incomplete variables at rounds 2, 3, and 4 are 20%, 35%, and 40%, respectively. For every simulated dataset, we conduct a multiple imputation procedure with 5 imputations using 10 imputation methods with different choices of imputation models. The imputation models include uni- or multivariate single- or multilevel generalized linear regressions, depending on implementation strategy (fully conditionals specification vs. joint modeling) and data format (wide vs. long). After imputation, we conduct three analyses: univariate generalized hierarchical model, latent growth model, and bi-variate generalized hierarchical model. We compare the estimates obtained from the multiple imputation procedure with those obtained from the original data with complete cases only. We summarize five metrics: (1) the relative bias, (2) the root of mean squared error (RMSE), (3) the 95% interval estimate width, (4) whether the interval estimate covers the estimate with complete data, and (5) the fraction of missing information (FMI) that measures the uncertainty in the imputed values for missing elements. Additionally, we record the computational time for a single imputation of each imputation model.

Keyword Categories:

Clinical: Longitudinal study, aging study
Genetics: Not Applicable (N/A)
Statistical: Multivariate missing data, multiple imputation, simulation study
Software: R
Related: Not Applicable (N/A)

References:

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Cao Y, Allore HG, Vander Wyk B, & Gutman R. Evaluation and Review of Imputation Methods for Multivariate Longitudinal data with Mixed-type Incomplete Variables. Under Review.

Component Files:

a. R program: MI_simulation_code.txt
b. R data sample: Sample_data_complete_wide.csv, Sample_data_complete_long.csv
c. PDF file explaining the entire example: MI_simulation_summary.pdf (the file you are reading)

Optimal Use

1. Read this Summary file completely; Component c is listed above.
2. Run the R program in concert with the data files; Components a & b above.

Application suggestions

The simulation studies showed that in longitudinal data with a small number of waves and a limited number of variables, when the analysis models comprise univariate regression models, FCS-standard is a computationally efficient method that results in precise and accurate estimates for both single and multilevel models. However, if the analysis models comprise multivariate multilevel models, FCS-LMM-latent is a valid statistical method that produces more accurate estimates at the cost of more intensive computations.