Introduction:

This document ties together the three different files that make up the GRASP example for performing different multiple imputation methods on the NHATS data. In this simulation study, we generated incomplete datasets following the two steps: (1) modeling the drop-out process using the original data under missing at random assumption; (2) predicting the drop-out status on data with complete cases only and removing observed values of individuals who are predicted to be drop-out at a certain time. We used two imputation strategies: fully conditional specification(FCS) and joint modeling (JM) for multiple imputation, and specified single-level or multilevel generalized linear regressions as imputation models depending on the data format and incomplete variables' type. After imputation, we performed a statistical analysis using the multiple imputed data and compared the results with the estimates obtained from the complete dataset. We summarized four metrics: (1) the relative bias $(\hat{\theta}^{meth} - \hat{\theta}^{comp})/\hat{\theta}^{comp}$, where $\hat{\theta}^{meth}$ is the estimate obtained after multiple imputation using one of the imputation methods; (2) the root of mean squared error (RMSE); (3) the average 95% interval estimate width; (4) the empirical coverage of 95% interval estimates. The sample data used in this simulation study was extracted from the National Health and Aging Trend Study (NHATS), which included four waves of observations on 5309 adults aged 65 years or older.

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. Unpublished manuscript.

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 entire example: MI_simulation_summary.pdf (file you are reading)

Optimal Use

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

Application suggestions

The simulation studies showed that FCS-LMM-latent had the best performance among all methods. Comparable performance are observed for FCS-Standard and general location model for data saved in a wide format. Using JM approach with multilevel modeling to impute the outcome variable can cause biased coefficient estimates of incomplete explanatory variables, because it does not account for the level-1 associations between the incomplete outcome and incomplete explanatory variables. FCS-GLMM methods had point estimates of regression coefficients with small biases, but it resulted in poor operating characteristics for point and interval estimates of subject-level variance when using a multilevel model for analysis.