

G37 – Research Fellows Meeting

Presenter's Abstract

20 November 2019

Nidhi Menon

Australian National University

Multiple Imputation of Missing Data in Multilevel Models

Background: Missing data are a common phenomenon in public health research. Multiple Imputation (MI) has been long recognised as an attractive approach to handle missing values. Statisticians are now advocating the use of MI as a gold standard in solving the missing data problem. Despite its early conception and its numerous advantages over the traditional ad hoc methods, there is still limited application of MI in public health research.

The theory of multiple imputation requires that imputations be made conditional on the sampling design. Not accounting for complex sample design features, such as stratification and clustering, during imputations can yield biased estimates from a design-based perspective.

Most datasets in public health research show some form of natural clustering (individuals within households, households within the same district, patients within wards, etc.). Cluster effects are often of interest in health research. These data structures are commonly observed in clinical and biostatistical settings where individuals are clustered within households, health care providers and so on. Real data example includes the three-level National Family Health Survey conducted in India with individuals nested within households, further nested within PSUs. Missing values can occur at any level in multilevel data, but guidance on multiple imputation in data with more than two levels is currently an open research question.

A random intercept model is simulated for a hierarchically clustered dataset with three levels. The outcome variable is a continuous random variable measured at the lowest level. Imputation of missing values at each level is done by using multiple imputation using chained equations (MICE) and joint modelling (JoMo). Standard multilevel regression models are fitted to the complete cases and the imputed datasets. This process is repeated 1000 times.

This study implements the Gelman and Hill approach for imputation of missing data at higher levels by including aggregate forms of individual-level measurements to impute for missing values at higher levels. The performance of popular methods of imputations, MICE and JoMo is compared to ad hoc procedures like available case analysis. Performance measures include bias in estimates, mean squared errors and probability coverage of confidence intervals. The study will highlight the strengths and limitations of imputation for variables in datasets with more than two levels.