Simulation Methods for Structural Equation Modeling with Missing Data

John W. Graham*

The Pennsylvania State University, United States. *jgraham@psu.edu*

Abstract:

Researchers heavily on simulations, most commonly Monte Carlo (MC) simulations, to address their questions, including questions relating to missing data analysis and design. With MC simulation, one starts with population parameter values and generates raw data. One then degrades the data set to produce missing values. The degraded data set is then analyzed using multiple imputation (MI), an ML approach (EM or SEM/FIML), or an older, ad hoc approach. Parameter estimates are obtained over thousands of replications, and averages are compared against known population parameter values. MC simulations are very useful for answering questions for which a closed-end solution is not possible.

Multiple Group SEM (MGSEM; Allison, 1987; Muthen, et al.,1987) is a non-MC alternative that produces ML estimates with missing data. It has been supplanted by SEM/FIML, but it has a valuable property: one operates directly on the population covariance matrix rather than on raw data, which must be generated thousands of times. MGSEM begins with different groups and covariance matrices representing different patterns of missing values. With MCAR it is easy to partition the covariance matrix into groups. Missing variances are replaced with 1; missing covariances with 0. For analysis, group 1 is set up as a typical 1-group model. Group 2 is set up as a typical 2-group model, except: factor variances, covariances, and regressions are constrained to be equal across groups; any factor loading, residual variance or covariance with data is estimated and constrained to be equal with group 1; any factor loading or residual covariance with missing data is fixed at 0; any residual variance with missing data is fixed at 1. MGSEM was applied with the two-method measurement (planned missingness) design (Graham et al., 2006), which models constructs with expensive (more valid) and cheap (less valid) measures. The design models response bias on the cheap measures, and is most efficient with complete cases on the cheap measure, but few cases with the expensive measure. Testing this design required 20,000 MC simulation replications per cell. But with MGSEM, the entire simulation took just a few analyses. The benefits in time saved, and in smoothness of the results curves were enormous. When research questions require raw data (e.g., with MI), MC simulation must be used. However, for questions that can use the population covariance matrix, MGSEM provides comparable (often better) simulation results at a small fraction of the cost in simulation time.