Dissertation Defense

Accounting for Complex Sample Designs in Multiple Imputation Using Finite Population Bayesian Bootstrap

Hanzhi Zhou

Professor Michael R. Elliott, Co-Chair
Professor Trivellore E. Raghunathan, Co-Chair
Professor Roderick J. A. Little
Professor Richard L. Valliant
Assistant Professor Brady T. West

Friday, December 13, 2013 at 10:00am

368 ISR, University of Michigan
Via video to 1208 LeFrak, JPSM at University of Maryland

Abstract

Existing fully parametric multiple imputation (MI) techniques typically assume simple random sample designs, relying on use of design-based estimators at the analysis stage to account for design effects. More complex methods which include the sample design in the formulation of the imputation model typically require strong model assumptions and expensive modeling and computation, and may still fail to be fully “congenial”. This dissertation develops an innovative multiple imputation framework (termed “two-step MI”) that accounts for complex sample designs through synthetic data generation and simple parametric models for imputation of missing values. We term this approach “two-step MI”, since we first generate posterior predictive distributions of the population that includes the (item-level) missing data, then fill in the item-level missing data using standard parametric MI techniques.

The first paper outlines the conceptual framework and develops a modified set of the standard MI combining rules for inference based on the new method. The focus is on the role of survey weights in conjunction with MI to adjust for item nonresponse. The new procedure uses a weighted finite population Bayesian bootstrap that generates posterior predictive distributions of the finite population that are free of complex design features. As a result, analysts need only to apply simple unweighted estimation methods to the imputed datasets, and, depending on the missingness mechanism, can ignore the sample design in the imputation procedure. Our simulation assuming a PPS sampling design shows that the proposed method achieves good frequentist properties in contrast to many alternative standard approaches that are not robust to model misspecification. The second paper extends the two-step MI to two-stage cluster sample design settings and develops two variations of the proposed procedure for simultaneously resolving clustering effects and “inversing” survey weights. Their performances are evaluated under a variety of simulation conditions in comparison with existing MI techniques. The third paper develops a general methodology to account for stratification effects in a highly stratified design. Quantile estimation and binary rare events data are investigated as well. Extensive analyses are conducted for real data applications on the Behavior Risk Factor Surveillance System (BRFSS), the Delta-V measure from NASS-CDS crash record data and Body Mass Index data from NHANES III.