Selecting Methods for Multiple Imputation of Missing Data

Abstract

Micha Fischer

July 14, 2022

Most data sets from sample surveys contain incomplete observations for various reasons such as respondent's refusal to answer questions. Unfortunately, most analysis tools assume complete data sets. When applying such tools to incomplete data, researchers are limited to using either complete observations or complete variables, which can have problematic consequences: biased estimates, decreased power in tests, and potential non optimal models. However, often, the challenges of missing data can be circumvented through sequential imputation (SI), an iterative procedure that imputes missing values variable by variable, generally based on the missing-at-random assumption (MAR). SI generates complete data sets that can be analyzed using standard analysis tools.

Different procedures can be used to perform SI, and each procedure can be applied in many different ways. These many options, however, can lead to subjectivity in the imputation process. Further, data is mainly analyzed with a substantive question in mind and missing data imputation might not be the primary focus of an analyst. To address these issues, previous studies compared different procedures to find the best way to apply SI. However, they often rely on one assessment strategy, e.g., simulated data only, and often compare only a small number of procedures. These shortcomings lead to findings with low generalizability. This dissertation tries to close this gap by comparing multiple parametric and non-parametric procedures for MI within the SI framework and tries to further automate and reduce subjectivity in the SI process.

Study One compares several parametric and non-parametric procedures within SI. The evaluation uses a simulation approach, analyzing data from 1) parametric models, 2) nonparametric models, and 3) a real survey data set, the publicly available version of the National Health and Nutrition Examination Survey (NHANES) data. The procedures to be compared include parametric and tree-based procedures. The first study finds that random forest and regression trees perform similarly and overall well. While Bayesian additive regression trees perform well in the non-parametric setting, it has the worst performance in the parametric scenarios investigated.

Study Two proposes a modified SI procedure in which the assessment of different procedures is automated. The study develops criteria for binary, nominal, and continuous incomplete variables to assess imputation methods within SI in an automated and objective fashion. The altered SI process is showcased in a simulation using NHANES data. This study provides methodology for a more automated SI procedure with included plausibility checks for a potential application to high-dimensional data sets with missing values, where specifying models via a human imputer is inefficient.

Study Three investigates the use and implications of incorporating response indicators (RIs) in covariates in the imputation process. This approach leads to imputation under a missing-not-at-random (MNAR) model and has been criticized in recent articles because of implausible assumptions. Despite this criticism, it is not yet apparent how widely used the approach is, because the literature stretches over multiple (sub)fields of computer science and statistics with different terminologies for the same procedure. A literature review provides insights into how to include RIs in predictors into models with different analysis goals. Furthermore, a targeted simulation showcases under which data situations and analysis goals this approach is sensible. The simulation shows that, under MAR, methods including RIs perform as well as those without them. In MNAR scenarios, methods including RIs can improve performance.