

Using Models to Inform Responsive Survey Design

Abstract

Responsive survey design (RSD) has gained increasing attention over the last decade. Although survey researchers have proposed some RSDs that rely on model predictions to optimize survey data collection during field work, several methodological challenges still need to be addressed. These challenges include incorporating incoming paradata into cost predictions, accounting for uncertainty in predictions, determining the optimal timing for implementation, and considering the quality of multiple estimates. This dissertation fills multiple gaps in the existing RSD literature via three papers. The first paper compares the ability of alternative dynamic time-to-event models to predict the number of future call attempts required until an interview or refusal. These models include a baseline model with only time-invariant covariates (discrete time hazard regression), accelerated failure time regression, survival trees, and Bayesian additive regression trees within the framework of accelerated failure time models. The number of call attempts is considered as a proxy cost indicator. Cost predictions are updated by fitting models on the training set for cases that are still unresolved at the time when the models are estimated for monitoring or intervening purposes. This approach accommodates additional paradata collected on each case up to that point. The second paper introduces a risk-conscious stopping rule that considers uncertainty in predictions of survey costs and errors. The proposed stopping rule is concerned with the problem of either maximizing data quality for a specific budget or minimizing the costs for a desired level of data quality. To implement a decision rule that stops a subset of cases in the data collection process, a survey manager not only needs to choose which set of cases to stop, but also when to stop them. Dynamically identifying the “optimal” timing for implementing a stopping rule that relies on predictions can be difficult since future outcomes are unknown during data collection. Implementing the stopping rule early may help to maximize cost savings, while decisions with reduced uncertainty can be made later as more data are collected. I analyze how the timing of implementing the risk-conscious stopping rule affects survey costs and errors at the end of data collection. The final paper proposes a multivariate stopping rule that considers the tradeoff between the data collection costs and the quality of multiple estimates. In multipurpose surveys, there may be data quality objectives that must be met for multiple estimates with a constraint on costs. The multivariate stopping rule is a weighted combination of several univariate stopping rules. Throughout this dissertation, I use real data from the Health and Retirement Study (HRS) to evaluate the cost prediction models and illustrate the proposed stopping rules.