A simulation study investigating regression techniques for respondent-driven sampling

Lisa Avery, PhD Candidate
Supervisor: Dr. Michael Rotondi

Abstract
Respondent driven sampling is a relatively new technique used to recruit participants from hard to reach populations. There are two aspects of these datasets that make undertaking regression analysis problematic: the unequal sampling probability of the participants and, the dependency between observations. A simulation study was performed to evaluate the validity of various regression models that could control for these two aspects of the data. Twelve networked populations, with varying levels of homophily and prevalence, based on a known distribution of a continuous predictor were simulated and 1000 RDS samples were drawn from each population. Weighted and unweighted binomial and Poisson general linear models, with and without various clustering controls were modelled for each sample to evaluate model validity. Type-I error rates were unacceptably high for most weighted regression models. When considering clustering, dependency within the data was in general inconsequential. Caution is warranted when undertaking regression analysis of RDS data. Even when reported degree is accurate, as in this simulation, low reported degree can unduly influence regression estimates. Unweighted regression is recommended.