Exploring the Effectiveness of Logistic Regression with Respondent-Driven Sampling Data

Katherine St. Clair (kstclair@carleton.edu), Bryan Kim (kimb2@carleton.edu), J. Liralyn Smith (smithj5@carleton.edu)

Carleton College, Northfield, MN

Introduction

Respondent-driven sampling (RDS) is a chain-referral type of sampling method primarily utilized for reaching hidden populations whose sampling frame is unknown (see, for example, Heckathorn 1997, Gile et al. 2018). With this sampling method, there is a tendency to oversample highly connected people who share similar characteristics (i.e., homophily). Much of the current research into RDS sampling methodology has focused on estimator properties for population means or proportions, while less work has been done on model performance using RDS data. In this poster, we will discuss results from a simulation study designed to analyze the estimation performance of a basic logistic regression model with RDS data. We found that estimator performance varied depending on how the population-level homophily was related to the response and explanatory variables. We also explored estimator properties of a random effects logistic model to account for clustering in a RDS sample, but this model performed worse than a basic logistic model in our study.

Respondent-Driven Sampling

The populations studied using RDS are often small but highly-connected populations that are hard to reach using conventional sampling methods. RDS sampling starts by purposefully recruiting and surveying a small number of individual, called seeds, from the population of interest. These seeds are given a small number coupons (often 3-5) that they give to other members of the population who are in their social network. Data is collected on these new recruits and they are given recruitment coupons to give to population members in their social network. These waves of incentivized peer-to-peer recruitment continues until the desired sample size is met. Figure 1 shows a simulated sample of size 100 with 7 seeds and 1-3 coupons used per recruit. Because recruitment for an RDS sample comes from an individual's social network, this sampling method can be useful when studying stigmatized populations like sex workers or drug users.

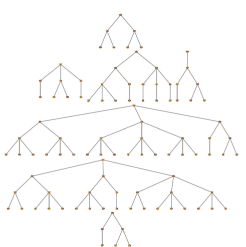

Figure 1: An example of an RDS recruitment tree with 7 seeds created using *RDStreeboot* (Baraff, 2016)

Modeling with RDS data

A sampling method where participants recruit further participants is decidedly non-random. Every recruiting participant is expected to find further participants from their existing contacts. Psychologists have noted that people tend to associate with people who have traits in common with them. This tendency is known as “homophily” and represents a significant violation of the default assumption of independence between sampled individuals.

RDS creates samples that cannot be assumed to accurately represent a random sample from the overall population. Most of the inferential work done with RDS data has focused on parameter point and interval estimation, often for proportions (see, for example, Verdery et al. 2015). Less work has been done to study how homophily affects our ability to model relationships.

Spiller (2009) proposed using a multilevel-model that includes random effects to account for the clustering of individuals who either have the same recruiter or who are in the same recruitment tree (i.e., come from the same seed). Some suggest using Generalized Estimating Equations (GEE) to account for the correlation between recruiter and recruit, while others either ignore the RDS design and use standard regression models or incorporate design weights to adjust model estimates and SE. To our knowledge, none of these approaches have been explored beyond a case study. Our research uses simulation to study the performance of a standard simple logistic regression model with RDS data.

Simulation Study

We used simulations to determine the impact of homophily from an RDS sample on estimators from a logistic regression model. Our simulation steps were:

1. Create a population of N=1000 nodes with a continuous explanatory trait (“age”). Construct a binary response trait (“health”) based on the logistic model:

$$health \sim Bern(\theta) \quad \log it(\theta) = -10 + 0.4(age)$$

2. Create a social network by adding edges (social connections) between nodes that depend on fixed homophily parameters (see below).
3. Take an RDS sample using 3 seeds, a max of 3 coupons and sample size of n=100. The number of available coupons that successfully recruit a new node is randomly determined (between 1-3). The R package RDStreeboot was used to generate the sample (Baraff 2016).
4. For the RDS sample, fit a logistic GLM for the response “health” given “age” and save estimates/SE/95% CI for the intercept and slope (age effect).
5. Repeat Steps 3-4 1,000 times.
6. Compare RDS estimates of intercept and slope to the population values (-10 and 0.4, respectively) and compare percent Bias, RMSE, and CI coverage.
7. Repeat Steps 2-6 with new homophily parameters to determine how properties from Step 6 depend on homophily.

Generating Homophily in a Population

In Step 2, we induce homophily into our population’s social network by increasing the likelihood of connection between similar nodes, either based on health, age, or both. We also created networks in which similar nodes were *less* likely to be connected, creating anti-homophily. This was accomplished as follows:

1. If x_i and x_j are the variable values of nodes i and j, then measure the distance between their values as

$$z_{ij} = \frac{|x_i - x_j|}{sd(x)}$$

2. For each pair (i,j), compute a similarity index

$$w_{ij} = \begin{cases} 1 & x_{ij} \leq 1 \\ -1 & x_{ij} > 1 \end{cases}$$

3. Repeat Steps 1-2 for the second variable so that both age and health have separate (univariate) similarity measures, w_{age} and w_{health}.
4. Fix homophily parameters α_{age} and α_{health}, then compute the probability of an edge (social connection) between nodes (i,j):

$$p_{ij} = \frac{e^{\alpha_{age} w_{age,ij} + \alpha_{health} w_{health,ij}}}{1 + e^{\alpha_{age} w_{age,ij} + \alpha_{health} w_{health,ij}}}$$

5. Use a Bernoulli draw to determine if an edge exists between nodes (i,j)

Changes to the homophily coefficients α_{age} and α_{health} make it more or less likely that similar nodes are socially connected in our simulated populations. For example, if:

- α_{age} is large and positive, nodes more similar in age are more likely to be connected
- α_{age} is near 0, no homophily exists with respect to age
- α_{age} is large and negative, nodes more different in age are more likely to be connected

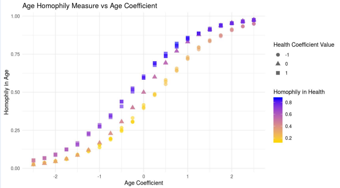

Figure 2: Homophily in age, measured as the proportion of edges that are between nodes with an age similarity index of 1, as a function of the age and health homophily coefficients. Homophily values above 0.5 indicate nodes more similar are more likely to be connected, values below 0.5 indicate the opposite.

Simulation Results

From this simulation study, there was a clear trend in bias when holding health’s coefficient value constant. As age’s homophily level increased and health’s coefficient value being held at a negative constant, bias generally decreased in value (Fig. 3, left). This relationship also held true in the opposite manner: as age’s homophily level increased and health’s coefficient value being held at a positive constant, bias generally increased in value (Fig. 3, right). When health’s coefficient value was held constant at 0, there was no clear pattern due to variation in bias across the range of age’s homophily levels (Fig. 3, middle).

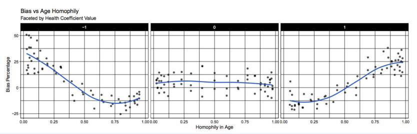

Figure 3: Bias percentages across age homophily faceted by negative, neutral, and positive health coefficients

It should be noted that the absolute value of bias percentage is the same relative to the strength of the type of homophily. For instance, bias at a homophily level measured around 0.6 (slight anti-homophily) had roughly the same magnitude as bias at a homophily level measured around 0.4 (slight homophily), albeit having different signs (respectively positive and negative).

We believe that these patterns could be products of the composition of the generated populations. Populations with the same homophily direction (i.e., both variables induced with anti-homophily or both variables induced with homophily) were found to have positive bias values, indicating that the logistic GLM estimator was systematically overestimating β_1. With an increased magnitude from matching homophily directions, more distinct groups of individuals were formed, resulting in less overlap in age for those who carried the health trait and those who did not (Fig. 4). For example, in Figure 4, we are more likely to see pockets of individuals who both have the same health outcome and are of similar ages. Consequently, based on our logistic regression model, an increase of 1 year in age would have a larger impact on the probability of an individual carrying the health trait relative to the truth in these populations.

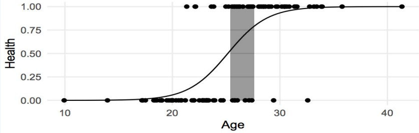

Figure 4: Example RDS sample from population with matching homophily directions

Conversely, populations with different homophily directions (i.e., one variable induced with anti-homophily and the other variable induced with homophily) were found to have negative bias values, indicating that the logistic GLM estimator was systematically underestimating β_1. With a decreased magnitude from counteracting homophily directions, less distinct groups of individuals were formed, resulting in more overlap in age for those who carried the Health trait and those who did not. For example, in Figure 5, we are more likely to see pockets of individuals who have the same health outcome (positive homophily for health) but are of dissimilar ages (negative homophily for age). As a result, in our estimated model, an increase of 1 year in age would have a smaller impact on the probability of an individual carrying the health trait relative to the truth in these populations.

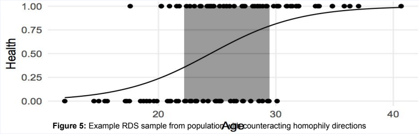

Figure 5: Example RDS sample from population with counteracting homophily directions

Simulation Results (cont.)

Root Mean Square Error (RMSE) captures the variability and bias in an estimator. We found RMSE inherently showed similar trends to those found when measuring bias. We also observed similar bias and RMSE trends for health homophily when holding the age homophily coefficient constant.

Following a negative quadratic function, coverage seemed to decline as the magnitude of anti-homophily or homophily in a variable increased, even when holding the other variable’s coefficient at any constant value (Fig. 6). We propose that these populations with low coverage values are attributed to small standard errors and high bias values stemming from the higher degree of group distinctiveness in these particular populations. When holding a variable’s coefficient constant at a value of 0, coverage was consistently around 0.95, regardless of the level of homophily in the other variable. This observation may be due to the fact that when homophily is absent in a variable, node connectedness was primarily determined by random chance, which was better accounted for in our basic logistic regression model.

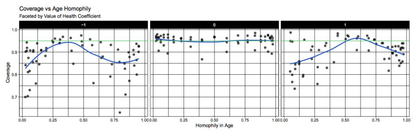

Figure 6: Coverage of confidence intervals for β_1 with a nominal level of 95% across age homophily faceted by negative, neutral, and positive health coefficients

Discussion and Future Work

Our simulations indicate that estimating the finite population regression coefficient using data collected from an RDS sample can lead to biased estimators with less than nominal confidence interval coverage rates. Bias is most extreme when there is the same “type” of homophily (homophily or anti-homophily) with respect to both the response and explanatory variables. Bias is slightly less extreme when the “opposite” homophily exists with the two variables. But, if two variables are positively related in the population, then they may be more likely have the same type of homophily direction influencing the RDS sample, leading to potentially more extreme bias in estimation.

We also obtained preliminary simulation results based on the random effects approach proposed by Spiller (2009). This model was similar to the logistic GLM, but included a random effect to account for clustering (similarities) between respondents who were recruited by the same individual. To successfully fit this model without numerical errors, we had to fix the number of coupons distributed by each recruiter at 3, rather than letting the number of coupons used vary as we originally had done. Even with this change, our simulations for a small number of homophily scenarios were prone to unstable estimation of β_1.

We are interested in studying the potential of this random effects model but make adjustments for more stable estimation. One solution would be to increase the number of coupons distributed per respondent to 5 rather than 3. Alternatively, we could increase the number of seeds and use a random effect at the recruitment tree level rather than at the recruiter level. This would result in fewer random effects to estimate with more individual observations per effect.

Acknowledgements

This simulation study was conducted as part of a 2019-2020 Senior Integrative Exercise from the Department of Mathematics and Statistics at Carleton College in Northfield, MN. Also contributing to this study was Katie Chavez, Jay Na, Aaron Prentice, and Jin Ruan. A special thank you to Mike Tie for providing computing support and Paula Lackie for assisting in data acquisition.

References

Abramovitz, D., Volz, E. M., Strathdee, S. A., Patterson, T. L., Vera, A. and Frost, S. DW. (2009). Using respondent driven sampling in a hidden population at risk of HIV infection: Who do HIV-positive recruiters recruit? *Sexually Transmitted Diseases*, 36(12):750.

Baraff, A. J. (2016). RDStreeboot: RDS Tree Bootstrap Method. R package version 1.0. https://CRAN.R-project.org/package=RDStreeboot.

Gile, K. J., Beaudry, I. S., Handcock, M. S., and Ott, M. Q. (2018). Methods for inference from respondent-driven sampling data. *Annual Review of Statistics and Its Application*, 5:65-93. https://doi.org/10.1146/annurev-statistics-031017-100704.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):174-199.

Spiller, M. W. (2009). Regression modeling using data from respondent driven sampling. Master’s Thesis. Cornell University.

Verdery, A. M., Merli, M. G., Moody, J., Smith, J. and Fisher, J. C. (2015). Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in China. *Epidemiology,* 26(5):661.