Respondent Driven Sampling

Katie Chavez, Bryan Kim, Jay Na, Aaron Prentice, Jin Ruan, J. Liralyn Smith

Goal

The goal of this paper is to give our readers a strong understanding of what Respondent Driven Sampling (RDS) is, its purpose, and how this method is used to find a population's estimates. To do so, the six people in this research group will use a previous Respondent Driven Sampling's framework as an outline for reproduction. For our study, we reproduced the previous study that uses Respondent Driven Sampling (Ramirez-Valles et al., 2005) to compare our estimates. After the comparison of certain estimates, we then describe the relationship of certain variables and estimates in greater depth through a multitude of simulations and illustrations. After the simulations, we took the previous study a step further to use an alternative model that was mentioned in the previous study, but not studied using simulations.

1 Introduction

1.1 Background

Respondent Driven Sampling (introduced by Heckathorn, 1997; see also Salganik & Heckathorn, 2004 and Volz & Heckathorn, 2008) is a "snowball" type of sampling method used to conduct studies on stigmatized populations. RDS is mainly implemented when the sampling frame is unknown due to individuals within the desired population being hidden within a larger population. It is often used for studies related to drug users, sex workers, or men who have sex with men. The design of RDS relies on individuals in each wave to drive the next wave of sampling through their social relationships and habits. This continuous "drive" is typically accomplished through a coupon method. Thus, RDS uses an individual's social relationships as a route to acquire data about hidden populations.

1.2 Finding the Sampling Frame

Since it is unlikely to find a complete representation of a stigmatized population, RDS starts with a convenience

sample of individuals that are within the target population. Typically, this sample includes between 5 and 10 individuals, but it can vary depending on the study. These individuals are the "0^{th}" wave, also called the parent seeds. Each parent seed will be interviewed to gather data and then given three coupons to distribute to individuals that are within the target population. These coupons represent the links between the individuals within the sample. If the individuals successfully give out their three coupons to individuals that are in the target population, then the individuals that recruited the next wave will receive some form of compensation. This method gives the individuals motivation to drive the next wave of recruits for the sampling frame. This revolving door of recruiting continues until the desired sampling size is met.

Because RDS utilizes a link-tracing design, RDS is a type of network sampling. Therefore, individuals are treated as nodes and their social relationships, identified by coupon exchanges, are edges that connect nodes. These social relationships are useful to investigate associations between certain characteristics. We can use the combination of nodes and edges to investigate the total number of nodes (individuals) known by a single node within the sampling population, which is also known as the node's degree. This value will be used to explore an assumption that individuals associate themselves with individuals of similar tendencies. If all nodes within the sampling frame know each other then we say that we have a connected network, but if the nodes do not completely know the sampling frame, then the network is unconnected.

1.3 Assumptions in RDS

1. Network is Undirected
2. Connected
3. Distribute Coupons at Random

The first assumption is that the network is undirected. This means relationships are reciprocal. Another assumption associated with the use of social relationships to gather data is that the network and target population are connected. RDS is wrapped around the belief that individuals associate themselves with others who are similar to themselves. This belief leads to the success of acquiring a full sampling frame. As a node recruits other nodes, they become connected through an edge. Thus, as the sampling size reaches the desired size, the nodes will all be connected in some sense. The last assumption is that nodes distribute coupons at random. Biased estimates are created when coupons are distributed non-randomly. Therefore, there is the assumption that, under the framework of Respondent Driven Sampling, the distribution of coupons between nodes is random, which leads to unbiased estimates. The term "unbiased estimate" will be unpacked further when we dive into homophily and other estimates associated with Respondent Driven Sampling.

1.4 An Empirical Example of RDS

One empirical example of RDS is Ramirez-Valles's (2005) study, in which he recruited a sample of 643 Latino men from Chicago and San Francisco who self-identified as gay, bisexual, or transgender (Latino Men Seeking Men). The main goal of his study is to examine the protective effects of community involvement in HIV for Latino men's sexual risk behavior. The two outcome variables used were whether participants have had unprotected sex in the past 12 months and their HIV status. Another researcher, Michael W. Spiller, used the same Latino Men Seeking Men (LMSM) data to explore the role of homophily in regression modeling (Spiller, 2009). For our study, we specifically investigated Spiller's method and attempted to reproduce his findings with the same data.

2 Spiller's (2009) Methodology

2.1 Clustering

Before reproducing the modeling results, we followed Spiller's method to first address the primary concern of RDS data. The primary concern of RDS data is its lack of independence among respondents. The independence assumption could be violated because participants recruited by the same person may be more similar than people recruited by different people. Spiller introduces clustering, a grouping of participants which share the same recruiter, as a solution for RDS data's lack of independence. For the LMSM data, Spiller identifies three potential clustering levels: shared recruiter, recruitment tree, and 3-digit zip code. Shared recruiter level classifies participants who recruited by the same person, recruitment tree classifies participants who could be traced back to the same parent seed, and 3-digit zip code classifies participants who were recruited from the same city. Spiller examines the levels of clustering by running one-way ANOVA tests and empty random effects models on the outcome variables. Both methods show similar results: unprotected sex and HIV status are strongly clustered at the shared-recruiter level (Table 1). For unprotected sex response variable, the one-way ANOVA tests show that shared-recruiter is the only level that displays as significant at $\alpha = 0.05$. Although clustering is observed in all three levels for HIV positive response variable, it is strongest at the shared-recruiter level shown in the result of the empty random effects models. The `Rho`, indicating the proportion of the total variance that is between clusters, is highest at the shared-recruiter level with a value of 0.332 (Table 1). Therefore, we followed Spiller's method to adjust the response variables at the shared-recruiter level when building logistic models.

Empty Random Effects Models				
Unprotected Sex		Recruiter	Tree	3-Digit Zip
	Rho	0.275	0.024	0.000
	Rho SE	0.127	0.030	0.000
HIV Positive		Recruiter	Tree	3-Digit Zip
	Rho	0.332	0.289	0.070
	Rho SE	0.082	0.099	0.062

Table 1: Spiller's results of empty random effects model tests on the response variables for clustering

2.2 Homophily

Homophily is defined as the tendency to associate with those similar to oneself. In terms of network structure, if nodes that are similar are more likely to be connected, then the network exhibits homophily. When the opposite is true, so that similar nodes are less likely to be connected, we denote the tendency as anti-homophily. Homophily has been observed by psychologists; people tend to associate with people who have things in common with them. For example, Carleton students are likely to be friends with other Carleton students. However, Carleton students are relatively less likely to be friends with people older than 90.

Spiller used the RDS Analysis Tool to determine the homophily values of each predictor variable relating to the model's dependent variable. Homophily values ranging from -1 (strong anti-homophily) to 1 (strong homophily). For the `Unprotected Sex` response variable, Spiller noted that participants' `HIV Status` was the only predictor variable that exhibited consistent non-zero homophily. For the `HIV Positive` response variable, participants' hard drug usage and relationship status were the only predictor variables that exhibited consistent homophily.

2.3 Homophily Results

We attempted to reproduce the same homophily table with the same RDS Analysis Tool used by Spiller (www.respondentdrivensampling.org). Unfortunately, we were not able to reproduce the same results due to the limitation of missing data and unspecified variables used by Spiller. For example, Spiller includes a predictor variable called `Club Drug Use` in his table. However, there are multiple variables in the LMSM data that are related to `Club Drug Use`. Therefore, we could not identify the same variable used in Spiller's analysis. Although Spiller also suggests that it is a better approach to examine homophily by site to ensure that the sample is not segregated by site, we did not split the data by city due to the limitation of time.

Unprotected Sex Homophily crossed with HIV Positive (Spiller)								
	Chicago				San Francisco			
	Unprotected=0		Unprotected=1		Unprotected=0		Unprotected=1	
HIV Positive	0	1	0	1	0	1	0	1
	-0.005	0.191	0.01	0.397	0.015	0.233	-0.303	-1

Table 2: Spiller's homophily table of unprotected sex crossed with HIV Positive using the RDS Analysis Tool

Unprotected Sex Homophily crossed with HIV Positive				
	Unprotected = 0		Unprotected = 1	
HIV Positive	0	1	0	1
	0.095	0.267	0.085	0.058

Table 3: Our results of unprotected sex crossed with HIV Positive using the RDS Analysis Tool

Considering these limitations, we still observe noticeable trends when we compare our results to Spiller's. When the homophily values were similar in the two cities from Spiller's homophily table, we also computed a similar homophily value in our results. For example, for participants who are HIV positive but did not have unprotected sex in the last 12 months, the homophily value is 0.191 from Chicago and 0.233 from San Francisco (Table 2). We get a similar homophily value of 0.267 in our result (Table 3). On the other hand, when the homophily values are different in the two cities, we get a much more moderate result. For example, for participants who are HIV positive and have had unprotected sex in the last 12 months, the homophily value is 0.397 from Chicago and -1 from San Francisco (Table 2). However, we get a value of 0.058 for the same group (Table 3). In addition, the homophily value of -1 from San Francisco seems to be an extreme statistic on the Spiller's table. Spiller suggests that a homophily value of -1 can also be an indication of instability due to small sample size. After some investigations in R, we noticed that this group of participants from San Francisco does have a relatively small sample size compared to other groups. Therefore, we conclude that we observed a much more moderate homophily value in our result because we increased the sample size by combining participants from the two cities.

Like the `Unprotected Sex` response variable, we also observed similar trends with the HIV positive response variables crossed with other predictors. For example, for participants who are HIV positive and not being in a relationship, the homophily value is 0.209 from Chicago and 0.175 from San Francisco (Table 4). We also get a similar homophily value of 0.21 in our result for the same group (Table 5). Overall, we tend to have much more moderate results than those presented in Spiller's table.

HIV Positive Homophily crossed with Relationship(Spiller)								
	Chicago				San Francisco			
	Unprotected=0		Unprotected=1		Unprotected=0		Unprotected=1	
Relationship	0	1	0	1	0	1	0	1
	-0.015	-0.025	0.209	0.194	-0.047	0.132	-0.175	0.046

Table 4: Spiller's results of HIV positive response variable crossed with relationship status using the RDS Analysis Tool

HIV Positive Homophily crossed with Relationship				
	HIV Positive = 0		HIV Positive = 1	
Relationship	0	1	0	1
	0.021	0.038	0.21	0.115

Table 5: Our results of HIV positive response variable crossed with relationship status using the RDS Analysis Tool

3 Simulation

In order to better understand the impact of non-random recruitment on the results of studies conducted with RDS, it is helpful to experiment with sampling from a population whose characteristics and connectedness we know. It is infeasible to gather such data about real populations, so we must create new, perfectly-understood populations. Because we govern the traits of each member of the population and the way that members are connected, we can study the impact of different connection parameters on the accuracy of models constructed from RDS samples taken from these custom populations. These simulations of RDS sampling on a connected population are invaluable in our study of the consequences of non-random recruitment for studies.

3.1 Population Generation

First, we create a population of imaginary individuals with a response variable, `Health`, that we model as a function of a predictor variable, `Age`. We specify the number of individuals that we want to generate, and then generate values of the predictor variable `Age` using a random generator that pulls from a normal distribution with mean of 25 and standard deviation of 5. We then generate the response variable, `Health`, by using a random number generator trained on a Bernoulli distribution with a p-parameter generated through

logistic regression. We provide a β_0 value and a β_1 for the logistic regression, so that we create a γ value by adding β_0 to the quantity β_1 multiplied by `Age`, and we then transform that γ into the p-parameter.

$$\gamma_{Health} = \beta_0 + \beta_1(Age)$$

$$p_{Health} = \frac{e^{\gamma_{Health}}}{e^{\gamma_{Health}} + 1}$$

Having created this gaggle of individuals, we then create a relationship matrix to connect members of the population to each other. To determine whether a connection exists between any two individuals, we first determine how similar they are to each other. We consider one variable at a time. For quantitative variables, we take the difference in that variable for the two individuals under consideration. We then divide this difference by the standard deviation of the variable overall, as calculated using every member of the population. If the absolute value of this quantity exceeds 1, we say that the individuals are dissimilar in this variable, and if the absolute value is less than or equal to 1, we say that the individuals are similar in this variable. For qualitative variables, we determine whether the two individuals are of the same category. If so, we say that they are similar in this variable, and if not, we say that they are dissimilar.

If we use y to denote the standardized difference between two nodes' values of some variable x, denoting the nodes as 1 and 2, then the equation is as follows:

$$y_{(1,2)} = \frac{|x_1 - x_2|}{SD(x)}$$

If we denote the similarity/dissimilarity value as w, then we calculate w in this way:

$$w_{(1,2)} = \begin{cases} 1 & y_{(1,2)} \leq 1 \\ -1 & y_{(1,2)} > 1 \end{cases}$$

We then pair each variable with some coefficient (denoted ζ to avoid confusion with the coefficients for the calculation of the log odds for the `Health` variable), which we use to weight that variable's contribution to homophily. The probability of a connection between two nodes is given by $p_{connection}$:

$$\gamma_{connection} = \zeta_0 + \zeta_1 w_{Age} + \zeta_2 w_{Health}$$

$$p_{connection} = \frac{e^{\gamma_{connection}}}{e^{\gamma_{connection}} + 1}$$

We use a Bernoulli random generator with a probability of success of $p_{connection}$ to determine whether the two nodes are connected or not. If the random generator outputs a 1, we say that the two individuals are connected, and if it outputs a 0, then the two individuals are not connected. We repeat this process for every possible pair of individuals, assuming that connections are mutual and that a person is not connected to themselves.

3.2 Measuring Homophily in Generated Populations

We wish to assess the extent to which our generated populations exhibit homophily. We thus need a metric to summarize the preferential association or non-association between nodes similar in some variable of interest. We opt to consider a proportional calculation, generating one value for each variable.

We first calculate the number of edges in the population. Then, for each variable, we determine the number of connections that are between individuals similar in that variable. The process is computationally intensive and scales linearly with the number of connections in the graph and the number of variables. We denote the homophily measure for some variable j as H_j.

We therefore assert that a homophily measure near 1 ($H_j \approx 1$ while $H_j \not\geq 1$) indicates near-perfect homophily in that variable j, and a homophily measure near 0 ($H_j \approx 0$ while $H_j \not\leq 1$) indicates near-perfect anti-homophily in j. In contrast, a homophily measure near 0.5 ($H_j \approx 0.5$) indicates an approximate absence of homophily and anti-homophily in j. It suggests that connections occur without any special regard to similarity or dissimilarity in j.

3.3 Coefficients and Homophily

In our simulations, we observe that we can adjust the value of each connection-likelihood coefficient (denoted ζ above, such that the coefficient associated with a particular variable j is denoted ζ_j) to induce homophily or anti-homophily in the generated population. A $\zeta_j \geq 2.5$ is sufficient to induce near-perfect homophily in j, while a $\zeta_j \leq -2.5$ is sufficient to induce near-perfect anti-homophily in j. A $\zeta_j \approx 0$ suffices to remove j from consideration in calculating whether two individuals will be connected, thereby inducing no homophily in j. Values of ζ_j greater in magnitude than 0 induce homophily or anti-homophily depending on whether they are greater than or less than 0, respectively. Because we measure homophily as a proportion, we see that small perturbations away from 0 have a more profound impact on the homophily measure H_j than a small perturbation away from, say, 2.5.

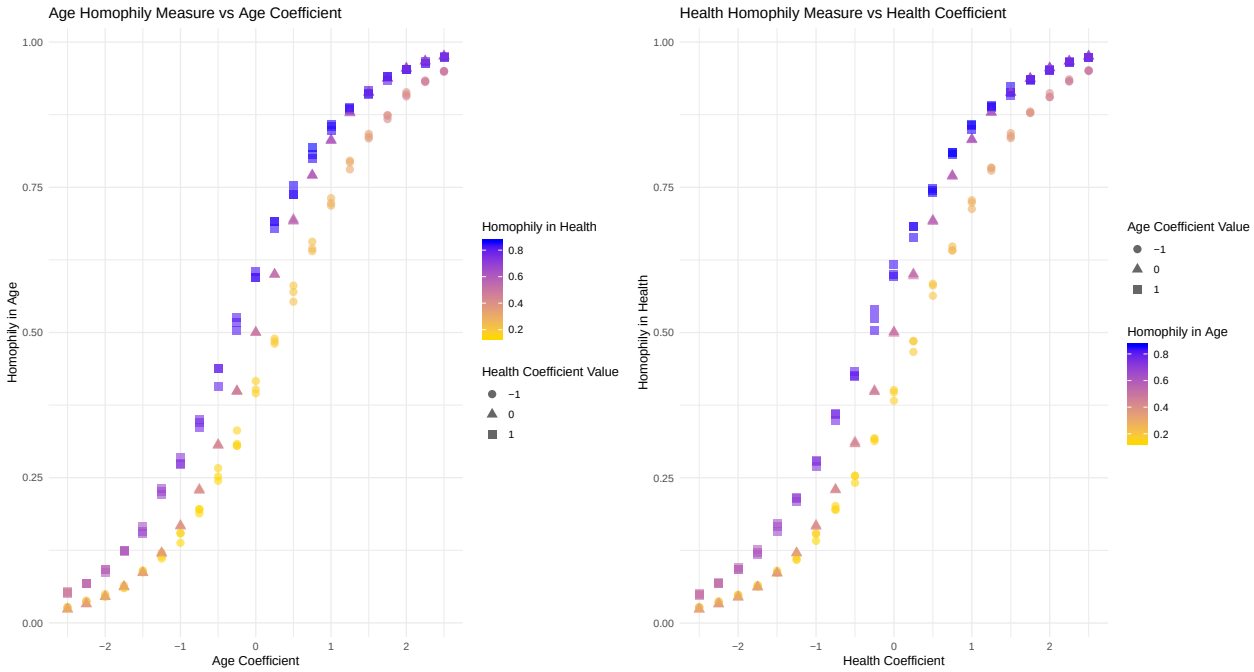

Figure 1: Coefficient-Homophily curves

These s-shaped curves confirm that we can induce homophily in a generated population by adjusting the values of the ζ coefficients that are used in connecting said population (Figure 1). Both variables' coefficients affect the homophily in a given variable, but the coefficient associated with the variable whose homophily is being measured has a stronger effect on that variable's homophily measure than the other.

3.4 Population Characteristics

In our statistical simulation study, each generated population consisted of 1,000 individuals, and the proportion of individuals who were "healthy" remained constant at approximately 0.5. In this study, the 50-50 split of individuals who did and did not carry the `Health` trait was created with a β_0 value (`Intercept`) of -10 and a β_1 value (`Age`) of 0.4. To analyze the estimator properties of a chosen statistical model, the level of homophily, determined by ζ_j (coefficients ranging from -2.5 to 2.5), in these populations was varied in one variable (`Health` or `Age`), ranging from strong anti-homophily (near 0) to strong homophily (near 1).

3.5 RDS Sampling Process

We proceeded to take a given population consisting of the desired parameters and conduct RDS using the R package *RDStreeboot* (Baraff, 2016). A single simulation took an RDS sample of 100 individuals, which was 10% of the original population (Figure 2). We determined that this sample size returned comparable results relative to larger sample sizes. The first three selected individuals within our RDS samples were always designated as the parent nodes. Known as seeds, these particular nodes comprised the initial wave of

recruitment in our simulations. With each successive wave, every node was granted a maximum of three coupons to recruit up to three others, which also paralleled Spiller's (2009) methods. For the purpose of promptly reaching our desired sample size, the number of coupons distributed by a singular node was never zero in our study; rather, there was an equal probability of $\frac{1}{3}$ that a node gave out one, two, or three coupons. Waves of recruitment were terminated when a sample size of 100 was reached. It is important to note that our simulations were implemented under ideal conditions, meaning that all of the nodes in each recruitment wave simultaneously gave out coupons to the nodes in the next recruitment wave. However, this is almost never the case in reality; there is usually some variability in how long an individual takes to recruit another individual (i.e., differing lengths of time that nodes spend to hand out coupons).

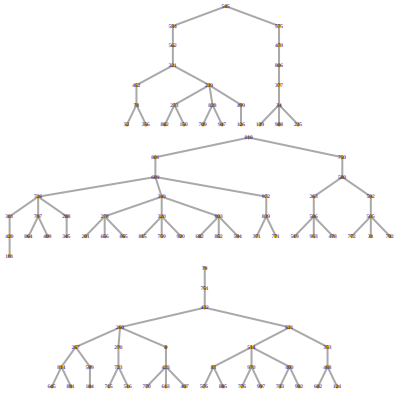

Figure 2: Tree network of an RDS sample of size 100

4 Logistic Regression Model

Once an RDS sample collected 100 individuals, we modeled this data using basic logistic regression. The model, similarly employed by Spiller (2009) in his thesis, is primarily used to estimate parameters modeling a binary dependent variable, which has a non-normal error distribution. In our study, the indicator responses of "0" and "1" were associated with "sick" and "healthy" in the `Health` variable. This generalized linear model forecasted the probability that an individual is "healthy" (`Health`) as a function of `Age`.

$$p_{Health} = \frac{e^{\beta_0 + \beta_1 Age}}{1 + e^{\beta_0 + \beta_1 Age}}$$

There have been previous evaluations of RDS point estimators, variance estimators, and violations of assumptions (Spiller et al., 2017). In contrast, we were most interested in examining the model's *coefficient* estimate of β_1 (`Age`), along with its standard error and 95% confidence interval for an RDS sample.

4.1 Post-Simulation Analysis

After 1,000 RDS samples of size 100 were taken from a population of size 1,000 and fitted with a basic logistic regression model, we examined the model's estimation performance on RDS data based on the quantitative measures of bias, root mean square error, and coverage.

$$Bias\,\% = \frac{E(\hat{\beta}_1) - \beta_1}{\beta_1} * 100$$

$$(\beta_1 = 0.4)$$

$$RMSE = \sqrt{Var(\hat{\beta}_1) + Bias^2}$$

$$CVGE = \%\, CIs\, containing\, \beta_1$$

The sample statistic, $\hat{\beta}_1$, represented the `Age` coefficient estimates in the RDS samples, while the population parameter, β_1, was the true `Age` coefficient (fixed at 0.4) that created a population with an approximately 50-50 split of sick and healthy individuals.

4.2 Simulation Results

To observe potential trends in RDS data with homophily measures, 352 unique populations with combinations of β_1 and β_2 coefficient values (ranging from -2.5 to 2.5; strong anti-homophily to strong homophily) were comprehensively analyzed in our statistical simulation study. The `Health` or `Age` coefficient value was held at a constant value of -1 (anti-homophily), 0 (no homophily), or 1 (homophily) in generated populations to gather a range of homophily levels under varying circumstances in the other variable.

4.2.1 Bias Percentage

We noticed similar trends in bias when holding coefficient values constant in either the `Health` or `Age` variables. As one variable's homophily level increased and the other variable's coefficient value being held at -1, bias generally decreased in value. This relationship also held true in the opposite manner: as one variable's homophily level increased and the other variable's coefficient value being held at 1, bias generally increased in value. When one variable's coefficient value was held constant at 0, there was no clear pattern due to variation in bias across the range of the other variable's homophily levels. However, regardless of the value of the `Age` variable's coefficient, there was extremely high bias (>50%) when there was very strong homophily (>0.9) in the `Health` variable (Figure 3).

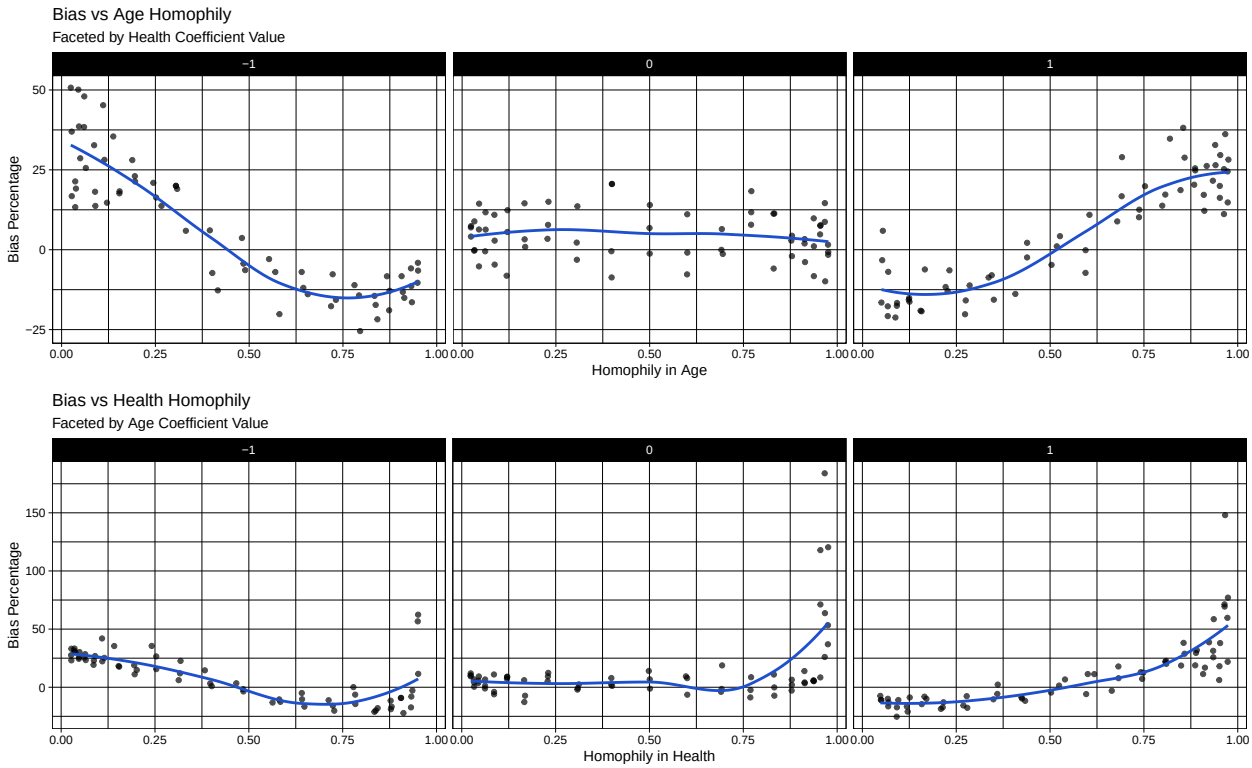

Figure 3: Bias Percentage

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 $\hat{\beta}_1$ 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. For example, with positive homophily, 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).

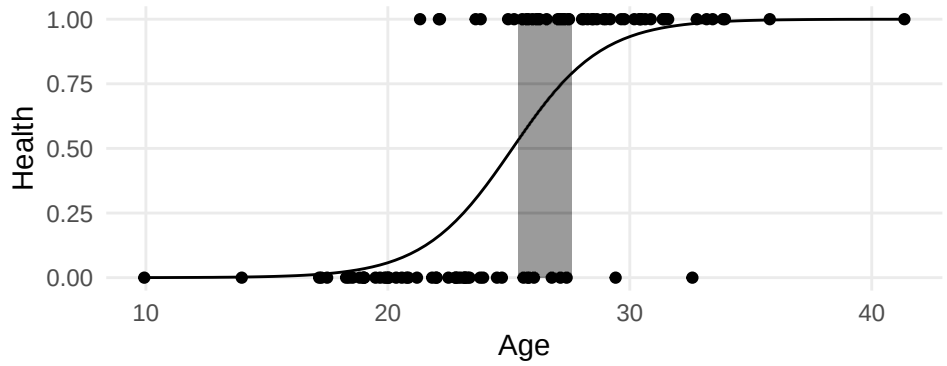

Figure 4: Example RDS sample from population with matching homophily directions

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 $\hat{\beta}_1$ 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, 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 basic 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).

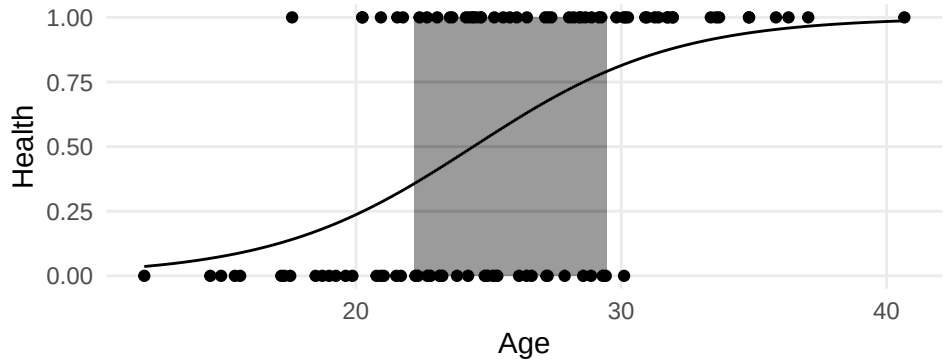

Figure 5: Example RDS sample from population with counteracting homophily directions

4.2.2 Root Mean Square Error (RMSE)

RMSE captures the variability and bias in the sample statistic, $\hat{\beta}_1$. Subsequently, RMSE inherently showed similar trends to those found when measuring bias (Figure 6).

Note that we scaled RMSE with the common logarithm in order to better visualize the relationship between RMSE and `Health` homophily levels. This was necessary due to the presence of the extreme bias values from strong health homophily cases interfering with the scaling of the data.

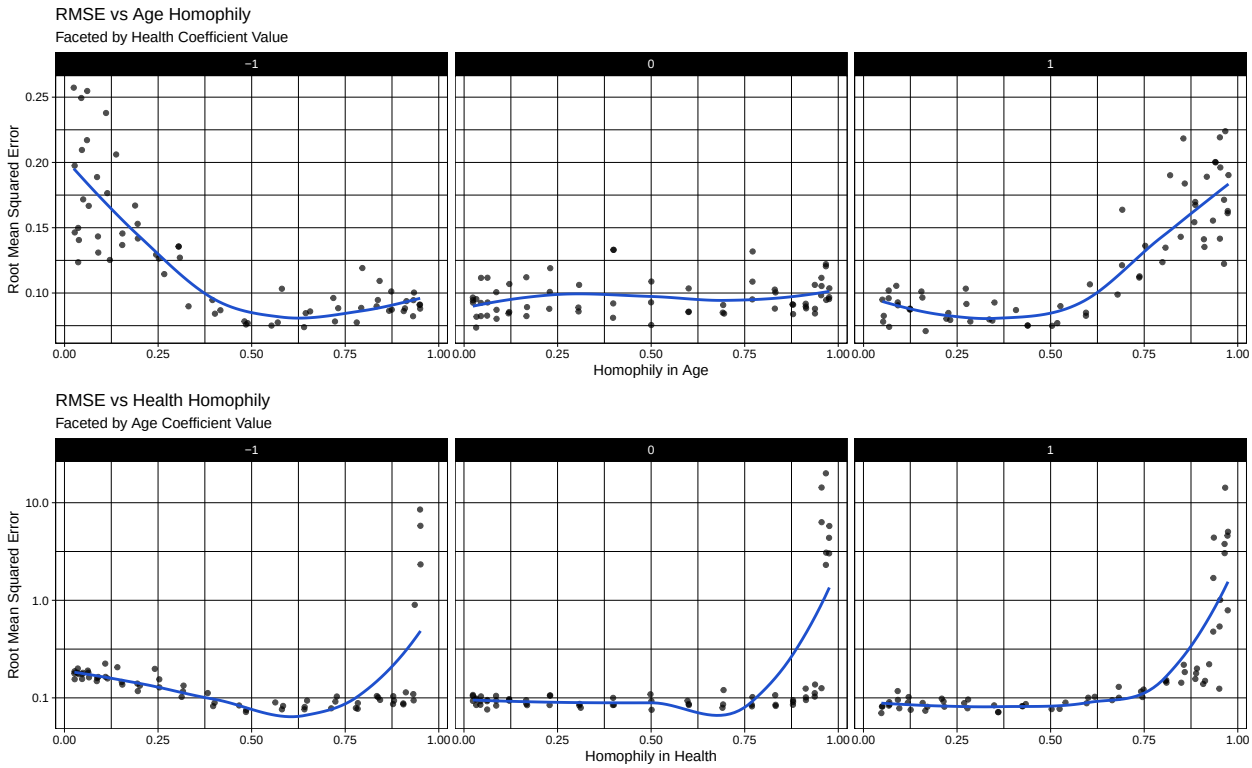

Figure 6: Root Mean Square Error

4.2.3 Coverage

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. 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 7).

5 Alternative Model

Because the basic logistic model did not work well when populations had strong homophily or strong anti-homophily, we decided to try a mixed effect logistic model. This is an alternative model that we can use in our simulations that has more than just a fixed effect of `Age` to predict the probability of a trait being

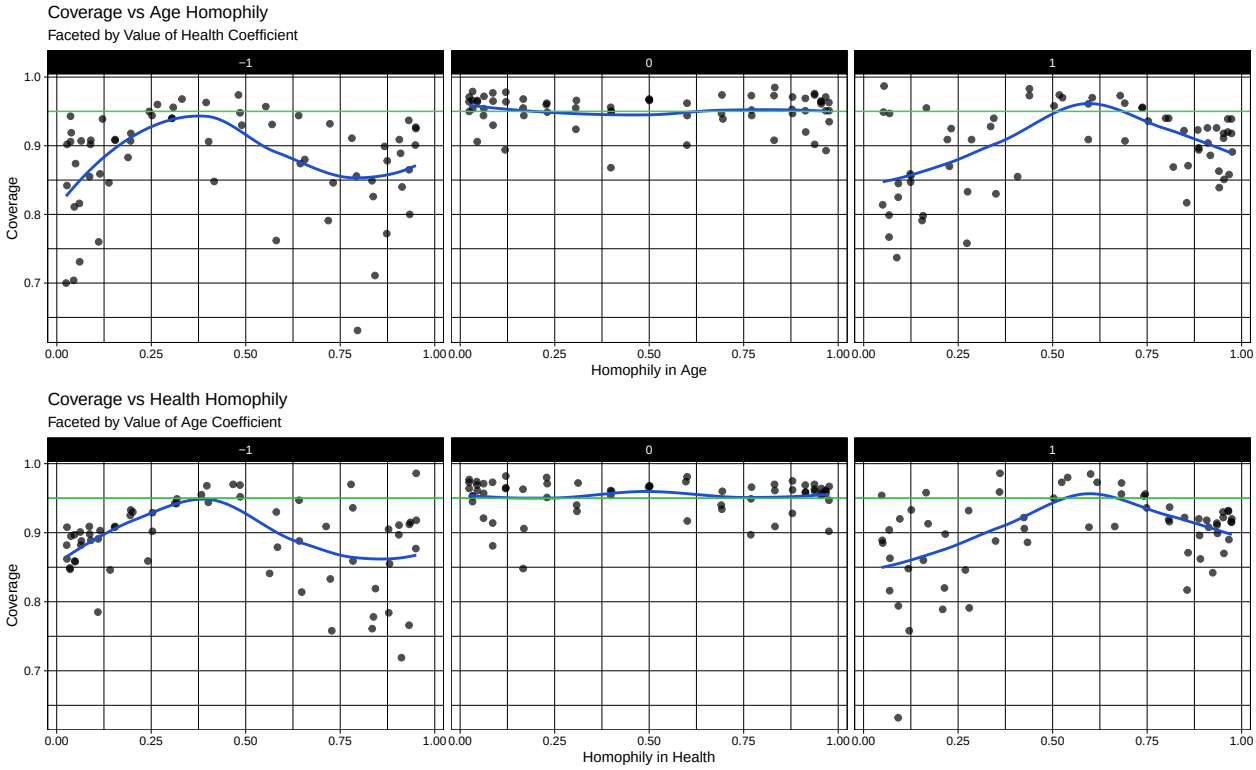

Figure 7: Coverage

prevalent. The motivation for this type of model came from the aforementioned Spiller (2009) paper, where Spiller used the model on the Latino Men Seeking Men data but did not perform simulations on it. This is because RDS data is not independent with homophily present, so any independence assumptions could be violated. What separates this model from the basic logistic regression model is that a random effect that has a normal distribution at 0 with some standard deviation σ. In our case, the random effect is put on the variable `Recruiter ID`, and it will affect the `Intercept` value. Therefore, in this model, people with the same age will have a different probability of `Health` if they were recruited by a different participant.

$$p_{Health} = \frac{e^{(\beta_0 + \alpha_{Recruiter}) + \beta_1 Age}}{1 + e^{(\beta_0 + \alpha_{Recruiter}) + \beta_1 Age}}$$

$$\alpha_{Recruiter} \sim N(0, \sigma)$$

5.1 Logistic vs Mixed Effect Model

We first wanted to run both logistic and mixed effect models within a population to compare when and why we might use one model over the other. In our simulations, we were able to control how much homophily was induced in the population and know our true odds ratio exactly, so it was clear to see the comparisons

between the two. We initially took one sample from a population we induced with high homophily and another sample from a population with zero homophily. We then calculated the odds ratios of the predicting variable from the maximum likelihood coefficient estimates we computed using the *lme4* package in R (Bates et al., 2015). The odds ratio is a way to quantify the strength of association between two variables, which in this case is `Health` and `Age`. The true odds ratio in our simulated data is 1.49, meaning that a year increase in `Age` is associated with an increase in the odds of having the `Health` trait by a factor of 1.49.

Spiller claims that adjusting for homophily with a mixed effect model can cause significant shifts to the coefficient values, which can be shown through their odds ratios. When a population exhibits high homophily that can strongly affect the outcome given the predicting variables, we see significant increases in our odds ratios. When homophily is not present, we see that the logistic model and mixed effect model produce similar results. In relation to our models from our simulated data, we see in Table 6 that a mixed effect model from the sample with no homophily yielded a slightly higher odds ratio than the logistic model, but the results were relatively similar. In contrast, we see a much more significant increase in the odds ratio in the mixed effect model from the sample with high homophily, demonstrating the kind of shift that can occur when homophily is present.

All models suggest `Age` is strongly associated with the `Health` outcome variable. It also shows us that the logistic models have odds ratios closer to the true odds ratio of 1.49 than the mixed effect models in both scenarios (Table 6). Through further analyses, we see that although the mixed effect models are further off, they are still very conducive with our results since they fit within two standard errors of the true value. However, this does suggest that there will be situations like this where simpler models yield more accurate estimations.

	Sample w/ Homophily		**Sample w/o Homophily**	
Model	*Logistic*	*Mixed Effect*	*Logistic*	*Mixed Effect*
Odds Ratio	**1.51**	**1.62**	**1.68**	**1.72**
Standard Error	0.09	0.13	0.1	0.13

Table 6: Modeling samples from simulated populations with a true odds ratio of 1.49

5.2 Reproducing Spiller's Models

Spiller believed in the idea that simpler was almost always better. He first proposed adjusting for homophily by using a simple logistic regression model, then treating the respondent's recruit's values as homophilous variables, thereon predictors to the model. We attempted to reproduce these results using the LMSM data

we obtained (Table 7). We see here that the respondent's gay identification and age are highly correlated to the outcome variable. When we examine the different potential recruiter-level predictors (i.e. recruiter's HIV status and relationship status), we also see that the recruiter's HIV status predictor variable yielded a significantly high odds ratio, or strong association, with the HIV status outcome variable. And so, if we were to estimate our predictions using logistic regression models, we would retain the `Recruiter HIV Positive` variable in the model, which would account for the homophily within the population.

Spiller	*Us*			*Logistic Models*					
Variable		**Mixed Effect Model**		**Model 1**		**Model 2**		**Model 3**	
Self ID (Gay ID)	**Odds Ratio**	**2.46**	**3.56**	**3.03**	**2.86**	**2.46**	**2.48**	**2.80**	**2.77**
	Standard Error	0.63	0.25	0.08	0.27	0.63	0.27	0.73	0.27
Relationship		**0.57**	**0.63**	**0.63**	**0.61**	**0.57**	**0.60**	**0.59**	**0.61**
		0.13	0.21	0.14	0.23	0.13	0.24	0.13	0.23
Income		**0.77**	**0.71**	**0.74**	**0.72**	**0.77**	**0.72**	**0.76**	**0.72**
		0.05	0.06	0.04	0.06	0.05	0.06	0.04	0.06
Age		**1.64**	**1.82**	**1.71**	**1.84**	**1.64**	**1.79**	**1.68**	**1.82**
		0.15	0.07	0.16	0.10	0.15	0.10	0.16	0.10
Age squared		**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Recruiter HIV positive						**2.59**	**2.97**		
						0.59	0.24		
Recruiter Relationship								**0.66**	**0.68**
								0.14	0.23

Table 7: Modeling Spiller's (2009) approaches to adjust for homophily

The other approach he suggested to adjust for clustering was the mixed effect model in which we would add a shared recruiter random intercept to our earlier model and compare coefficient results. As you can see from both Spiller's and our reproduced results, there was no significant shift in coefficient values when running the mixed effect model against a logistic model. Based on this observation and with the idea that simpler is almost always better, both Spiller and us concluded that we would reject all models that include random intercepts.

It is important to note that the LMSM data we used is the same data Spiller used in his study. We were able to produce very similar results, however they do slightly differ for two main reasons. First, Spiller did not specify in his paper exactly how he decided to aggregate the categories within the `Gay ID` variable, and so we had to group them together as we saw fit. As a result, some respondents may have been grouped differently in our tests. Secondly, a few of the predictor variables that Spiller included in his models were renamed since

his 2009 publication and could not be found. Subsequently, we included less variables than Spiller in our models, which resulted in some predictor variables yielding more weight and higher odds ratios.

5.3 Simulation Results

We then used the mixed effects model for simulating on the populations we created. The populations generated for our mixed effect logistic model used very similar parameters to the populations that were generated for our generalized linear models. The one difference in this model is that every participant recruited 3 volunteers to take part in the sample with 100% success. We did this because we wanted each participant to have enough recruits to allow for the random effect to be estimable by the mixed effect model. If a participant only handed out one coupon, then a random effect could not be applied to the recruiter. The rest of the parameters remained the same when generating our 32 populations for the random mixed effects model, for which we ran 1,000 simulations of sample size 100. After the simulations were completed, we then calculated the same post-simulation statistics that we used when simulating with the basic logistic regression model.

Since the homophily in `Health` had such a strong correlation with the post-simulation statistics, we decided to eliminate outliers and only look at populations with a `Health` homophily coefficient of 0. Therefore, any change in the homophily in `Health` measure is a byproduct of the change in the homophily in `Age`.

5.3.1 Bias Percentage

There does not appear to be an obvious pattern between homophily in `Age` and bias, and the points appear to be randomly scattered in Figure 8. However, the relationship between homophily in `Health` and bias as a percentage of the true value of the coefficient estimate of `Age` can be displayed as a parabolic curve. The bias is high when homophily and anti-homophily are very strong. As the bias starts approaching 50% or lower, the homophily measure on `Health` gets closer to random chance.

5.3.2 Root Mean Square Error (RMSE)

First, we see that root mean square error against homophily in `Age` in Figure 9 looks very similar to the previous figure with bias. Again, there is no discernible pattern, and the relationship appears to be randomly scattered. Also, for root mean square error against homophily in `Health`, we get a very similar parabolic pattern to the figure with bias. This is sensible because bias is a component of computing root mean square error, so it is reasonable that both post-simulation statistics would have the same pattern. Also, the root mean square error values are higher in strong homophily and anti-homophily cases for `Health` than they are in `Age`.

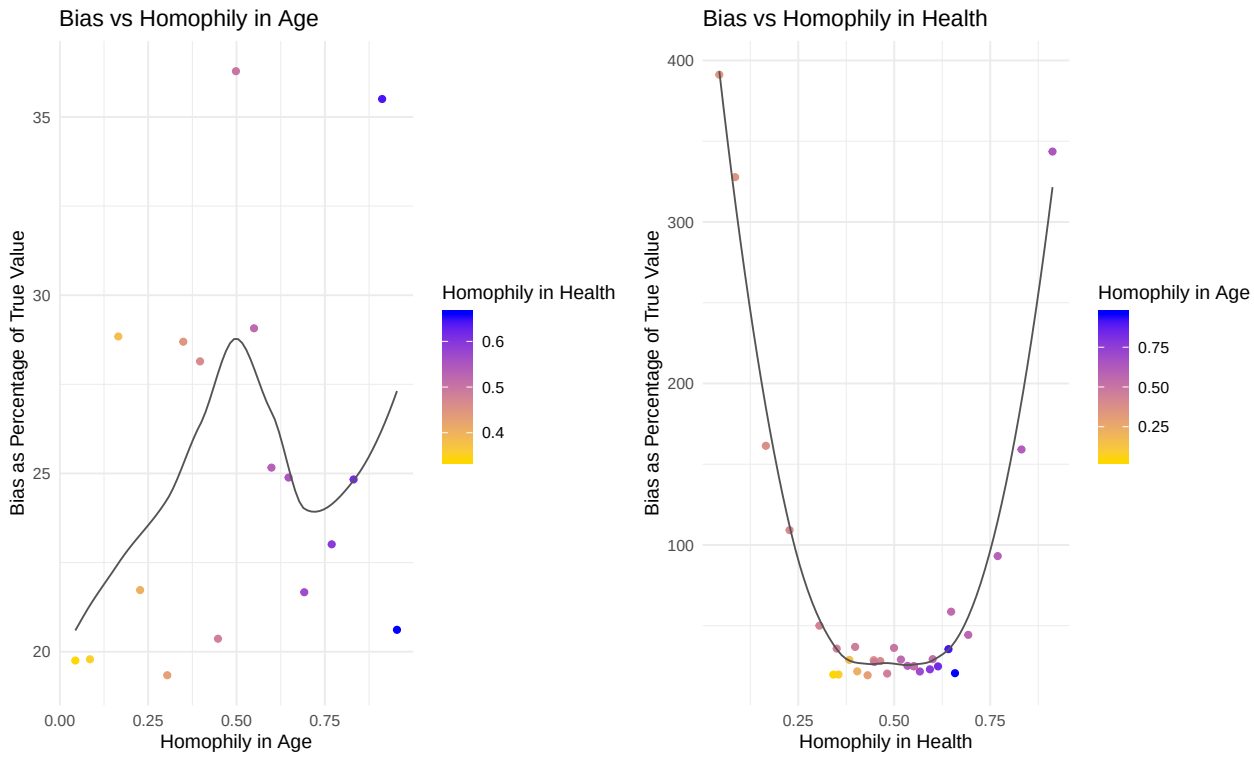

Figure 8: Bias Percentage

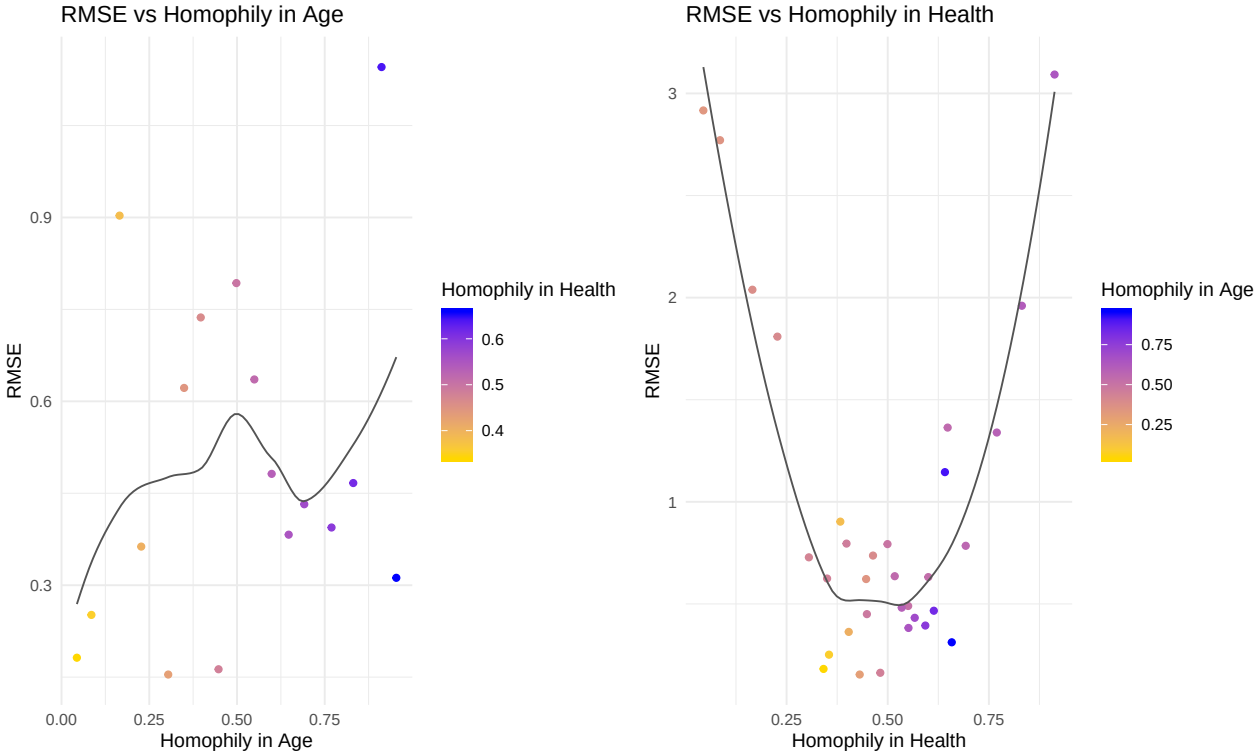

Figure 9: Root Mean Square Error

5.3.3 Coverage

Figure 10 shows that our coverage with populations having homophily in `Age` is always above our 95% coverage line. We also get this unique pattern, where in general, having slight homophily or anti-homophily in `Age` actually makes for a slightly higher coverage rate with the exception of the one case below 96%. The relationship between coverage and homophily in `Health` is an upside down parabolic curve, which is the opposite of the previous homophily in `Health` figures. Every population with a homophily value between .25 and .75 had a coverage above 95%. Then, as the homophily measure moves away from .50 and becomes less random, our coverage starts to drop significantly, falling below 80%. This does not come as a surprise, as these are the populations that had a notably higher bias, which would impact how the confidence intervals are calculated.

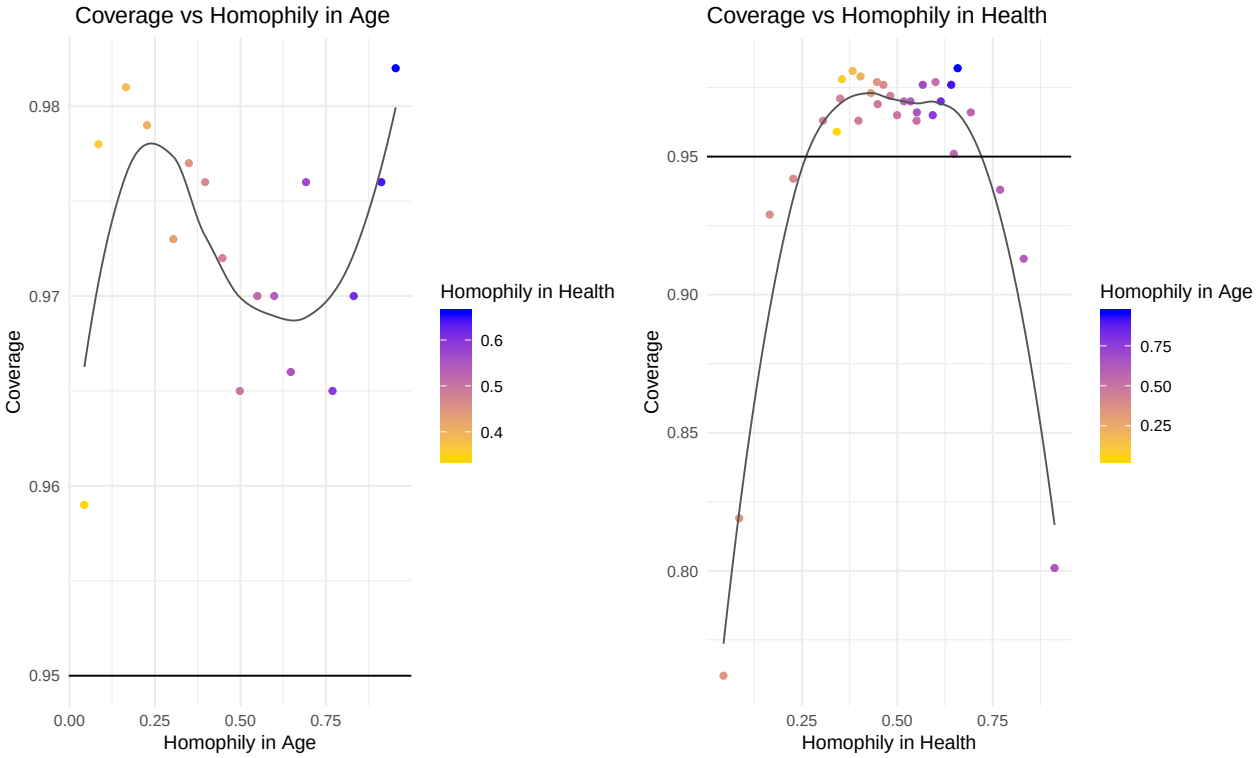

Figure 10: Coverage

6 Discussion

Based on our statistical simulation study, we believe that a basic logistic regression model does *not* perform well in estimating a variable's coefficient with increased levels of anti-homophily and homophily in RDS data. Furthermore, homophily in the `Health` variable seemed to be more influential compared to homophily in the `Age` variable in generated populations, potentially because `Health` was the binary response variable while `Age` was the explanatory variable in this specific model.

The figures for the mixed effects model support the evidence from the basic logistic models that homophily in the response variable, `Health`, has a stronger effect on the post-simulation statistics than the explanatory variable, `Age`. There is also no evidence to suggest that the mixed effects model is better than a basic logistic model for RDS in this scenario, as Spiller's study suggested.

The simulations done for the mixed effects model are still in a preliminary stage due to various factors. One issue was time; we simply lacked enough time to perform as many simulations as we did for the basic logistic model. We would have liked to have as many simulations completed for the mixed effect model as we did for the basic logistic model. Another issue we encountered was that the deviance in the random effect was so small that the simulation would stop due to convergence in our mixed effect simulations. Also, the mixed effects model was intended to try to account for high homophily and anti-homophily populations, but it actually performed worse than the basic model in these situations. It may be worth exploring the mixed effects model further to determine why it does not perform as well in such circumstances.

Homophily is very important to understand when doing Respondent Driven Sampling. It is difficult to measure because there is no one right or wrong way of calculating it, and as a result, we had to create our own method of doing so. However, it is worthwhile to take the time to learn how much homophily is present within a population of interest. Researchers must be aware that having too much homophily or anti-homophily could result in a violation of independence assumptions.

Looking to the future, grasping Respondent Driven Sampling and interpreting homophily can be very useful. Respondent Driven Sampling is a relatively new method of sampling, and there are not many related publications on it. Therefore, by learning more about Respondent Driven Sampling and the potential factor of homophily in a population, we can work on answering questions about hard-to-reach populations.

7 References

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Fellows, I. E. (2018). Respondent-Driven Sampling and the Homophily Configuration Graph. *Wiley Online Library, John Wiley & Sons, Ltd.* doi:10.1002/sim.7973

Ramirez-Valles, J., Heckathorn, D. D., Vazquez, R., Diaz, R. M., & Campbell, R. T. (2005). From networks to populations: The development and application of respondent-driven sampling among IDUs and Latino gay men. *AIDS and Behavior, 9*(4), 387-402.

Respondent Driven Sampling. www.respondentdrivensampling.org/

Spiller, M. W. (2009). Regression modeling of data collected using respondent-driven sampling. *American Sociological Association Annual Meeting*; San Francisco, CA.

Spiller, M. W., Gile, K. J., Handcock, M. S., Mar, C. M., & Wejnert, C. (2017). Evaluating variance estimators for respondent-driven sampling. *Journal of Survey Statistics and Methodology, 2017,* smx018. doi:10.1093/jssam/smx018