PHP2550: Practical Data Analysis

Assignment 4: Simulation

Antonella Basso

November 20, 2022

1. Agent-Based Modeling Paper

First, read the paper *Improving the impact of HIV pre-exposure prophylaxis implementation in small urban centers among men who have sex with men* by Gantenberg et al. available on Canvas (`hiv_prep_abm.pdf`). This paper uses agent-based modeling to determine how to allocate PrEP most effectively. Then, respond to the questions below.

a. Write a 2-3 paragraph summary of the paper.

b. Why do you think agent-based modeling was used to address this question?

c. Give an overview of the model structure (what does an agent represent? how do agents interact?) and the measures/outcomes reported about the simulation study.

d. What were some of the limitations of the model and approach?

Solution

Research Question: "Which statewide PrEP allocation strategies yield the largest reduction in new diagnosed cases of HIV—particularly as it pertains to the country's NHAS target—given a set range of known demographic and behavioral factors within the MSM population aged 15-74?"

a. In a research attempt to improve current interventions for HIV prevention, this paper introduces a simulation-based approach for identifying the pre-exposure prophylaxis (PrEP) prescribing strategies that maximize the population-level impact of statewide implementation. In particular, researchers discuss the development of a stochastic agent-based model (ABM) that, in simulating the spread of HIV under various PrEP allocation conditions within a virtual agent population of size $N = 25,000$, paints a rough picture of what we could expect to observe in the population of men who have sex with men (MSM) between the ages of 15 and 74 in the state of Rhode Island under similar prevention strategies. The agent-based modeling study thus provides insight into the conditions of PrEP allocation that may best decrease the rate of HIV transmission, specifically by approximately 25% over a 10-year period, following the 10-year target reduction of new diagnoses established by the US National HIV/AIDS Strategy (NHAS) in 2010.

Results showed that prioritizing uninfected MSM with more than 10 yearly sexual partners, especially in settings with low PrEP coverage, significantly reduces the number of new cases, compared to other allocation strategies. More importantly, it was shown that in addition to increasing statewide PrEP exposure, engaging MSM with comparatively large numbers of sexual partners could result in the 10-year HIV transmission decreases that satisfy the aforementioned NHAS target. Specifically, investigators found in simulating each prescribing scenario that at least a 25% decrease was achieved in new infections "when PrEP coverage was sustained at 15% of the HIV-negative population over 10 years"—the largest reduction in new cases at the lowest level of PrEP coverage occurring in the case where PrEP engagement was focused on the MSM population with the greatest number yearly sexual partners (> 10). Notably however, the sustained population-level PrEP coverage needed to meet the national goal "is substantially higher than current levels of PrEP uptake". Thus, researchers argue "the importance of engaging MSM at especially high risk of HIV infection" when PrEP allocation strategies are implemented at either the state or city level, to maximize its impact on the population while using up as few resources as possible.

b. The most likely reason investigators decided upon an ABM in response to the primary research question can be attributed to the fact that they are interested in analyzing the effects of various policies on a population (i.e., the effects of statewide PrEP allocation strategies on MSM aged 15-74)—specifically, those which arise form unit-unit interactions rather than from individual-level exposures. That is, given the causal, yet system-oriented, nature of the problem, an ABM is best able to facilitate the analysis of the kind of complex network researchers seek to simulate in a way that properly addresses the interdependence of individuals. Moreover, the fact that the study focuses on a population-level exposure to not just one, but multiple prescribing strategies, a simulation-based approach like this one allows researchers to explore their effects under several specified conditions of interest—something that would simply not be possible in a non-virtual setting, given constraints like time, costs, and the inability to observe the effects of multiple PrEP allocation strategies on the same population for example. Additionally, the fact that ABMs are stochastic, such an approach allows investigators to model various patterns of HIV transmission probabilistically, thereby allowing them to estimate the target population and corresponding policy effects more effectively than other, potentially non-stochastic, methods.

c. With the motivating idea that identifying the PrEP allocation strategies that minimize statewide HIV transmission in a virtual setting would increase its effectiveness in the real world, investigators set out to facilitate such an environment via an ABM wherein $N = 25,000$ artificial agents were stochastically assigned a range of fixed and variable characteristics to simulate Rhode Island's population of MSM aged 15-74 according to statewide patient data from a PrEP clinic, distributions discussed in relevant studies, and other external sources (for approximating sexual tendencies and relationships we could expect to observe in practice). Specifically, each agent (i.e., unit or individual) in the model was assigned the following fixed and updated attributes at each time-step.

Fixed Attributes:

- ***Sexual Role Preference*** (insertive-only, receptive-only, or versatile), which partially constrains the pool of individuals with which an agent is likely to have a sexual encounter (e.g., an insertive-only agent would not be allowed to pair with another exclusively insertive agent in the model).
- ***Mean Number of Annual Sex Partners***, where actual partner numbers vary (annually), but are based on agents' corresponding static distributions to allow for year-year behavioral variation while maintaining sexual tendencies regarding partner acquisition.
- ***Mean Annual Sex Frequency per Partner***, where the number of encounters an agent has with each partner is derived from taking the (pairwise) mean of their desired annual per-partner sex frequencies.

Time-Updated Attributes:

- ***Age***, for which agents have a 50% chance of pairing with units within their own age group and a likelihood of pairing with individuals outside their age group that decreases as a function of age difference.
- ***PrEP Status***, where agents initiate or discontinue treatment probabilistically.
- ***HIV Status***, where uninfected agents are tested at a starting rate (based on Rhode Island data), which is then tuned to yield diagnoses in roughly 82% of the simulated population; infected agents aware of their HIV statuses are subject to decreased risks of transmission (independent of treatment status) "based on the general observation that HIV-infected MSM reduce certain risk behaviors post-diagnosis"; and the probability of transmission from a diagnosed HIV-infected agent to a uninfected unit in a serodiscordant partnership is scaled by a fixed value of 0.5.

- ***Status of Viral Suppression***, where suppression may be achieved via antiretroviral treatment (ART) administered stochastically to diagnosed individuals, "with a target rate of viral suppression among HIV-infected agents of 45%".

Given that the study incorporates a "death" component into the model, agents who exit the simulation are also replaced by new uninfected units from the same age group with stochastically assigned attributes of interest, for the purpose maintaining the same population size. Moreover, as is the primary goal of the study, investigators constructed a discrete-time stochastic ABM, calibrated to Rhode Island's statewide HIV prevalence from 2009–2014, to simulate patterns of HIV trasmission within the virtual agent population (representing MSM aged 15–74) under the following five PrEP allocation scenarios for comparative analysis, subject to the conditions discussed above.

PrEP Allocation Strategies:

1. no allocation (for baseline comparison);
2. random allocation;
3. allocation to "the current patient population";
4. allocation to the population of MSM with over 5 annual sexual partners; and
5. allocation to the population of MSM with over 10 annual sexual partners.

Lastly, investigators estimated the number and proportion of infections prevented, as well as the corresponding person-years, for each of these scenarios under various levels of PrEP coverage. In the same order as above, a few of these results are given below, assuming 15% PrEP coverage over a 10-year period, where incidence rate pis measued as incidences per 1,000 at-risk person-years.

PrEP Allocation Strategy Results:
(Median New Infections (#), Median Infections Averted (#), Incidence Rate)

1. 826, 0, 3.51
2. 654, 176, 2.77
3. 612, 218, 2.59
4. 595, 235, 2.52
5. 555, 275, 2.35

As discussed in part (a), centering PrEP engagement around the at-risk population of MSM with more than 10 sexual partners per year, is proves to be the allocation strategy that yields the best results with regards to HIV prevention.

d. Researchers discuss several limitations specific to their chosen model. We categorize and detail them as follows.

Assumption Limitations:

- Given that the data used to model the MSM population of interest is itself (inevitably) limited, the model falls short in that it does not adequately account for potential sources of individual heterogeneity that drive observed processes between subjects. Although probabilities are placed over varying characteristics for this purpose, they still do not explicitly address, for example, differences in age-specific sex frequency nor in condom use patterns by partner type, sexual role preference, and/or HIV serostatus. Moreover, not accounting for behaviors and interactions such as non-dyadic encounters and brief partnerships could have also influenced inconsistencies between observations, results, and the model's implicit assumptions.

- It is possible for the model to have slightly underestimated PrEP efficiency, given that PrEP assignment was inherently ignorant to different risk behaviors among prescribed agents such as condom use. Similarly, (even if only minimally) the fact that an agent's actual number of sex partners varies annually suggests that agent pools for specific PrEP allocation scenarios are also subject to change on a yearly basis. That is, for example, suppose that an agent with a mean number of annual sex partners below 10—who therefore has a risk of infection less than does one with a much higher mean number of annual sex partners—that, in a given year, happens to display an actual number of sex partners greater than 10 and is chosen for PrEP by virtue of having been placed in the pool of higher-risk individuals eligeable for prescription. It follows that the estimated positive impact of PrEP on the population according to the model, in this case, will be smaller than the ("true") impact we (likely) would have observed had an agent at higher-risk of infection been chosen instead. For this reason, researchers argue that decreased levels of PrEP coverage (i.e., 5% and 10%) are likely to be the most accurate in estimating the true effects of PrEP allocation that is concentrated around MSM with at more than 10 annual sex partners.

- A meaningful phenomenon discussed in a relevant study, which investigators did not consider when constructing the model, involves behavioral risk patterns attributed to individuals recently perscribed PrEP, such as decreased condom use and/or increased partner acquisition rates, for example. Such behaviors, which put agents initiating PrEP at greater risks for infection, could (indirectly) influence the model's treatment effect estimates, especially as it does not implement HIV screening intervals for agents on PrEP.
- It is also possible for the model to have overestimated HIV incidence, given the assumption that all sexual partnerships and encounters occur within the bounds of the state being modeled (Rhode Island, in our case), which is inconsistent with the fact that MSM (in Rhode Island) typically engage in sexual partnerships with MSM from other states. Similarly, the model does not account for cases in which agents are prescribed PrEP from out-of-state providers, demonstrating another potential disagreement between the model's observations and its rather strict assumptions.

Calibration Limitations:

- While diagnoses (in Rhode Island) seldom fluctuate, the fact that model predicts rising HIV incidence can be partially attributed to its having been calibrated to rising prevalence meanwhile imposing steady mortality rates among those infected.
- Many of the limitations discussed regarding the model's assumptions have additional implications for its calibration. Namely, despite the fact that sensitivity analyses for sex frequency and annual sex partner count were conducted to account for differences in PrEP policy effect estimates that may arise from (resulting) varying patterns of HIV-transmission, the downfall, as researchers claim, lies in that "these analyses also featured markedly different dynamics than those underlying the main model".

In addition to these limitations specific to the study model, there are some drawbacks to the ABM approach more broadly that are important to consider as it related to the study's findings and implications. We summarize these as follows.

- As mentioned by investigators in the paper, one of the limitations of ABMs, and other parametric modeling approaches, is "the considerable uncertainty surrounding model parameters". Specifically, given the possibility that the populations used to estimate condom use and sex frequency parameters do not actually resemble the Rhode Island MSM population, it isn't certain whether the model's outcomes and/or inferences are free of bias—namely, that which arises from using external populations to estimate model parameters. For this reason, investigators argue the importance of collecting as much and as detailed information as possible across relevant settings when parameterizing ABMs.
- Aside from uncertainty regarding parameter estimates and potential bias due to data scarcity and/or misrepresentation, approaches like ABMs can be difficult to validate in the presence of unmeasured confounders (i.e., when modeling unobserved associations). In turn, this could both bias our results and lead to invalid inferences. Thus, as mentioned previously with regards to the model's assumption limitations, it is crucial to account for as many potential unobserved sources of variation and bias as possible by either relaxing assumptions or being explicit about removing corresponding processes in the model.

2. Simulating Bias and Uncertainty from Model Selection

The point of this question is to illustrate that reported regression coefficients and associated standard errors can be incorrect when ignoring the iterative nature of actual model building. This question illustrates this concept in a simple case. Consider a simple linear regression model.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{where } \epsilon_i \sim \mathrm{N}(0,1)$$

Suppose you have m observations where $x = -1/\sqrt{m}$ for half of the observed observations and $x = 1/\sqrt{m}$ for the other half. Your interest centers on estimating β_1.

Consider the following estimation procedure. You fit the model and test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ at level of significance α. If the test fails to reject H_0, set the estimator to 0. If the test rejects, set the estimator to the OLS estimator $\hat{\beta}$. Call this estimator $\hat{\beta}_1^\alpha$.

Evaluate the performance of $\hat{\beta}_1^\alpha$ for estimating β_1 by setting up a simulation study. Justify the design of your simulation study using the ADEMP framework seen in class. Further, interpret the results.

Solution

Our proposed simulation study, with an **Aim** to illustrate the importance of iterative processes when building regression models for estimating parameters, involves implementing the $\hat{\beta}_1^\alpha$ algorithm to iteratively estimate the **Estimand** of interest β_1, allowing for varying values of K, m and α, where

- β_1 is the slope coefficient for X in the linear model above;
- K is the number of iterations in the simulation;
- m is the number of observations in the data; and
- α is the significance level at which we test the null hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ when fitting the model at each iteration k to decide whether to set $\hat{\beta}_{1k}^\alpha$ equal to 0 or to the OLS estimator $\hat{\beta}_{1k}$.

The **Data-Generating Process**, **Methods**, and **Performance Measures** aspects of the ADEMP framework are summarized by the following algorithm.

Algorithm: Coefficient Estimator ($\hat{\beta}_1^\alpha$)

```
hat_b1_est_sim(seed, iter, m, alpha)
```

- **Inputs:** Where,
  - `seed` is the seed to be set for each simulation for reproducibility purposes;
  - `iter` is the desired number of iterations K;
  - `m` is the number of observations m in the data to be generated; and
  - `alpha` is the chosen significance level α at which to test the standard null hypothesis for estimating β_1 at each iteration k.
- **Outputs:** Where,
  - `b0_vec` is a length K vector of estimated intercepts $\beta_{0k} = \hat{\beta}_{0k}$ irrespective of α;
  - `b1_vec` is a length K vector of estimated coefficients $\beta_{1k} = \hat{\beta}_{1k}$ irrespective of α;
  - `sim_b1_est` is a length K vector of estimated coefficients $\beta_{1k}^\alpha = \{0, \hat{\beta}_{1k}\}$ subject to specified α in the estimation procedure detailed below;
  - `hat_b1` is the expectation of $\hat{\beta}_1^\alpha$ (i.e., the mean of β_{1k}^α);
  - `mse` is the mean squared error (MSE) of $\hat{\beta}_1^\alpha$, given by MSE$= \frac{1}{K}\sum_{k=1}^{K}(\beta_{1k}^\alpha - \beta_1)^2$; and
  - `bias` is the bias of $\hat{\beta}_1^\alpha$, given by $Bias = \mathrm{E}[\hat{\beta}_1^\alpha] - \beta_1$.

- **Estimation Procedure:**

  For each iteration $k = 1, 2, ..., K$ (i.e., `k in 1:iter`) do the following.

  1. Define the X vector of length m such that $x_i = \begin{cases} -\frac{1}{\sqrt{m}} & \text{for} \quad i = 1, 2, ..., \frac{m}{2} \\ \frac{1}{\sqrt{m}} & \text{for} \quad i = \frac{m}{2} + 1, , ..., m \end{cases}$
     (i.e., `x = c(rep(-1/sqrt(m), m/2), rep(1/sqrt(m), m/2))`).
  2. Sample m values from the standard normal distribution to obtain the m-length vector for ϵ such that $\epsilon_i \sim \mathrm{N}(0, 1)$ (i.e., let `eps = rnorm(m)`).
  3. Set "true" values for β_0 and β_1 to generate Y—parameters to be subsequently estimated. For this study, we let $\beta_0 = 4$ and $\beta_1 = 0.7$ (i.e., `t_b0 = 4` and `t_b1 = 0.7`), hence the reason we do not include them as user-defined parameters.
  4. Compute the length m vector for Y following the regression equation. That is, for each x_i and ϵ_i let $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, such that $Y = 4 + 0.7X + \vec{\epsilon}$ (i.e., `y = t_b0 + t_b1*x + eps`).
  5. Create a $2 \times m$ data frame for use in the linear model with vectors X and Y (i.e., `data <- data.frame(x, y)`).
  6. Use the data to estimate the β coefficients using the `lm` and `summary.lm` functions (i.e., `model <- lm(y ~ x, data)` and `model_summary <- summary(model)`), subsequently appending them to their corresponding output vectors, ignoring α.
  7. Obtain the β_{1k}^{α} vector via the $\hat{\beta}_1^{\alpha}$ estimator as follows. If p-value `Pr(>|t|)` $\leq \alpha$, then set $\beta_{1k}^{\alpha} = \beta_{1k}$, otherwise set $\beta_{1k}^{\alpha} = 0$ (i.e., using `p = model_summary$coef[2,4]`, `ifelse(p <= alpha, b1, 0)`). Subsequently, append β_{1k}^{α} to the corresponding output vector `sim_b1_est`.

  Compute the $\hat{\beta}_1^{\alpha}$ estimate `hat_b1` by taking the mean of the β_{1k}^{α} vector `sim_b1_est` (i.e., `mean(sim_b1_est)`).

  Obtain the performance measures for $\hat{\beta}_1^{\alpha}$.

  - `mse` using MSE$= \frac{1}{K}\sum_{k=1}^{K}(\beta_{1k}^{\alpha} - \beta_1)^2$ (i.e., `sum((sim_b1_est - t_b1)^2)/iter`); and
  - `bias` using $Bias = \mathrm{E}[\hat{\beta}_1^{\alpha}] - \beta_1$ (i.e., `hat_b1 - t_b1`).

It should be noted that since varying m (i.e., letting m be 10, 100, and 1,000) had no noticeably significant effect on the simulation estimates, in addition to varying α, we analyze our estimates under varying values of K. Hence, our simulations are initialized as follows.

Each simulation,

- uses the same random seed value `seed`$= 4$ for reproducibility;
- assumes "true" coefficients $\beta_0 = 4$ and $\beta_1 = 0.7$ for estimation;
- generates data sets of $m = 100$ observations;
- tests regression coefficient null hypotheses at significance levels $\alpha = \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$, where $\alpha = 1$ implies ignoring the constraint altogether (i.e., accepting the OLS estimate β_{1k} or rejecting the null hypothesis every time);
- considers either 100, 500, or $1,000$ iterations K; and
- evaluates performance via MSE and bias.

Specifically, we consider MSE and bias to evaluate the performance of our simulation-based estimator $\hat{\beta}_{1k}^{\alpha}$ as each provides favorable insight into how well model estimates fit the data on average. Specifically, while bias tells us how far off the average estimated value generally is from the true value $\beta_1 = 0.7$, MSE sheds light on the amount of error produced by the model used to fit the data overall. Although it is typical in practice to accept models with low MSE to maximize prediction accuracy, this isn't always guaranteed. As both bias and variance contribute to MSE, lower values are often achieved by minimizing both bias and variance as much as possible. However, as we see in the simulation results below, testing with smaller α, which leads to preferable MSE values, does not give optimal estimates.

Table 1: Simulation Results

	K=100			K=500			K=1,000		
α	$\hat{\beta}_1^\alpha$	MSE	Bias	$\hat{\beta}_1^\alpha$	MSE	Bias	$\hat{\beta}_1^\alpha$	MSE	Bias
0.001	0.0000	0.4900	-0.7000	0.0205	0.5317	-0.6795	0.0180	0.5313	-0.6820
0.005	0.0000	0.4900	-0.7000	0.0436	0.5660	-0.6564	0.0482	0.5804	-0.6518
0.010	0.0249	0.5171	-0.6751	0.0591	0.5848	-0.6409	0.0696	0.6080	-0.6304
0.050	0.1478	0.6005	-0.5522	0.2127	0.7112	-0.4873	0.2128	0.7671	-0.4872
0.100	0.2734	0.6528	-0.4266	0.3222	0.7836	-0.3778	0.3269	0.8278	-0.3731
0.500	0.6496	0.7116	-0.0504	0.6444	0.8946	-0.0556	0.6351	0.9558	-0.0649
1.000	0.6809	0.7253	-0.0191	0.6844	0.8963	-0.0156	0.6763	0.9607	-0.0237

As both bias and variance contribute to MSE, lower values are often achieved by minimizing both bias and variance as much as possible. However, these results show that testing with smaller values of α, despite decreasing error, can also bring our model to underfit the data more significantly. This is likely due to the fact that being less restrictive with the coefficients we accept into the model more drastically increases the estimator's variance such that, on average, we get closer to the target estimand, but not without seeing an increase in MSE as well. Noticeably, as bias is decreased by the increase in variance, our estimates for β_1 get closer to the true value 0.7.

Moreover, looking at estimates with regards to our chosen number of iterations, we see that, as was the case with α, increasing values of K coincide with more accurate coefficient estimates as well as with larger MSEs. Despite such increases not being as drastic, this nonetheless reinforces the importance of implementing iterative model building processes when estimating parameters.

Although in practice, we are often willing to make estimators biased for the purpose of minimizing error, this study demonstrates how too much bias can lead us further away from the true parameters we wish to estimate as variance decreases. For this reason, it is important to consider the trade-off between bias and variance and not immediately resort to standard hypothesis testing for estimating regression coefficients, as it may produce incorrect results given the underlying structure of the data. Lastly, this study shows us that the iterative nature of model building calls for simulation-based methods like this to identify phenomena that do not necessarily align with our traditional statistical intuitions.

Code Appendix

```
## Libraries
library(tidyverse)
library(lme4)
library(lmtest)
library(latex2exp)
library(kableExtra)
```

```
### Simulation Function
hat_b1_est_sim <- function(seed, iter, m, alpha){

  set.seed(seed) # seed

  # output vectors
  b0_vec <- c() # b0 estimates (ignoring alpha)
  b1_vec <- c() # b1 estimates (ignoring alpha)
  sim_b1_est <- c() # (hat) b1 estimates (at sig. level = alpha)

  # coefficient estimation procedure
  for (i in 1:iter){

    # given
    x <- c(rep(-1/sqrt(m), m/2), rep(1/sqrt(m), m/2)) # x variable
    eps <- rnorm(m) # normally distributed errors

    # setting values for "true" b0 and b1 to generate y given m, x, and eps

    # true parameters
    t_b0 <- 4 # set b0 = 4
    t_b1 <- 0.7 # set b1 = 0.7

    # generating data (y)
    y <- t_b0 + t_b1*x + eps
    data <- data.frame(y=y, x=x)

    # regression model
    model <- lm(y ~ x, data=data) # lm for estimating b1
    model_summary <- summary(model) # summary

    # all estimated model coefficients
    b0 <- model_summary$coef[1,1]
    b1 <- model_summary$coef[2,1]

    b0_vec <- c(b0_vec, b0)
    b1_vec <- c(b1_vec, b1)

    p <- model_summary$coef[2,4] # p-value

    # b1 estimates (at sig. level = alpha)
    b1_est <- ifelse(p <= alpha, b1, 0)
    sim_b1_est <- c(sim_b1_est, b1_est)
  }
  # hat b1
  hat_b1 <- mean(sim_b1_est)

  # performance measures (MSE and bias)
  mse <- sum((sim_b1_est - t_b1)^2)/iter
  bias <- hat_b1 - t_b1

  return(list(b0=b0_vec,
              b1=b1_vec,
              sim_b1_est=sim_b1_est,
              hat_b1=hat_b1,
              mse=mse,
              bias=bias))
}
```

```
### Simulations (varying alpha & m)

# potential values for alpha
alpha_values <- c(0.001, 0.005, 0.01, 0.05, 0.1, 0.5)

## Group 1 Simulations
# 100 iterations for 100 observations using each alpha value (above)
# using the same seed for reproducibility
sim1_a001 <- hat_b1_est_sim(seed=4, iter=100, m=100, alpha=alpha_values[1])
sim1_a005 <- hat_b1_est_sim(seed=4, iter=100, m=100, alpha=alpha_values[2])
sim1_a01 <- hat_b1_est_sim(seed=4, iter=100, m=100, alpha=alpha_values[3])
sim1_a05 <- hat_b1_est_sim(seed=4, iter=100, m=100, alpha=alpha_values[4])
sim1_a1 <- hat_b1_est_sim(seed=4, iter=100, m=100, alpha=alpha_values[5])
sim1_a5 <- hat_b1_est_sim(seed=4, iter=100, m=100, alpha=alpha_values[6])

## Group 2 Simulations
# 500 iterations for 100 observations using each alpha value (above)
# using the same seed for reproducibility
sim2_a001 <- hat_b1_est_sim(seed=4, iter=500, m=100, alpha=alpha_values[1])
sim2_a005 <- hat_b1_est_sim(seed=4, iter=500, m=100, alpha=alpha_values[2])
sim2_a01 <- hat_b1_est_sim(seed=4, iter=500, m=100, alpha=alpha_values[3])
sim2_a05 <- hat_b1_est_sim(seed=4, iter=500, m=100, alpha=alpha_values[4])
sim2_a1 <- hat_b1_est_sim(seed=4, iter=500, m=100, alpha=alpha_values[5])
sim2_a5 <- hat_b1_est_sim(seed=4, iter=500, m=100, alpha=alpha_values[6])

## Group 3 Simulations
# 1000 iterations for 100 observations using each alpha value (above)
# using the same seed for reproducibility
sim3_a001 <- hat_b1_est_sim(seed=4, iter=1000, m=100, alpha=alpha_values[1])
sim3_a005 <- hat_b1_est_sim(seed=4, iter=1000, m=100, alpha=alpha_values[2])
sim3_a01 <- hat_b1_est_sim(seed=4, iter=1000, m=100, alpha=alpha_values[3])
sim3_a05 <- hat_b1_est_sim(seed=4, iter=1000, m=100, alpha=alpha_values[4])
sim3_a1 <- hat_b1_est_sim(seed=4, iter=1000, m=100, alpha=alpha_values[5])
sim3_a5 <- hat_b1_est_sim(seed=4, iter=1000, m=100, alpha=alpha_values[6])
```

```
### Simulation Results

## (1) K=100
# hat b1 estimate for each alpha
```

```
hat_b1_est1 <- c(sim1_a001$hat_b1, sim1_a005$hat_b1,
                 sim1_a01$hat_b1, sim1_a05$hat_b1,
                 sim1_a1$hat_b1, sim1_a5$hat_b1)

# hat b1 MSE for each alpha
hat_b1_mse1 <- c(sim1_a001$mse, sim1_a005$mse,
                 sim1_a01$mse, sim1_a05$mse,
                 sim1_a1$mse, sim1_a5$mse)

# hat b1 bias for each alpha
hat_b1_bias1 <- c(sim1_a001$bias, sim1_a005$bias,
                  sim1_a01$bias, sim1_a05$bias,
                  sim1_a1$bias, sim1_a5$bias)

# b0 and b1 estimates ignoring alpha
c(mean(sim1_a001$b0), mean(sim1_a001$b1)) # closest to true values

## (2) K=500
# hat b1 estimate for each alpha
hat_b1_est2 <- c(sim2_a001$hat_b1, sim2_a005$hat_b1,
                 sim2_a01$hat_b1, sim2_a05$hat_b1,
                 sim2_a1$hat_b1, sim2_a5$hat_b1)

# hat b1 MSE for each alpha
hat_b1_mse2 <- c(sim2_a001$mse, sim2_a005$mse,
                 sim2_a01$mse, sim2_a05$mse,
                 sim2_a1$mse, sim2_a5$mse)

# hat b1 bias for each alpha
hat_b1_bias2 <- c(sim2_a001$bias, sim2_a005$bias,
                  sim2_a01$bias, sim2_a05$bias,
                  sim2_a1$bias, sim2_a5$bias)

# b0 and b1 estimates ignoring alpha
c(mean(sim2_a001$b0), mean(sim2_a001$b1)) # closest to true values

## (3) K=1,000
# hat b1 estimate for each alpha
hat_b1_est3 <- c(sim3_a001$hat_b1, sim3_a005$hat_b1,
                 sim3_a01$hat_b1, sim3_a05$hat_b1,
                 sim3_a1$hat_b1, sim3_a5$hat_b1)

# hat b1 MSE for each alpha
hat_b1_mse3 <- c(sim3_a001$mse, sim3_a005$mse,
                 sim3_a01$mse, sim3_a05$mse,
                 sim3_a1$mse, sim3_a5$mse)

# hat b1 bias for each alpha
hat_b1_bias3 <- c(sim3_a001$bias, sim3_a005$bias,
                  sim3_a01$bias, sim3_a05$bias,
                  sim3_a1$bias, sim3_a5$bias)

# b0 and b1 estimates ignoring alpha
```

```
c(mean(sim3_a001$b0), mean(sim3_a001$b1)) # closest to true values

## Results Data Frame
b1_1 <- c(hat_b1_est1, hat_b1_est_sim(seed=4, iter=100, m=100, alpha=1)$hat_b1)
b1_mse1 <- c(hat_b1_mse1, hat_b1_est_sim(seed=4, iter=100, m=100, alpha=1)$mse)
b1_bias1 <- c(hat_b1_bias1, hat_b1_est_sim(seed=4, iter=100, m=100, alpha=1)$bias)

b1_2 <- c(hat_b1_est2, hat_b1_est_sim(seed=4, iter=500, m=100, alpha=1)$hat_b1)
b1_mse2 <- c(hat_b1_mse2, hat_b1_est_sim(seed=4, iter=500, m=100, alpha=1)$mse)
b1_bias2 <- c(hat_b1_bias2, hat_b1_est_sim(seed=4, iter=500, m=100, alpha=1)$bias)

b1_3 <- c(hat_b1_est3, hat_b1_est_sim(seed=4, iter=1000, m=100, alpha=1)$hat_b1)
b1_mse3 <- c(hat_b1_mse3, hat_b1_est_sim(seed=4, iter=1000, m=100, alpha=1)$mse)
b1_bias3 <- c(hat_b1_bias3, hat_b1_est_sim(seed=4, iter=1000, m=100, alpha=1)$bias)

b1_df <- data.frame(alpha=c(alpha_values, 1),
                    b1_1, b1_mse1,b1_bias1,
                    b1_2, b1_mse2, b1_bias2,
                    b1_3, b1_mse3, b1_bias3)

b1_df <- as.data.frame(apply(b1_df, 2, round, 4))

# NOTICE: estimate gets closer to true value with increasing alpha values
```