PLSC 30600: Problem Set 1

YOUR NAME

January 5, 2022

This problem set is due at **11:59 pm on Tuesday, January 17th**.

Please upload your solutions as a .pdf file saved as "Yourlastname_Yourfirstinitial_pset1.pdf"). In addition, an electronic copy of your .Rmd file (saved as "Yourlastname_Yourfirstinitial_pset1.Rmd") must be submitted to the course website at the same time. We should be able to run your code without error messages. In addition to your solutions, please submit an annotated version of this `.rmd` file saved as "Yourlastname_Yourfirstinitial_pset1_feedback.rmd", noting the problems where you needed to consult the solutions and why along with any remaining questions or concerns about the material. In order to receive credit, homework submissions must be substantially started and all work must be shown. Late assignments will not be accepted.

Problem 1

In this problem we will examine what information the data might provide us regarding the magnitude or direction of a treatment effect if we are only willing to make a consistency or SUTVA assumption with respect to the potential outcomes and a positivity/overlap assumption on the probability of treatment.

Consider our standard causal inference setup with a binary treatment and a binary outcome. Y_i denotes the observed outcome for unit i, $Y_i \in \{0, 1\}$. D_i denotes the observed treatment for unit i. $Y_i(d)$ denotes the potential outcome we would observe if i were assigned treatment value d. By consistency, we have: $Y_i(d) = Y_i$ if $D_i = d$. By positivity, we have $0 < Pr(D_i = 1) < 1$. In other words, treatment is not deterministic and each unit could have received either treatment or control.

Our estimand is the average treatment effect $\tau = E[Y_i(1) - Y_i(0)]$. We will focus here on causal identification and work with the population expectations $E[Y_i(d)]$ and $E[Y_i]$ and conditional expectations $E[Y_i(d)|D_i]$ and $E[Y_i|D_i]$, setting aside the question of estimation.

Part A

Write an expression for the average treatment effect in terms of the difference in observed means between treatment and control $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$ and a bias term.

Part B

Which components of the bias term can be observed from the data and which ones cannot? Can the average treatment effect be point-identified from the data alone? Explain why.

Part C

What is the smallest possible value of the bias term? What is the largest possible value of the bias term? (remember that Y_i is a binary value that can either be 0 or 1).

Part D

Using your answer from part C, write the upper and lower bounds of the average treatment effect in terms of the observable conditional expectations $E[Y_i|D_i = 1]$, $E[Y_i|D_i = 0]$. Is there any value of the observed difference-in-means for which the bounds only contain positive or only contain negative values? What does this tell us about whether we can learn about the direction of a treatment effect from the observed data alone?

Problem 2

Despite its significant importance to many political debates, there are few causal estimates of the effect of expanded healthcare insurance on healthcare outcomes. One landmark study, the Oregon Health Insurance Experiment, covered new ground by utilizing a randomized control trial implemented by the state government of Oregon. To allocate a limited number of eligible coverage slots for the state's Medicaid expansion, about 30,000 low-income, uninsured adults (out of about 90,000 wait-list applicants) were randomly selected by lottery to be allowed to apply for Medicaid coverage. Researchers collected observable measures of health (blood pressure, cholesterol, and blood sugar levels), as well as hospital visitation and healthcare expenses for 6,387 selected adults and 5,842 not selected adults.

For this problem, you will need the `OHIE.dta` file. The variables you will need are:

```
ohie <- haven::read_dta("OHIE.dta")
```

The variables you will need are:

- `treatment` - Selected in the lottery
- `ohp_all_ever_admin` - Ever enrolled in Medicaid from matched notification date to September 30, 2009 (actually had Medicaid insurance)
- `weight_total_inp` - Survey weight
- `tab2bp_hyper` - Outcome: Binary indicator for elevated blood pressure (defined a systolic pressure of 140mm Hg or more and a diastolic pressure of 90mm Hg or more)
- `tab2phqtot_high` - Outcome: Binary indicator for a positive screening result for depression (defined as a score of 10 or higher on the Patient Health Questionaire - 8)
- `tab4_catastrophic_exp_inp` - Outcome: Indicator for catastrophic medical expenditure (total out-of-pocket medical expenses $\geq$ 30% of household income)
- `tab1_gender_inp` - gender (0 - Male, 1 - Female, 2 - Transgender)
- `tab1_age_19_34_inp` - Age 19-34
- `tab1_age_35_49_inp`- Age 35-49
- `tab1_race_black_inp` - Race/ethnicity is Black
- `tab1_race_nwother_inp` - Race/ethnicity is non-White/other
- `tab1_race_white_inp` - Race/ethnicity is White
- `tab1_hispanic_inp` - Hispanic/Latino

We'll start by subsetting the data down to only those observations where all three of the outcomes and the seven covariates are non-missing

```
# Which observations have all non-missing outcomes + covariates
ohie$nonmissing <- complete.cases(ohie %>% dplyr::select(tab2bp_hyper, tab2phqtot_high,
                                                         tab4_catastrophic_exp_inp, tab1_gender_inp, tab
                                                         tab1_age_35_49_inp, tab1_race_black_inp, tab1_
                                                         tab1_race_white_inp, tab1_hispanic_inp))
```

```
# Subset down to complete cases
ohie_complete <- ohie %>% filter(nonmissing == 1)
```

Note that because of the methods used to recruit treated and control individuals for observation, all analyses are weighted using the known sample selection weight `weight_total_inp` (see the paper for more details on how this was constructed). Your analyses below should incorporate these weights.

Part A

Using the complete case data and the pre-treatment covariates, assess whether you think randomization of the selection lottery was successfully carried out. Explain why or why not.

Part B

Estimate the average intent-to-treat effect on each of the three separate outcomes: elevated blood pressure, depression, and catastrophic medical expenditure. Provide a 95% asymptotic confidence interval for each and assess, for each outcome, whether you would reject the null of no ITT at $\alpha = .05$. Briefly discuss your findings and provide a substantive interpretation of your results.

Part C

Estimate the average effect of being selected in the lottery on actual enrollment in Medicaid. Provide a 95% confidence interval and determine whether you would reject the null of no average effect of selection on enrollment at $\alpha = .05$. Based on your results, discuss whether you think selection in the lottery had a meaningful effect on treatment uptake.

Part D

Suppose that a researcher instead chose to estimate the effect of Medicaid enrollment using a "per-protocol" analysis - comparing participants assigned to treatment (selected in the lottery) who did enroll in Medicaid to those assigned to control (not selected) who did not enroll. Use this "per-protocol" analysis to estimate the average treatment effect of Medicaid enrollment on depression, provide a 95% asymptotic confidence interval and compare your results to the ITT estimate from Part B.

Does the "per-protocol" analysis provide an unbiased estimator of the average treatment effect of Medicaid? Explain why or why not.

Problem 3

In this problem, you will use simulation to learn about the sampling variance of the difference-in-means estimator for the ATE under different randomization schemes.

Assume the following data-generating process:

We observe sample of $N = 100$ observations. Each unit is assigned treatment $D_i = 1$ with some probability $Pr(D_i = 1)$. We will assume that the outcome is generated by $Y_i = \tau D_i + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, 1)$. We will assume a constant, additive treatment effect of $\tau = 2$ for the sake of the simulation.

Part A

Suppose treatment was assigned via independent Bernoulli trials with a constant probability of treatment $Pr(D_i = 1) = .5$ and $D_i \perp\!\!\!\perp D_j$ for all units $i \neq j$. Using a monte carlo simulation and assuming the data-generating process above, find the variance of the sampling distribution of the simple difference-in-means estimator (use `60637` as your random seed set at the beginning of the code fragment and use 10000 monte carlo iterations).

Part B

Now consider a completely randomized experiment where $N_t = 50$ units receive treatment and $N_c = 50$ units receive control. In this setting, the marginal probability of treatment is $\mathbb{P}(D_i = 1) = .5$ but D_i is not independent of D_j. Using a monte carlo simulation for this assignment process, find the variance of the sampling distribution of the simple difference-in-means estimator (again, use `60637` as your random seed set at the beginning of the code fragment and use 10000 monte carlo iterations). Compare your variance to the variance under the data-generating process from Part A and discuss why they may differ.

Part C

Sometimes when designing an experiment, it is impossible to completely randomize over the entire sample of respondents since respondents arrive in a sequence. For example, experimenters fielding online surveys do not observe the entire sample and sometimes have to randomly assign treatments in a "just-in-time" manner.

Efron (1971) suggests an alternative approach to independent bernoulli randomization that biases the coin depending on how many units have previously been assigned to the treatment group versus the control group.

Consider the randomization scheme where treatment is assigned sequentially for units 1 through 100 according to their order. In other words, treatment for unit 1 is randomly assigned. Then treatment for unit 2 is randomly assigned depending on the value of the treatment for unit 1, and so on... Let $\tilde{N_{t,i}}$ denote the number units treated prior to unit i, $\tilde{N_{c,i}}$ the number of units under control prior to unit i and $\tilde{Z}_i = \tilde{N_{t,i}} - \tilde{N_{c,i}}$ or the difference in the number of treated and control groups. By definition, $\tilde{Z}_1 = 0$ since there are no treated or control units when the first unit is assigned.

Define the probability of treatment $Pr(D_i = 1)$ for the ith unit as

$$Pr(D_i = 1) = \begin{cases} \pi & \text{if } \tilde{Z}_i < 0 \\ 0.5 & \text{if } \tilde{Z}_i = 0 \\ (1 - \pi) & \text{if } \tilde{Z}_i > 0 \end{cases}$$

Intuitively, the assignment mechanism biases the probability of receiving treatment upward if there are fewer treated than control and biases it downward if there are more treated than control at the time of assignment.

Let $\pi = .9$. Using a monte carlo simulation for this assignment scheme, find the variance of the sampling distribution of the simple difference-in-means estimator (use `60637` as your random seed set at the beginning of the code fragment and use 10000 monte carlo iterations). Compare your variance to your result in Part A and your result in Part B. Discuss any differences you observe.

Part D

Using your simulation results from Part C, is the difference-in-means estimator using this assignment scheme unbiased for the average treatment effect $\tau = 2$?

Part E

Intuitively, what will happen to the sampling variance if π is set to be less than .5? (You don't need to use a simulation to answer this, but you are welcome to use one if it would help).

Problem 4

Do international election monitors reduce the incidence of electoral fraud? Hyde (2007) studies the 2003 presidential election in Armenia, an election that took place during a period where the incumbent ruling party headed by President Robert Kocharian had consolidated power and often behaved in ways that were considered undemocratic.

The full citation for this paper is

> Hyde, Susan D. "The observer effect in international politics: Evidence from a natural experiment." *World Politics* 60.1 (2007): 37-63. At the time of the election, OSCE/ODIHR election monitors reported widespread electoral irregularities that favored the incumbent party such as ballot-box stuffing (pp. 47). However, we do not necessarily know whether these irregularities would have been worse in the absence of monitors. Notably, not all polling stations were monitored – the OSCE/ODIHR mission could only send observers to some of the polling stations in the country. Since in the context of this election only the incumbent party would have the capacity to carry out significant election fraud, Hyde examines whether the presence of election observers from the OSCE/ODIHR mission at polling stations in Armenia reduced the incumbent party's vote share at that polling station.

For the purposes of this problem, you will be using the `armenia2003.dta` dataset

The R code below will read in this data (which is stored in the STATA .dta format)

```
### Hyde (2007) Armenia dataset
armenia <- read_dta("armenia2003.dta")
```

This dataset consists of 1764 observations polling-station-level election results from the 2003 Armenian election made available by the Armenian Central Election Commission. The election took place over two rounds with an initial round having a large number of candidates and a second, run-off election, between Kocharian and the second-place vote-getter, Karen Demirchyan. We will focus on monitoring and voting in the first round. The specific columns you will need are:

- `kocharian` - Round 1 vote share for the incumbent (Kocharian)
- `mon_voting` - Whether the polling station was monitored in round 1 of the election
- `turnout` - Proportion of registered voters who voted in Round 1
- `totalvoters` - Total number of registered voters recorded for the polling station
- `total` - Total number of votes cast in Round 1
- `urban` - Indicator for whether the polling place was in an urban area (0 = rural, 1 = urban)
- `nearNagorno` - Indicator for whether the polling place is near the Nagorno-Karabakh region (0 = no, 1 = yes)

Part A

Hyde describes the study as a "natural experiment," stating:

> "I learned from conversations with staff and participants in the OSCE observation mission to Armenia that the method used to assign observers to polling stations was functionally equivalent to random assignment. This permits the use of natural experimental design. Although the OSCE/ODIHR mission did not assign observers using a random numbers table or its equivalent, the method would have been highly unlikely to produce a list of assigned polling stations that were systematically different from the polling stations that observers were not assigned to visit. Each team's assigned list was selected arbitrarily from a complete list of polling stations." (p. 48) What makes this study a "natural experiment" and not a true experiment? What assumption must the study defend in order to identify the causal effect of election monitoring that would be guaranteed to hold in a randomized experiment?

Part B

For the purposes of this part, assume election monitors were assigned as the author describes - in a manner "functionally equivalent to random assignment." Assume that this is true. Using the difference-in-means estimator, estimate the average treatment effect of election monitoring on incumbent vote share in round 1. Provide a 95% asymptotic confidence interval using the Neyman variance estimator and interpret your results. Can we reject the null of no average treatment effect at the $\alpha = 0.05$ level?

Part C

Evaluate the author's identification assumptions by examining whether the treatment is balanced on three pre-treatment covariates: the total number of registered voters, whether a polling place was in an urban area, and whether the polling place was located near the Nagorno-Karabakh region (Kocharian's home region and a disputed territory between Armenia and Azerbaijan). Discuss your results. Are they consistent with the author's description of "as-if random" assignment? Do you believe that your estimator from Part B is unbiased for the true average treatment effect?