DS-UA 201
Causal Inference
Instructor: Parijat Dube
Due: July 21, 2022

Homework 1- 100 points

General Instructions

This homework must be turned in on Gradescope by July 21st 2022, 11:59pm. It must be your own work, and your own work only—you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone. Your homework submission must be written and submitted using Rmarkdown. No handwritten solutions will be accepted. You should submit:

1. A compiled PDF file named yourNetID solutions.pdf containing your solutions to the problems.
2. A .Rmd file containing the code and text used to produce your compiled pdf named yourNetID solutions.Rmd. Note that math can be typeset in Rmarkdown in the same way as Latex.

Please make sure your answers are clearly structured in the Rmarkdown file:

1. Label each question part(e.g. 3.a).
2. Do not include written answers as code comments.
3. The code used to obtain the answer for each question part should accompany the written answer.

Problem 1 - *Hit or Miss?* 25 points

One longstanding debate in the study of international relations concerns the question of whether individual political leaders can make a difference. Some emphasize that leaders with different ideologies and personalities can significantly affect the course of a nation. Others argue that political leaders are severely constrained by historical and institutional forces. Did individuals like Hitler, Mao, Roosevelt, and Churchill make a big difference? The difficulty of empirically testing these arguments stems from the fact that the change of leadership is not random and there are many confounding factors to be adjusted for.

In this exercise, we consider a natural experiment in which the success or failure of assassination attempts is assumed to be essentially random. Each observation of the CSV data set `leaders.csv` contains information about an assassination attempt. Table 2.8 presents the names and descriptions of variables in this leader assassination data set. The `polity` variable represents the so-called *polity score* from the Polity Project. The Polity Project systematically documents and quantifies the regime types of all countries in the world from 1800. The polity score is a 21-point scale ranging from 10 (hereditary monarchy) to 10 (consolidated democracy). The `result` variable is a 10-category factor variable describing the result of each assassination attempt.

1. **(5 points)** How many assassination attempts are recorded in the data? How many countries experience at least one leader assassination attempt? (The `unique()` function, which returns a set of unique values from the input vector, may be useful here.) What is the average number of such attempts (per year) among these countries?

2. **(5 points)** Create a new binary variable named `success` that is equal to `1` if a leader dies from the attack and `0` if the leader survives. Store this new variable as part of the original data frame. What is the overall success rate of leader assassination? Does the result speak to the validity of the assumption that the success of assassination attempts is randomly determined?

3. **(5 points)** Investigate whether the average polity score over three years prior to an assassination attempt differs on average between successful and failed attempts. Also, examine whether there is any difference in the age of targeted leaders between successful and failed attempts. Briefly interpret the results in light of the validity of the aforementioned assumption.

4. **(5 points)** Repeat the same analysis as in the previous question, but this time using the country's experience of civil and international war. Create a new binary variable in the data frame called `warbefore`. Code the variable such that it is equal to `1` if a country is in either civil or international war during the three years prior to an assassination attempt. Provide a brief interpretation of the result.

5. **(5 points)** Does successful leader assassination cause democratization? Does successful leader assassination lead countries to war? When analyzing these data, be sure to state your assumptions and provide a brief interpretation of the results.

Problem 2 - *ITE and ATE* 30 points

Consider a dichotomous treatment variable D (1: treated, 0: untreated) and a dichotomous outcome variable Y (1: death, 0: survival). Let the treatment be receiving a heart-transplant. Table 1 is a list of heart patients with their potential outcomes and observed treatment.

1. **(4 points)** Write down the Individual Treatment Effect (ITE) and observed outcome (assuming consistency) for each patient. List the members of the population which has a causal effect of the treatment.

2. **(4 points)** Calculate Average Treatment Effect (ATE) using the potential outcomes. Can we conclude that *"null hypothesis of no average causal effect"* true (i.e., ATE is null)?

3. **(4 points)** Calculate the average of ITEs for the population. Is ATE always the same as the average of ITEs for a population?

4. **(4 points)** When there is no causal effect for any unit in the population, we say the *sharp causal null hypothesis* is true? Does sharp causal null hypothesis imply the null hypothesis of no average causal effect? Does absence of ATE imply absence of ITEs in a population?

5. **(4 points)** Assuming consistency holds, estimate the mean associational difference in this case, i.e., $E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$. Is there association between treatment and outcome?

Consider an alternate scenario where there are two different treatment options available for heart patients: Treatment A and B. The health services are controlled by the government and they have a limited budget to spend on this treatment. Further they can only adopt any one of the two

Homework 1- 100 points

Unit	D_i	$Y_i(0)$	$Y_i(1)$
Rheia	0	0	1
Kronos	0	1	0
Demeter	0	0	0
Hades	0	0	0
Hestia	1	0	0
Poseidon	1	1	0
Hera	1	0	0
Zeus	1	0	1
Artemis	0	1	1
Apollo	0	1	0
Leto	0	0	1
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	0	1
Aphrodite	1	0	1
Cyclope	1	0	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

Table 1

treatments, not both of them. For a population of 10 million patients, 2 million would die within five years if they receive treatment A, and 3 million would die within five years if untreated. For the same population, 1 million would die within five years if they receive treatment B, and 3 million would dies within five years if untreated.

1. **(2 points)** What is the ATE for treatments A and B?
2. **(4 points)** On average, how many patients need to be treated under treatment A to save 1 life? Similarly, on average How many patients needs to be treated under treatment B to save 1 life?
3. **(4 points)** Let the cost of treatment A for 1 patient be \$10 thousand and the cost of treatment B for 1 patient be \$4 thousand. If the total budget to spend is \$10 million, which treatment option should the government prefer in order to save as many lives as possible in the next five years.

Problem 3 - *ATT and ATE* 20points

Suppose we have a sample of observations, each assigned a binary treatment $D_i \in 0, 1$ with $D_i = 1$ indicating a unit is treated and $D_i = 0$ indicating the unit is assigned control. Assume $0 < Pr(D_i = 1) < 1$. We observe an outcome Y_i for each observation. Define potential outcomes

$Y_i(1)$ and $Y_i(0)$, denoting the outcome observed for unit i if it were assigned treatment ($Y_i(1)$) or control ($Y_i(0)$) respectively. Assume that $Y_i(1)$, $Y_i(0)$ are iid from the same distribution, which implies that: $E[Y_1(d)] = \cdots = E[Y_n(d)] = \mu(d)$ and $Var[Y_1(d)] = \cdots = Var[Y_n(d)] = \sigma^2(d)$

In class we have talked about the Average Treatment Effect (ATE), which is defined as $\tau = E[Y_i(1) - Y_i(0)]$. Another very common effect of interest to researchers is the Average Treatment Effect on the Treated (ATT), which is defined as:

$$\tau^t = E[Y_i(1) - Y_i(0)|D_i = 1].$$

1. **(4 points)** What is the interpretation of the ATT? Give your description of what this effect means and how it is different from the ATE.

2. **(8 points)** Assume that:

   - Consistency holds for all treatment levels: $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$
   - Weak ignorability holds **only** for the control outcome, i.e.: $Y_i(0) \perp\!\!\!\perp D_i$, and it is **not true** that: $Y_i(1) \perp\!\!\!\perp D_i$

   Show that consistency of all treatment levels, and weak ignorability of the control condition (the assumptions just made) are enough to identify the ATT, i.e., show that:

   $$\tau^t = E[Y_i(1)|D_i = 1] - E[Y_i(1)|D_i = 0]\ .$$

3. **(8 points)** Write and simplify the difference between the ATE and the ATT under the same assumptions as in previous part. What additional assumption is necessary for this difference to be 0, and for the ATT to be equal to the ATE? Why is this assumption enough?

Problem 4 - *Bernoulli Trial and ATE* 25 points

Under the same setting as Problem 3, suppose that D_i is assigned to the n units in a Bernoulli trial, that is each unit receives treatment independently with probability $Pr(D_i = 1) = p$.

Part a (10 points) Recall that the number of treated units, N_t is defined as $N_t = \sum_{i=1}^n D_i$. Recall that in the case of a Bernoulli trial, N_t is a **random variable**.

1. What distribution does D_i follow?

2. What is $E[N_t]$?

3. What is $Var[N_t]$?

4. Suppose that we wanted the **expected** number of treated units in our Bernoulli trial to be the same as the number of treated units as a completely randomized experiment with n_t treated units. What value of p should we choose?

Part b (15 points) After conducting the experiment as described above, we wish to estimate the ATE. To do so, we employ the following estimator:

Homework 1- 100 points

$$\hat{\tau}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} (Y_i \frac{D_i}{p} - Y_i \frac{1 - D_i}{1 - p})$$

Show that under consistency, positivity, and ignorability for all treatments this estimator is unbiased for the ATE.