PROBLEM SET 1

Causal Inference (HT 2025)

Due by 5:00 PM on Friday of Week 4

Instructions:

You must submit two files –

1. A PDF file with your write-up and results, including neatly formatted tables and figures.
2. An R Mardkdown file that generates this PDF, and contains your code for analysis.

Ensure that the files are named PS1.pdf and PS1.Rmd. Do not include any identifying information to ensure an anonymous grading process. Canvas will automatically assign anonymised numbers.

Collaborating with other students is permitted, under three conditions:

1. You must first try to answer each problem on your own before meeting with classmates.
2. Every student is required to produce their own code and write-up. No copy-pasting!
3. You must indicate who you collaborated with: students working together should come up with a random five digit number and mention this group ID in their PDF submissions.

Problem 1

Total points: 40

We will examine how randomized experiments work by creating an imaginary experiment. Use the dataset `a` from the file called `experiment.Rda`. For each individual unit (i) in our sample, the dataset contains the potential outcome under control (Y_i^0 or Y_{0i}) and the potential outcome under treatment (Y_i^1 or Y_{1i}) in the columns `a$y0` and `a$y1`, respectively. This is a purely hypothetical scenario. In reality, we never observe potential outcomes under both treatment and control for the same units: we can only observe one of them (the fundamental problem of causal inference). By creating a randomized experiment with this dataset, we'll demonstrate how experiments overcome this fundamental problem.

1. Find the "true" Average Treatment Effect across all units. *[5 points]*

Next, we'll implement a randomized experiment on this sample of 100 units. We randomly assign half of the units to treatment and half to control by creating a new variable indicating treatment status (D_i). As discussed in class, we can do this using the following steps:

- Input the command `set.seed(1)` so that results are replicable
- Assign each unit a random number between 1 and 100 – create a new column in the dataset named `rand`, using the `sample()` command and a vector of the numbers 1 to 100
- Re-order the dataset from lowest to highest value of rand using the code:
  `a <- a[order(a$rand),]`
- Create a treatment variable named `tr` that equals 1 for the first 50 units and 0 for the second 50 using the code `c(rep(1,50), rep(0,50))`

For this randomized experiment:

2. Conduct a test to assess whether the treatment and control groups have the same average potential outcomes under control. Has randomization succeeded in creating treatment and control groups with equivalent potential outcomes under control? Why? *[5 points]*
3. Estimate the Average Treatment Effect based on your experiment. How similar is it to the "true" Average Treatment Effect? Explain. *[10 points]*

Now, let's see how the experimental procedure performs over repeated randomizations. To simulate repeated randomizations, use the following steps:

- Step I. Create a function that takes in the dataset `a`, carries out a randomized experiment, and reports the estimated ATE.
  - You can do this by combining: code provided above to implement a randomized experiment **without** using `set.seed()`, and your code for the estimated ATE from Task (3).
- Step II. Find the results of 10,000 randomized experiments, i.e., run your function 10,000 times, and store these results in a variable. You can do this using the `replicate()` command.

4. What is the average estimated ATE across your 10,000 experiments? Does this suggest that your estimator is unbiased? Why? *[10 points]*
5. Repeat Task (4), calculating the mean difference in potential outcomes under control (`a$y0`) between the treatment and control groups instead of the ATE. What is the mean difference from your 10,000 experiments? What does this signify? *[10 points]*

Problem 2

Total points: 20

Past research suggests that ballot secrecy influences turnout. A recent field experiment sent emails to a random group of nonvoters around the 2014 election in Mississippi, reminding them that their vote was secret. Use the file called `ballot.csv` that contains the data collected through this experiment. Table 1 contains a description of variables in this dataset.

Table 1: Variables in ballot.csv

Variable	**Description**	**Type**
vote2014	Did the respondent vote in 2014? 1 = Yes, 0 = No	(Outcome)
mail	1 = Received mail, 0 = Did not receive mail	(Treatment)
vote_year	Voting in the relevant year	
d_age	Respondent's age	
d_gen_female	Respondent's gender: 1 = Female, 0 = Male	
d_race_blk	Respondent's race: Black	
d_race_hsp	Respondent's race: Hispanic	
d_race_other	Respondent's race: Other	
never_voted	Respondent has never voted	

For this exercise, we wish to establish if the results of this experiment hold when we focus on the female subsample (instead of the full sample).

1. Confirm that the randomization process was successful by making sure that women in treatment and control groups are similar in all relevant aspects, e.g., age, ethnicity, non-voting habits. Show your results either using a figure or by producing a publishable table. *[10 points]*
2. Estimate the Average Treatment Effect and test whether the effects you calculated are robust to the inclusion of covariates. Report all results in a single table. Is it necessary to include covariates when calculating the ATE? *[10 points]*

Problem 3

Total points: 40

A key problem incumbents encounter in civil wars is a lack of information to combat insurgency. Insurgents exploit information asymmetries at the local level to hide and become a difficult target for incumbents. In the absence of such information, incumbents often resort to indiscriminate violence, via large-scale reprisals against entire villages suspected to host insurgents. One example of indiscriminate violence is Aerial bombardment. Due to the nature of insurgency, bombing frequently occurs in and around settled areas, and leading to many civilian casualties.

Using data from the Vietnam War, Kocher, Pepinsky and Kalyvas (2011) examined the effect of bombings on Viet Cong support. In particular, they looked at the impact of the September 1969 bombings on hamlet control in December 1969. The data is available in the file `Vietnam_matching.dta`. A description is provided in Table 2.

Table 2: Variables in Vietnam_matching.dta

Column name	**Description**
bombed_969	Number of bombings per hamlet in September 1969.
bombed_969_bin	Was there a bombing in the hamlet in September 1969? 1 = Yes, 0 = No
std	Rough terrain
ln_dist	Hamlet's distance from closest international boundary (logged)
score	Development index score
lnhpop	Hamlet's population (logged)
mod2a_1adec	Enemy Military Model (2A) in December 1969.
mod2a_1ajul	Enemy Military Model (2A) in July 1969.
mod2a_1admn	District-wise average Hamlet Control before September 1969.

Note:
The "Enemy Military Model (2A)" or Hamlet Control is a rating of the presence and activity of Viet Cong military units in the vicinity of each hamlet on a 5-point scale, where: 1 = fully government controlled, 2 = moderately government controlled, 3 = contested, 4 = moderately insurgent controlled, and 5 = fully insurgent controlled.

The data comes from various sources. The United States compiled a gazetteer of South Vietnamese hamlets, identified their geographic coordinates, and conducted a census. District Senior Advisors (DSAs, army officers ranking major or above) were assigned to complete detailed questionnaires for every village and hamlet in their zones of operation. Some of these questionnaires were compiled monthly, others quarterly. DSAs were detached from U.S. units to live and work in the districts they

rated. The RVN (Republic of Viet Nam) had 261 districts with a median area of 377 kilometers squared, or about one-fourth the size of the median U.S. county. There was a median of 36 hamlets per district in 1969.

In this exercise, we will explore how matching models work, using the dataset in `Vietnam_matching.dta`. We will apply the matching procedure to estimate the effect of *experiencing a bombing* in September 1969 on insurgency control in December 1969.

1. Which covariates should you use in the matching procedure and why? *[10 points]*
2. Choose a matching estimator – briefly describe your choice. Assess balance in pre-treatment covariates between treated and control units, before and after matching. What is the measure you used for assessing balance? *[10 points]*
3. Estimate the causal effect of interest with matching. Are the results from your matching analysis different from using a simple OLS regression *without* covariates? If there is a difference: how do you explain it? *[10 points]*
4. How does the comparison between matching and OLS regression change when you include covariates (control variables)? Report the models in a single table and discuss. *[10 points]*