Homework 4

Due date: 10/22/2024

October 10, 2024

# Introduction

This exercise sheet contains a series of problems designed to test and enhance your understanding of the topics covered in the course. Please ensure that you attempt all problems and provide detailed solutions where necessary. If you have any questions or need clarification, feel free to reach out your TA.

# Exercises

## Exercise 1: Fire VS Water

Read the **fireVswater pokemon.csv** available on your canvas page. The dataset presents the weight (*Y* ) of a sample of water and fire pokemons (*W*) across the first *nX* = 7 generations (*X*).

1. Show with R if the groups (Water-Gen1, Fire-Gen1, Water-Gen2, ...) are balanced.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

The subgroups within each generation are not balanced because they feature notably different covariate means except for those in the first generation.

1. Compute the mean and the standard deviation for each sub-sample (Water-Gen1, Fire-Gen1,Water-Gen2, ...). Plot the evolution of the means for fire and water pokemons through generations and highlight any specific trend.

A screenshot of a computer code

Description automatically generated

A screenshot of a data table

Description automatically generated

A graph of different colored squares

Description automatically generated

Water type Pokémon weigh slightly more than fire type Pokémon on average. That said, fire type Pokémon weigh a lot more than water type Pokémon in the seventh generation.

1. We want to study the difference of weight between fire and water types. The following estimator has been proposed: . Describe in words the estimator and use R to apply it to the dataset.

This estimator is calculating the difference between the mean weight of fire Pokémon and the mean weight of water Pokémon. This estimator does not condition on generation.

A screenshot of a computer code

Description automatically generated

1. Another estimator has been proposed: )). Describe the estimator in words and use R to apply it to the dataset.

This estimator is looking at the average difference in mean weight between fire type and water type Pokémon across all generations.

A screenshot of a computer program

Description automatically generated

1. Another estimator has been proposed: )). Describe the estimator in words and use R to apply it to the dataset.

This estimator is calculating the difference in mean weight between fire type and water type Pokémon by multiplying it by the number fire type Pokémon in each generation and then averaging this value over the total number of fire type Pokémon across all generations. The purpose of this estimator is to calculate how much fire type Pokémon weighs on average compared to water type Pokémon.

A screenshot of a computer program

Description automatically generated

1. Compare the estimates obtained in (c), (d), and (e). Are they the same? If they are not tryto deduct why (*Hint: (a) and plotting the data could be useful*).

The estimators in (c), (d), and (e) are all different from each other. The estimate in (e) is much larger than the estimates in (c) and (d). This is because there are covariates in the data and none of the groups are balanced. Therefore, there is a covariate shift. Furthermore, each estimator is measuring something different. Estimator 1 measures the difference in mean weight between fire type and water type Pokémon without conditioning on generation. Estimator 2 measures the difference in mean weight between fire type and water type Pokémon by conditioning on the generation. Finally, estimator 3 measures on average how the weight of fire type Pokémon compare to the weight of water type Pokémon across all generations.

1. Using bootstrapping create a confidence interval for the estimate resulting from estimator (e).

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

## Exercise 2: Village Leaders and Sanitation Budget

In 2004, Chattopadhyay and Duflo studied the budgetary consequences of having women, rather than men, lead village councils in West Bengal and Rajasthan, India. Each village either receives treatment (has a woman council leader) or is untreated (has a man council leader), then we observe the share of the local council budget that is allocated to providing drinking water. For the purposes of this problem, pretend we can observe the potential outcomes of the villages—how much of their budget is allocated toward water sanitation—when their council is led by a woman (*Yi*(1)) or a man (*Yi*(0)). These potential outcomes are shown in the table below, as well as columns for true (unobserved) individual treatment effects, *τi*, an indicator of whether or not a village receives treatment (*Di*), and the *Yi* we observe based on this treatment status.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Village | *Yi*(0) | *Yi*(1) | *τi* | *Di* | Observed *Yi* |
| Village 1 | 10 | 15 | 5 | 1 | 15 |
| Village 2 | 15 | 15 | 0 | 0 | 15 |
| Village 3 | 20 | 30 | 10 | 0 | 20 |
| Village 4 | 20 | 15 | -5 | 0 | 20 |
| Village 5 | 10 | 20 | 10 | 0 | 10 |
| Village 6 | 15 | 15 | 0 | 0 | 15 |
| Village 7 | 15 | 30 | 15 | 0 | 15 |
| Average | 15 | 20 | 5 | 0.14 | 15.71 |

1. Fill in the table with each village’s individual treatment effect *τi* and add its average to the bottom of the table.
2. Using your omniscient powers of observing all potential outcomes, what is (1)? What is (0)? What is

The first estimator is looking at the mean budget allocation if all the villages were treated, which is equal to 20. The second estimator is looking at the mean budget allocation if none of the villages were treated, which is equal to 15. Finally, the last estimator is looking at the average treatment effect by looking at the difference between the treatment and the control groups.

1. It appears that only Village 1 was assigned to treatment. Fill in the observed *Yi*s in the table based on this treatment assignment.
2. What estimator will you use to calculate the average treatment effect estimate using observed*Yi*s? Calculate your estimate.

I will use the third estimator to calculate the average treatment effect because it calculates the difference in mean budget allocation between villages that were treated and villages that were untreated. Then I can simply do this for observed Yis.

A math equations on a piece of paper

Description automatically generated

1. Is your estimator from part (d) causally identified? Why or why not?

My estimator from part (d) is not causally identified because the treatment is not randomized, and it is not conditioned on any covariate variable. Therefore, this estimator is missing additional information from which to draw a more accurately representative estimate.

## Exercise 3: Conditional Randomization Practice

Below is a table of data from an experiment testing a new diet dog food on chihuahua weight. The outcome of interest, *Yi*, is the chihuahua’s weight in pounds after the experiment. *Ti* indicates whether the chihuahua was treated with the new diet food (1) or not (0); this was randomized conditioned on the health of the chihuahua at the beginning of the study (*Zi*). We also collected information about the chihuahua’s fur type, *Xi*, which we coded as 1 if the chihuahua had long hair and 0 if it had short hair.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Chihuahua | *Zi* | *Xi* | *Yi* | *Ti* |
| 1 | Obese | 0 | 8 | 1 |
| 2 | Obese | 1 | 6 | 1 |
| 3 | Obese | 1 | 6 | 0 |
| 4 | Overweight | 0 | 5 | 1 |
| 5 | Overweight | 1 | 6 | 1 |
| 6 | Overweight | 0 | 6 | 0 |
| 7 | Overweight | 1 | 7 | 0 |
| 8 | Overweight | 0 | 5 | 0 |
| 9 | Healthy Weight | 0 | 4 | 1 |
| 10 | Healthy Weight | 1 | 3 | 1 |
| 11 | Healthy Weight | 0 | 4 | 1 |
| 12 | Healthy Weight | 0 | 5 | 0 |
| 13 | Healthy Weight | 1 | 3 | 0 |
| 14 | Healthy Weight | 1 | 3 | 0 |

1. Calculate the *CATE*ˆ for each level of *Xi*. Calculate ˆ*π*(*X*) for each level of *Xi*.

A white paper with writing on it

Description automatically generated

1. Are these estimators *CATE*ˆ *X* causally identified? Why or why not? If so, calculate a standardized *CATE*ˆ *X*.

These estimators are not causally identified because our sample was not randomly conditioned on fur type.

1. Calculate the *CATE*ˆ for each level of *Zi*. Calculate ˆ*π*(*Z*) for each level of *Zi*.

A piece of paper with math equations

Description automatically generated

1. Are these estimators *CATE*ˆ *Z* causally identified? Why or why not? If so, calculate a standardized *CATE*ˆ *Z*.

These estimators are causally identified because our sample was randomly conditioned on the health of the chihuahua at the beginning of the study.

1. Calculate the difference in sub-sample mean weight among treated and untreated chihuahuas(i.e. ˆ*µ*(*T* = 1) − *µ*ˆ(*T* = 0)). Is this estimator causally identified? Why or why not?

A white board with text

Description automatically generated with medium confidence

This estimator is not causally identified because treatment is not random, and it is not being conditioned for any potential covariates.

## Exercise 4: Voter Turnout Experiment Analysis

Use the data ”GGLsample.csv” for a sample of the data from Gerber, Green, and Larimer’s voter turnout experiment. Each voter was marginally randomized into treatment or control. The relevant variables are sex, which records the individual’s sex; treatment, which records whether the individual received the ’Neighbors’ treatment– a postcard indicating their neighbors’ voting records and that their neighbors will be informed of whether they vote in the upcoming election– or control; bintreat, a binary recoding of treatment; and voted, the outcome of interest– whether the individul voted in the election following the experiment.

1. Estimate the *CATE*ˆ *male*. Is this estimator causally identified? Why or why not?

A screenshot of a computer

Description automatically generated

This estimator is not causally identified. While this estimator adjusts for the relevant covariate of sex, therefore reducing the bias in the assignment of treatment, there are still too many other variables to adjust for such as education and income. Therefore, there is too much information we do not know. As such, we cannot determine whether the treatment is truly randomized and our estimator is not causally identified.

1. Use a bootstrap estimated sampling distribution to create a 95% confidence interval for the*CATEmale* and interpret your interval in plain language.

A screenshot of a computer

Description automatically generated

If we repeated our *CATE* estimator many times with a sufficiently large sample size, our intervals constructed this way would include a slight increase between 0.013 and 0.114 voting outcome for males when conditioning on sex approximately 95% of the time.

1. Estimate the *CATE*ˆ *female*. Is this estimator causally identified? Why or why not?

A screenshot of a computer

Description automatically generated

This estimator is not causally identified. While this estimator adjusts for the relevant covariate of sex, therefore reducing the bias in the assignment of treatment, there are still too many other variables to adjust for such as education and income. Therefore, there is too much information we do not know. As such, we cannot determine whether the treatment is truly randomized, and our estimator is not causally identified.

1. Use a bootstrap estimated sampling distribution to create a 95% confidence interval for the*CATEfemale* and interpret your interval in plain language.

A computer screen shot of a computer

Description automatically generated

If we repeated our *CATE* estimator many times with a sufficiently large sample size, our intervals constructed this way would include a slight increase between 0.04 and 0.144 voting outcome for females when conditioning on sex approximately 95% of the time.

1. Estimate the *ATE*ˆ . Is this estimator causally identified? Why or why not?

A screenshot of a computer code

Description automatically generated

This estimator is not causally identified because it does not adjust for covariates. Therefore, the treatment in this estimator is not random and cannot be causally identified.

1. Use a bootstrap estimated sampling distribution to create a 95% confidence interval for the *ATE* and interpret your interval in plain language.

A screenshot of a computer code

Description automatically generated

If we repeated our *ATE* estimator many times with a sufficiently large sample size, our intervals constructed this way would include a slight increase between 0.04 and 0.114 voting outcome approximately 95% of the time.

# Submission Instructions

Please submit your completed exercises by **October 22** through **gradescope**. Ensure that your solutions are well-organized, clearly written, and include all necessary calculations and explanations. Questions about submission should be directed to your TA.

# Helpful Resources

To better assist you in the completion of this exercise sheet, we suggest you to review the following material:

* **Lecture 3** - bootstrapping estimated sampling distributions and confidence intervals
* **Lecture 6** - estimators
* **Lecture 9** - covariate shift and different estimator options;
* **Lecture 10** - causal identification for marginally and conditionally randomized experiments;
* **Lab** - practicing all of the above