FIN 550: Big Data Analytics

Problem Set #1

Select whether this is an individual or group submission. No more than 3 members per group. Beyond the fact that all group members may submit the same answers, each submission must be separate work.

⧠ Individual Submission  
⧠ Group Submission.

List group member names: Ives He, Xuzhe Yu, Hanbin Yan

**Problem set deliverables**

You should submit the following three files as part of your problem set solution:

1. A completed version of this file, containing group member names and solutions to Problems 1 and 2.
2. A file named “Case-Executive-Summary.pdf” with the executive summary report for Problem 3.
3. An R script named “Case-Code.R” for Problem 3.

# Causal Treatment Effects (15 points)

**Table 1: Health Outcomes and Treatments**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Esther** | **John** | **Timothy** | **Ruth** | **Daniel** |
| Potential outcome if not treated: *Y0i* | 4 | 3 | 2 | 4 | 5 |
| Potential outcome if treated: *Y1i* | 4 | 5 | 4 | 5 | 5 |
| Treated | No | Yes | Yes | No | No |
| Observed health outcome | 4 | 5 | 4 | 4 | 5 |
| Treatment effect | 0 | 2 | 2 | 1 | 0 |

1. Based on the information given, fill your answers in the blanks of Table 1 (imaginary table!) for a group of five individuals. Note that health outcome is measured by an index 1-5, where 1=poor and 5=excellent.

2. What is the average treatment effect among individuals who are treated?

Average Treatment Effect = (Y1\_John - Y0\_John + Y1\_Timothy - Y0\_Timothy) / 2 = (2 + 2) / 2 = 2

3. Calculate the difference in group means between the treatment and control groups. Is this a measure of average causal treatment effects? Why or why not?

The mean outcome for the treated group is: (5 + 4) / 2 = 4.5

The mean outcome for the control group is: (4 + 4 + 5) / 3 = 4.33

Difference in Means = 4.5 - 4.33 = 0.17

This measure is not a valid measure of the average causal treatment effect because it does not account for potential confounding variables, which could bias the results.

4. Calculate selection bias in the prior measure (difference in group means between the treatment and control groups).

Selection Bias = Average Treatment Effect - Difference in Means = 2 - 0.17 = 1.83

5. Using only data on actual health outcomes, how could we eliminate selection bias? (Assume we can do whatever we want, including forcing people to be treated or not or getting information on more individuals can collect health outcomes on more individuals if desired.)

To eliminate selection bias, we can use random assignment to treatment and control groups. Random assignment minimizes the impact of unobserved confounding variables, ensuring that any differences observed between the groups are likely to be due to the treatment itself. Alternatively, increasing the sample size and adjusting for covariates can also help reduce selection bias.

# True/False (15 points)

For each of the following points, state whether the **boldface statement** is true or false, and explain why in 1-3 sentences. No credit will be awarded without a valid explanation. The questions are meant to be straightforward in the sense that you should be able to apply basic concepts covered in class to determine the answer.

1. John and Joe are identical twin brothers separated at birth. John and Joe have the same IQ and other innate abilities. Also, their adoptive parents are identical in terms of income and education levels. John and Joe just reunited after 25 years and found out that John’s earnings are 20% higher than Joe’s. John has a college degree while Joe doesn’t. **As a result, this difference in earnings reflects the causal effect of college on earnings.**

Statement: This difference in earnings reflects the causal effect of college on earnings.

Answer: False.

Explanation: While college may influence earnings, we cannot conclude causation solely based on this observation. Differences in earnings could also be attributed to other environmental or social factors that were not controlled for, despite their similarities.

1. Research claims that dental insurance is a primary factor that determines dental service utilization. In order to estimate the effect of dental insurance on utilization, the following linear regression results were produced using a random sample from a pool of individuals who had private dental insurance (i.e., treatment group). The control group was randomly picked from a pool of individuals without dental insurance. **The results provide evidence of the causal effect of dental insurance on dental care utilization.**

|  |  |  |  |
| --- | --- | --- | --- |
| Outcome: Number of dental services received | | | |
|  |  | | |
| Controls | Coefficient | | Standard Errors |
| **Dental Insurance**  **= 1 if individual has insurance**  **= 0 if not** | **1.14** |  | **(0.28)** |
| Nonwhite | -0.04 |  | (0.31) |
| Female | -0.01 |  | (0.26) |
| Education | 0.59 |  | (0.25) |
| Married | 0.42 |  | (0.33) |
| Have children | -0.39 |  | (0.24) |
| Employed | 0.87 |  | (0.25) |
|  |  |  |  |
| Sample size | 1,157 | | |

Statement: The results provide evidence of the causal effect of dental insurance on dental care utilization.

Answer: True.

Explanation: Since the study used a random sample with both a treatment and control group, the observed difference in utilization can be attributed to the presence of dental insurance, suggesting a causal relationship. Random assignment helps to control for confounding factors.

1. Many things may impact an individual’s earnings. These include an individual’s innate ability, their education and developed skills, their social network, and apparent luck. For example, it is hard to know if students who take “FIN 550: Big Data Analytics” ultimately make more money because the course causes them to earn more, or because of all the other things about these students that are exceptional to begin with. **However, if it were possible to randomly assign some Master’s students into FIN 550 while randomly blocking others, the difference in future earnings and promotions between these two groups would reflect the average causal effect of the course.**

Statement: If it were possible to randomly assign some Master’s students into FIN 550 while randomly blocking others, the difference in future earnings and promotions between these two groups would reflect the average causal effect of the course.

Answer: True.

Explanation: Randomly assigning students to take the course would allow us to control for pre-existing differences. This design would likely yield an accurate estimate of the course’s causal effect on future earnings and promotions.

1. When an estimate is not equal to the true value of the estimand (the quantity of interest), we can conclude that the estimate is biased.

Statement: We can conclude that the estimate is biased.

Answer: False.

Explanation: An estimate that differs from the true value may result from random error or sampling variability, not necessarily bias. Bias occurs when there is a systematic deviation in the estimation process, not merely a difference from the true value.

1. In a randomized experiment, checking for balance refers to checking that a similar number of individuals were assigned to the treatment and control groups.

Statement: Checking for balance refers to checking that a similar number of individuals were assigned to the treatment and control groups.

Answer: False.

Explanation: Checking for balance involves ensuring that covariates (e.g., age, gender, income level) are similar across treatment and control groups, not merely that the group sizes are similar. This balance is essential for attributing any difference in outcomes to the treatment rather than other factors.

# Illinois Workplace Wellness Study (70 points)

Complete the data case, “Does Workplace Wellness Work? The Illinois Workplace Wellness Study.” The case is available on Canvas. The case deliverables—an executive summary and R script—should be included with your problem set solutions.