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: Javeria Malik, Ahmed Bilal, Shehzadi Mahum Agha

**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**

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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **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. What is the average treatment effect among individuals who are treated?

(2 + 2)/2 = 2

1. 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?

**Treated Group Mean: (**5+4)/2 = 4.5

**Control Group Mean: (**4+4+5)/3= 4.33

**Difference:** 4.5 – 4.33 = 0.17

No, the difference in the outcomes cannot be understood as the true causal treatment effect because the observed outcomes include selection bias. Individuals who received the treatment would be systematically different from those who did not.

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

**Selection Bias =** Y01(Treatment Group) – Y0i(Control Group)

(3+2)/2 – (4+4+5)/3 = 1.83

1. 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.)

We could use the following techniques to eliminate selection bias:

* Randomized Control Trial (RCT) to forcefully assign individuals to either a treatment or control group randomly to make sure the differences are only for the treatment in both groups which will then balance both observed and unobserved factors across treated and untreated groups to avoid selection bias.
* We should ensure that the division between treatment and control group is completely random

# 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.**

**-> False**

Although John and Joe have many inherent similarities, the only difference of the college education is not enough to conclude causal effects of college education on earnings. Earnings can be different for several reasons like location, intelligence, skillset, industry, etc., so these aspects lack control making it incorrect to conclude a causal effect of college on earnings based solely on this comparison.

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 | | |

**-> True**

The linear regression includes a treatment group with dental insurance and a control group without it, and the coefficient for "Dental Insurance" is 1.14 with a standard error of 0.28. Given the random assignment of individuals to treatment and control groups, we can attribute the observed difference in utilization to the presence of dental insurance, supporting a causal interpretation.

We can further find out 95% confidence interval to check effect of dental insurance on dental service utilization. The formula for the 95% confidence interval is:

Coefficient ± 1.96 × Standard Error

Substitute the values:

1.14 ± 1.96 × 0.28 = 1.14 ± 0.5488

So, the 95% confidence interval is approximately:

[ 0.5912, 1.6888]

This interval does not include zero, reinforcing that the effect of dental insurance on dental care utilization is statistically significant and likely positive.

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.**

**-> True**  
Randomly assigning master's students to take FIN 550 versus not taking it would eliminate selection bias by ensuring that both groups are comparable in all respects except for their course enrollment. As a result, any difference in future earnings and promotions could be attributed to the effect of the course, providing an estimate of the average causal effect.

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.

**-> False**  
An estimate can differ from the true value due to random sampling error, not necessarily because of bias. Bias refers to a systematic difference between the expected value of an estimator and the true value, whereas random differences in individual estimates from the true value are normal and expected in finite samples.

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.

**-> False**  
In a randomized experiment, checking for balance involves verifying that baseline characteristics (such as age, gender, and other relevant variables) are similar across treatment and control groups, not just the number of individuals in each group. Balance in characteristics helps ensure that any observed treatment effects are not confounded by pre-existing differences between the groups.

# 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.