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:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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?  
Ans: 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?  
(1) 1/6. Treatment group means : 9/2, control groups means : 13/3. “9/2 - 13/3 = 1/6”  
(2) No, this is not a measure of the average causal treatment effects. This calculation is the difference in the average observed health outcomes between the treatment group and the control group. The average causal treatment effects is Avg(Y1i | Ti = 1) – Avg(Y0i | Ti = 1).

4. Calculate selection bias in the prior measure (difference in group means between the treatment and control groups).  
Selection bias = Avg(Y0i | Ti = 1) – Avg(Y0i | Ti = 0).   
Avg(Y0i | Ti = 1) : (3+2)/2 = 5/2  
Avg(Y0i | Ti = 0) : (4+4+5)/3 = 13/3  
5/2 – 13/3 = -11/6

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.)  
randomization。把所有人打散，把原本有保險的，迫使一半的人保險，一半的人不保險，原本沒保險的，同理。

# 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. Because there are still some factors could not be controlled, such as living location. Hence, it’s difficult to obtain the causal treatments effect from the individual.
2. 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 | | |

False. Because it included the others factors, we can not ensure the causal effect of dental insurance on dental care utilization.

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. This experiment randomly assigning and blocking master student is called randomization. And in the randomization, it can eliminate selection bias. In addition, we can use balance test to make sure whether randomization worked.
2. 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, there are several reasons leading estimate is not equal to estimand, such as a biased estimator, insufficient sample size, or sampling error.
3. 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, the number of individuals is one of the indicators needed to be checked, and there are others possible factors indicators needed be checked, such as age, education, etc.

# 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.