FIN 550: Big Data Analytics
Problem Set #1 Problem 1 and 2

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:

Hsiang Lee, Chu-Chun Ku, Yi-Ting Lee

1 Causal Treatment Effects (15 points)

Table 1: Health Outcomes and Treatments

	Esther	John	Timothy	Ruth	Daniel
Potential outcome if not treated: Y_{0i}	4	3	2	4	5
Potential outcome if treated: Y_{1i}	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?

   ATE= (0+2+2+1+0)/5=1
   TOT=(2+2)/2=2
   TOT>ATE because treatment effects are higher on average for treated individuals (i.e., John and Timothy) compared to treatment effects across the entire group. This could occur because individuals with higher benefits from the treatment are the ones who chose to get it.

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?

   Treatment group: (5+4)/2 = 4.5
   Control group: (4+4+5)/3 = 4.3
   Difference: 0.2
   This is not the measure of average causal treatment effects. We don't know if these people share other characteristics besides the treatment, so selection bias might happen.

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

   Treatment group: (3+2)/2 = 2.5
   Control group: (4+4+5)/3 = 4.3
   Difference: -1.8

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.)

   Use random assignment and ensure a sufficiently large sample size.

2 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. The income difference between John and Joe alone cannot confirm a causal effect of college education. Although they share similar IQ and family background, other factors may influence their earnings, such as local job market, life experience …etc. Without random assignment, this comparison may involve selection bias.

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. While the dental insurance dummy is statistically significant, this significance could have been driven mainly by the fact that individuals with bad teeth or risk aversion are the ones who choose to get dental insurance (i.e., selection bias), and they used dental services significantly more than individuals without dental insurance. This does not mean that dental insurance "caused" the treatment group to use more dental services.

3. 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. By randomly assigning Master's students to either take the FIN 550 course or be excluded from it, we can make sure that students in the treatment and control groups are randomly assigned. This helps remove other possible factors, so we can better measure the direct causal effect of the course on income.

4. 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. We cannot conclude that the estimate is biased, because sometimes the difference between a single estimate and the true value is due to random variation, not bias. We should increase the sample size to see whether the average of the estimates remains biased or gets closer to the estimand.

5. 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 other possible indicators needed to be checked, such as age, education, etc.