New York University Shanghai

Problem Set 7

BUSF-SHU 210: Business Analytics (Spring 2019)

1. Omitted Variable Bias

A new treatment for a disease is introduced and we want to examine its effectiveness compared with the existing standard of care (control). Let $W = 0, 1$ denote control or treatment, respectively; and $Y = 0, 1$ denote the outcome. $Y = 1$ refers to that the (standard or new) treatment is effective, whereas $Y = 0$ refers to that the (standard or new) treatment is ineffective.

You run an experiment with a large sample size of 10000 patients, 5000 in the control group and the other 5000 in the treatment group. The results of the experiment are reported in the following table (here the numbers are the number of individuals in each category):

	Treatment ($W = 1$)	Control ($W = 0$)
Male Effective ($Y = 1$)	1500	375
Male Ineffective ($Y = 0$)	2250	875
Female Effective ($Y = 1$)	1000	2625
Female Ineffective ($Y = 0$)	250	1125

Please answer the following questions:

(a) What is the estimated average treatment effect for this experiment $\widehat{ATE}$?

(b) Now we consider male and female patients separately. What is the estimated average treatment effect for male, $\widehat{ATE}_M$? What is the estimated average treatment effect for female, $\widehat{ATE}_F$?

(c) Compare the results of (i) and (ii), what interesting phenomenon have you found? Can you explain why such phenomenon would occur?

(d) What additional assumption(s) do we need about the experiment, in order to have a solid conclusion that this treatment is effective for both male and female?

2. Regression and Matching in Observational Studies

We examine the effect of college education on earnings. The data set `Schooling.csv` contains the earning, college, and IQ score information of randomly selected 10000 people.

- $LogEarning$ = Log of the individual's salary
- $College$ = Whether this individual has a college degree. $College = 1$ (resp. 0) means the individual has (resp. does not have) a college degree.
- IQ = The IQ score of this individual at age 18 (right after high school).

Hint: It is suggested that you normalize the IQ score of individual by subtracting its mean and dividing by its standard error, since the IQ score itself is recorded according at the percentile of an individual within a population. See https://en.wikipedia.org/wiki/Intelligence_quotient for more references.

(a) Run a linear regression using $LogEarning$ as the dependent variable and $College$ as the independent variable. What is the estimated treatment effect of college education based on this model? Do you think this is an unbiased estimation? If not, is it an under-estimation or over-estimation? Why?

(b) If you think the model built in part (i) gives us a biased estimation of the treatment effect of college education, please build another model to, partially, correct this bias. What is the estimated treatment effect of college education based on your new model? Please also discuss the rationale behind how you build this model.

(c) You believe that the IQ of an individual would influence the effect of college education on earning. Build a model that could reflect such influence. What is the estimated treatment effect of college education based on this model?

(d) In order that the regression analysis in (i), (ii), and (iii) could give us a solid causal inference conclusion, what additional assumption(s) we should have about the data set?

(e) Under the assumption of part (d), use propensity score matching to estimate the causal effect of college education on the (log of) future earning. Please remember to check the balance of treatment and control groups before and after matching.

3. Causal Inference with Instrumental Variable

Facebook is considering establishing a new feature that could facilitate the process of searching and connecting to a new friend. The company runs an online randomized experiment to examine the **whether users' adoption of this new feature could improve their satisfaction**. To do so, Facebook randomly selects a sample of 10,000 users. Within this sample, a random group of users are selected into the treatment group. Facebook sends an encouragement message to each individual in the treatment group. The encouragement message advocates the new feature and encourages the user to adopt it. The rest of the users in the sample are in the control group, to which Facebook sends nothing. Facebook cannot control which user will adopt their new feature, but believes that the adoption of the new feature is positively correlated with receiving the encouragement message. Users who are initially more satisfied with Facebook before the experiments will be more likely to adopt the new feature. At the end of the experiment, Facebook conducts a survey that asks each user in the experiment to report his/her satisfaction level of Facebook then. We assume that each user truthfully reports his/her satisfaction level of this online social media.

The data associated with the experiment described above is stored in `NewFeature.csv`. Each row in this data set represents a user. The data has 3 variables:

- *Encouragement*: An indicator of whether the user has received the encouragement. $Encouragement = 1$ if the user received the encouragement; $Encouragement = 0$ if the user did not receive the encouragement.
- $Satisfaction$: The reported satisfaction level. A higher value means the user is more satisfied. 0 means not satisfied at all; 10 means fully satisfied.
- *Adoption*: An indicator of whether the user has adopted the new feature. $Adoption = 1$ if the user adopted the new feature; $Adoption = 1$ if the user did not adopt the new feature.

Please briefly answer the following questions:

(a) Run a linear regression with $Satisfaction$ as the outcome and *Adoption* as the covariate. Assume $\hat{\beta}_A$ is the fitted coefficient of *Adoption*. What is the value of $\hat{\beta}_A$? Is $\hat{\beta}_A$ an unbiased estimation of the causal effect of *Adoption* on $Satisfaction$? Why or why not? If not, will this be an overestimation or an underestimation of the true causal effect?

(b) Facebook uses *Encouragement* as an instrumental variable to establish the causal effect of *Adoption* on $Satisfaction$. Discuss why *Encouragement* could be a valid instrumental variable in this setting.

(c) Use the two-stage least-square method to estimate the unbiased causal effect of *Adoption* on $Satisfaction$. Please obtain the estimation results using two different approaches, one with two linear regression steps, and the other with a single integrated step. We use $\hat{\gamma}_A$ to denote the true causal effect of *Adoption* on $Satisfaction$. What is the value of $\hat{\gamma}_A$? How do you interpret $\hat{\gamma}_A$?