Fall 2022
SS 340
Loujaina Abdelwahed

Submit your work at the beginning of class on Tuesday (September 27). You may work with other students but the write-ups must be unique.

Please submit your code (.do file if you use Stata or .ipynb if you use Python). Please annotate your code and explain what you are doing in each line. Please make sure that your code is easily replicable.

1. The attached file layoffs.csv includes monthly data on the number of employees laid off by companies in the retail sector. The variables included are: **(30 points)**

   - company: company name
   - notice year: year of the layoff notice
   - notice_month: month of the layoff notice
   - layoffs: number of employees laid off

   The dataset has 321 observations in total.

   a. Present summary statistics of all the number of the variable layoffs. Think about how to present your output to the reader (who is me), who hasn't seen the dataset before. Make sure to describe the variable of interest in writing.

   b. Test the hypothesis that the average number of monthly layoffs in retail over the past three years is 240 employees. Use Stata or Python for this – no need to conduct the test manually.

   c. Present summary statistics of the layoff variable by year, for each of the three years 2020, 2021 and 2022.

   P.S. In numbers b, please state your null and alternative hypotheses. You can conduct your test using any software and just eyeball the p-values. But in your write up, please write out the null and the alternative hypotheses and then your conclusion about whether you accept or do not accept the null. Also please provide a complete interpret the result in your own words.

2. Attached is a CSV file that includes data on 800 individuals from different states. The variables included are: **(40 points)**

   - personid: individual identifier
   - educ: years of education completed by each individual
   - white: dummy variable equal to 1 if the person is white and 0 otherwise
   - age
   - income

- cigs: number of cigarettes smoked per day
- cigpric: state cigarette price – cents per pack
- restaurn: dummy variable = 1 if person resides in a state that has restrictions on smoking in restaurants

a. Present summary statistics of all the relevant variables in the dataset. Think about how to present your output to the reader (who is me), who hasn't seen the dataset before. Make sure to write one paragraph explaining and describing the characteristics of your sample.

b. What fraction of your sample are smokers? Explain how you got to the answer.

c. Now we want to try to have a more intuitive summary statistics that tells the reader more about the differences across different groups. Re-do the summary statistics for smokers and non-smokers separately. Again, think about how to present your results. Write one-to-two paragraphs explaining the difference in characteristics between smokers and non-smokers.

d. Test the hypothesis that the level of education is similar for smokers and non-smokers. Please state the null and alternative hypotheses and compute the t-statistic by hand using the summary statistics that you generated in the table above.

P.S. In number d, please state your null and alternative hypotheses. You can conduct your test using any software and just eyeball the p-values. But in your write up, please write out the null and the alternative hypotheses and then your conclusion about whether you accept or do not accept the null. Also please provide a complete interpret the result in your own words.

e. Using the data you have, show that non-white individuals are more likely to be smokers.

f. Do you think there is a relationship between income and the number of cigarettes that a person smokes? Show visually

d. Do you think there is a relationship between the price of cigarettes and the number of cigarettes that a person smokes? Show visually

3. Suppose that Amazon is considering introducing a rewards program. The idea of the program is that the value of each purchase on Amazon.com is converted into points. For each 1000 points, the customer can convert the 1000 point into 10 dollars to spend on any good or service from Amazon or use it to pay part of Prime membership fees. The management believes that the rewards program will have two main benefits: (1) increase sales through amazon.com since customers will be interested in collecting points; and (2) maintain and potentially expand the number of Prime members. Enrollment in the rewards program is not

automatic. Customers have to enroll themselves in the program through their personal accounts. Suppose that Amazon has been implementing a pilot version of this program for 3 months now.

Amazon turns to its economists' team to study whether the introduction of the rewards program will help them achieve their objectives. You are the economist responsible for making this assessment.

In the pilot version of the program, some customers who signed up. Amazon (like all tech companies) has infinite amount of data about each and every customer. They know the customer's demographic information, where they live, what they buy, …etc. For the rest of the problem assume that in Amazon, you have access to any data that you need.

(10 points each)

a. What are potential comparisons that you can make in order to assess whether introducing the rewards program helps increase sales?

b. Please explain why these comparisons do not tell us the real causal effect of the impact of the rewards program on Amazon sales.

   P.S. I know there is selection bias. Please explain in detail what are the sources of the bias and what is the potential direction of the bias. Is it upward bias or downward bias? Explain the direction for each of the sources separately and then in the end try to make a conclusion about the general direction of the bias.

c. Suppose that the Chief Economist gave you the green light to do whatever it takes to study the causal impact of the rewards program. What would you do to get the best answer? Please explain your approach in detail and explain why this approach gives us the causal impact.