Fall 2021
SS 340
Loujaina Abdelwahed

Submit your work at the beginning of class on Monday (Nov 1). You may work with other students, but the write-ups must be unique. Please save your answers in .docx format.

This problem set is based heavily on the paper Almond, Chay, and Lee (2005). Although they use a different methodology to estimate the causal effect, which we have not covered, we are interested in the same question about the relationship between maternal smoking during pregnancy and infant birthweight. We use the same dataset that the authors used.

This problem set is based on problem sets from Ken Chay, John DiNardo, Jordan Matsudaira, and Ben Ost.

The goal of this assignment is to examine the research question: what is the effect of maternal smoking during pregnancy on infant birthweight. The data for the problem set is an extract of all births from the 1993 National Natality Detail Files for Pennsylvania. Each observation represents an infant-mother match. The data in csv format, pennbirthweight0.csv.

There should be 161,493 observations in the data, with 55 variables. You also have the codebook of the dataset, which explains the variable names and help you figure out what each variable is measuring.

The data here are "real" and quite imperfect, which will help to simulate the unpleasant experience of real-world data work. Part of the goal of the problem set is to gain experience dealing with "dirty" data. I have NOT done any effort to pre-process the data for you. You should refer to the codebook for the data to help you.

I also have not attempted to solve the problem set myself. I will deal with your write-ups like I am reviewing a paper for a journal. I have not seen the data before. The only thing I see is your analysis.

Please note that this problem set will take some time, so you should start early!

Most of the work will be writing code. For some questions I want you to answer questions or create a table and I have bolded these portions. For example, for question 1, all you have to do is write code. You are not expected to produce any output for your solutions for question 1, but for question 2 part b, you are expected to present output and/or respond to the question in your typed write-up. For portions that aren't bolded, I still expect that you examine the question and may be write a few sentences about how you thought of the question.

Please make your answers reasonably pretty (e.g. don't just take a snapshot of your stata or python output). Figures and tables should have headings and variables should be labeled reasonably. In other words, in your write up, I do not want to see the variable codes (e.g. dbrwt). LABEL the variables.

Fall 2021
SS 340
Loujaina Abdelwahed

1. Process the data using the codebook for guidance.
    a. Fix missing values. In the data set, several variables take on a value of, say, 999 if missing. These need to be fixed because they will throw off all your calculations. Instead, variables with missing values should be set to "." (refer to the codebook for missing value codes)
    b. Be careful of variable definitions e.g. the lung cancer variable (and others) is coded as 1 or 2, not 0 or 1. You might want to code them as binary 0/1.
    c. Produce an analysis dataset that drops observations with missing values. (If this were a real research project you would not indiscriminately drop observations like this. You would consider why observations were missing, how this might bias your estimates, and whether missing data imputations might be worthwhile.) In this case, you can simply drop observations with missing values.
    d. In creating your analysis dataset, you should only keep the variables you plan to use in your analysis. It is good practice to label all your variables as well.
    e. Produce a summary table describing the sample means of the variables in your analysis dataset.

2. Now you will estimate the impact of smoking during pregnancy on infant birth weight.
    a. Compute the mean difference in birthweight by smoking status.
    b. Under what circumstances can one identify the average treatment effect of maternal smoking by comparing the unadjusted difference in mean birth weight of infants of smoking and non-smoking mothers?
    c. **Create a table that convinces me either that the assumption from part (b) is valid or not valid in the current context.** HINT: A useful table in any paper is one that describes the average values of the covariates for the observations and then presents these means for the subsets of people who do and do not receive the treatment. The table should also statistically test whether the mean of each covariate is different between treatment and control. For your table, you could have 4 columns and roughly 10 rows. The rows are based on the covariates you think are most important in thinking about the assumption from part (b). The first column replicates part of the table you made in 1.e (means of the covariates.) The second and third columns show the means for the covariates split by treatment vs control. The fourth column performs a t-test for each covariate. Just report p-values for your test.
    d. Based on you results from 2.c, make a guess at the direction of the bias. Recall, that you need to consider both the covariance between the treatment and the covariates and the likely direct impact of the covariates on the outcome.
    e. Using a simple linear regression, estimate the impact of smoking on birth weight. Use robust standard errors.
    f. Start adding covariates to your simple regression. You should choose your covariates very carefully and only include what you think are "good controls". **What is your estimate of the impact of smoking on birthweight? Explain in terms of the omitted variable bias formula why the coefficient from the regression increased or decreased relative to the means comparison from 2.a.**

g. Consider how the estimate is different depending on whether you include a control for whether the infant is a girl or a boy. **Explain why this control doesn't change the coefficient estimate for the impact of smoking on birth weight yet is highly significant. Why is it still beneficial to include "useless" controls?** Add in other useless controls.

h. Add in a control that you consider to be a very "bad control". **Explain why the control is bad, and provide an interpretation of how the coefficient on smoking changed.** For future parts of the problem set, you should include the "useless" controls and "good" controls, but you should exclude the bad controls to avoid over controlling. This is true for research more generally.

3. You have analyzed the impact of smoking on birthweight using OLS. **State the assumption behind OLS and comment on whether you think the assumption is plausible in this context. Based on the analyses you have performed, what would you conclude regarding the causal impact of smoking on birthweight?**