SDS 384 Spring 2020: Homework #1

Due: Tuesday, February 18 by 9:00am

Please submit your homework solutions as a single .pdf file via the canvas website. If you have handwritten solutions to some problems, please scan them and turn them in via the canvas website. If you wish to turn in a combination of typed and handwritten solutions **please merge them as a single .pdf file**.

Remember that you are encouraged to work in groups for the homework assignments, but that **each student must turn in their own solution** that cannot be a simple copy of the others from your group. If you work in a group, please list the names of your group members on the assignment.

(1) The canvas site provides an excerpts from the published HEI Research Report 148, *Impact of Improved Air Quality During the 1996 Summer Olympic Games in Atlanta on Multiple Cardiovascular and Respiratory Outcomes* by Peel et al.. The first four pages are excerpts from an overall summary and critique of the report, which provide general background. The next five pages are excerpts from the more detailed research report. In both cases, the text appearing in red boxes should be sufficient to answer the following questions.

   (a) For both the analysis of ozone concentrations and the analysis of ED visits, what are the potential outcomes defining the effects of interest?

   (a) For both the analysis of ozone concentrations and the analysis of ED visits, provide a possible violation of the "No Multiple Versions of Treatment" (or "consistency") part of SUTVA.

   (a) For both the analysis of ozone concentrations and the analysis of ED visits, describe the assignment mechanism and provide one reason why it may not be *unconfounded*.

(2) Describe an example of an assignment mechanism that is ignorable but not unconfounded and support your argument with statement(s) about the assignment mechanism.

(3) For this exercise, you will follow Chapter 11 of the Imbens and Rubin textbook, but use a different data set to implement some of the techniques we have learned about in class. While the textbook describes analysis of a social program called the Saturation Work Initiative Model (SWIM) program, in this exercise you will analyze the `lalonde` data available in the `R` package `MatchIt` with the command `data(lalonde)`. Use the `help(lalonde)` command to get a basic description of the data. Use the variable called `treat` to denote the randomized treatment assignment, use `re78` as the outcome of interest, and use `re74` as the pre-training earnings.

   (a) Produce a table analogous to Table 11.2 of the textbook. Use the same test statistics

as in Section 11.3, and conduct the tests for the entire data set (corresponding to the All column in Table 11.2) and stratified by the `nodegree` variable (corresponding to the two rightmost columns of Table 11.2).

(b) Unlike in Section 11.3 of the textbook, provide a Fisher interval for the analysis of the entire data set (i.e., not stratified by `nodegree`).

(c) Produce a table analogous to Table 11.3 of the textbook, but with estimates for the same three categories as in part (a) (All and both levels of `nodegree`).

(d) Fit a regression model with interactions as in Section 11.5 of the textbook, using the following variables as covariates: `age, educ, black, hispan, married, nodegree`. Report the population average treatment effect from this model and compare it with the effect estimates from a regression model that adjusts only for the treatment indicator.

(e) The second paragraph of Section 11.6 describes an analysis assuming a normal distribution for the two potential outcomes with the correlation between $Y_i(0)$ and $Y_i(1)$ fixed to be 1.0 and unknown mean and variance parameters. Perform the described model-based inference for the model with no covariates using the priors specified in the text and report your estimate of the treatment effect.