SDS 384 Spring 2020: Homework #2

Due: Tuesday, March 24 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.

This assignment centers around a data set very similar to the one used in the Papadogeorgou et al. DAPSm paper. The data contain information on power plants operating in the United States in 2002 and 2014, and are available on the Canvas site in the file `annualEGUs.csv`. Specifically, the units in the data are Electricity Generating Units (EGUs) in 2002 and 2014, some of which were treated with a particular technology to reduce their emissions of NO_x, an important precursor to harmful air pollution. The technology is a Selective Catalytic Reduction or Selective Non Catalytic Reduction System, (SnCR). The outcome of interest is the level of NO_x emissions. Several other characteristics are measured on each power plant. Table 1 lists the variables that you will use for this analysis (you can ignore any other variables you see in the data). For all analyses of these data, log transform the `Outcome` variable.

Table 1: Description of relevant variables in the `annualEGUs.csv` data.

Variable name	Description
`Tx`	Whether the EGU has an SnCR installed in that year
`Outcome`	Annual emissions of NO_x in tons
`totOpTime`	Number of hours operated during the year
`HeatInput`	Measure of the amount of fuel burned
`pctCapacity`	Average percent of total operating capacity actually operated
`Phase2`	Indicator of participation on Phase II of the Acid Rain Program
`avgNOxControls`	Average number of other NO_x emissions controls (besides SnCR)
`coal_no_scrubber`	Indicator of whether the EGU burns coal as primary fuel and does not have an SO_2 scrubber installed
`coal_with_scrubber`	Indicator of whether the EGU burns coal as primary fuel and has an SO_2 scrubber installed
`EPA.Region`	Which of 9 EPA defined regions in which the EGU is located

(1) Separately for 2002 and 2014, conduct an unadjusted "crude" analysis comparing the average NO_x levels for treated and untreated units. Evaluate whether the observed covariates are balanced in this unadjusted analysis.

(2) In this exercise you will use a variety of propensity score methods to estimate the causal effect of having an SnCR in a given year on NO_x emissions in that year, under the assumption that the covariates listed in Table 1 are sufficient to adjust for confounding (i.e., that having an SnCR installed is conditionally unconfounded with respect to NO_x emissions). For all parts of this exercise:

- Use logistic regression with all of the variables in Table 1 (besides `Tx` and `Outcome`) included as covariates to estimate the propensity score.
- Be sure to check covariate balance for each analysis
- Conduct each analysis separately for 2002 and 2014, and comment (in ~ 3 sentences) on the differences between the analyses in the two years
- I strongly suggest you read up on the following R packages to conduct these analyses: `MatchIt`, `survey`, `ipw`, `twang`

(a) When you arrive at a propensity score model, plot the histograms of the estimated propensity scores in treated and untreated units.

(b) Conduct a 1-1 nearest neighbor propensity score matching procedure without replacement.

(c) Conduct a 1-1 nearest neighbor propensity score matching procedure without replacement and a caliper set to 0.1 standard deviations of the estimated propensity score distribution.

(d) Conduct an analysis that subclassifies units based on the estimated propensity score

(e) Conduct an IPW analysis using weights $\frac{W_i}{\hat{e}(X_i)} + \frac{1-W_i}{1-\hat{e}(X_i)}$ and be sure to include a visual summary (e.g., histogram) of the estimated weights.

(f) Conduct an IPW analysis using stabilized weights and be sure to include a visual summary (e.g., histogram) of the estimated weights.

(3) Describe in ~ 5 sentences why the answers you obtained with the different propensity score methods in Exercise (1) were different from one another.

(4) Repeat Exercise (1e), but use a more advanced prediction model (your choice) to estimate the propensity score. Describe (~ 3 sentences) any differences.