Problem Set 5

Prof. Conlon

Due: 4/4/20

Packages to Install

The packages used this week are

- ggplot2
- data.table (data tables are computationally efficient and IMHO easier to work with)
- rdd (package for regression discontinuity designs)
- estimatr (tidy version of lm)
- knitr (make pretty tables using kable command)
- extraDistr (package with extra distributions)

```
ERROR: DATA DIRECTORY NOT CORRECT
data_dir variable set to:  /Users/christopherconlon/Documents/applied_metrics/data/
```

Problem 1 (Coding Exercise)

Using the Lalonde dataset and the **cobalt** package finish the exercise from the slides.

That is:

Consider three possible matching techniques

1. Caliper matching on a single variable (pick the best one)
2. 4 nearest neighbor matching.
3. Propensity Score matching using a logit
4. Propensity Score matching using a kernel

For each matching approach:

a. Create a balance table. For each pretreatment covariate, include comparisons for treated and untreated units in terms of the mean and standard deviation. Report a test, for each covariate, of the hypothesis that the difference in means between treatment conditions is zero.

b. For each covariate, plot its distribution under treatment and control

c. Estimate the ATT and/or ATE of participating in the job training program

d. Can you estimate both ATE or ATT? Why or why not?

Problem 2 (Coding Exercise)

The dataset for this exercise comes from a paper by Benjamin Olken entitled "Monitoring Corruption: Evidence from a Field Experiment in Indoneisa". The paper evaluates an attempt to reduce corruption in road building in Indonesia. The treatment we focus on was "accountability meetings". These meetings were held at a village level, and project officials were probed to account for how they spent project funds. Before construction began, residents in the treated villages were encouraged to attend these meetings. The dataset is called "olken.csv".

The outcome we care about is **pct.missing**, the difference between what officials claim they spent on road construction and an independent measure of expenditures. Treatment is given by **treat.invite** such that:

$$\text{treat.invite} = \begin{cases} 1 & \text{if village received intervention} \\ 0 & \text{if village was control} \end{cases}$$

We have the following four pre-treatment covariates:

- head.edu : the education of the village head
- mosques : mosques per 1000 residents
- pct.poor : the percentage of households below the poverty line
- total.budget : the budget for each project

We now have the following questions:

a. Create a balance table. For each pretreatment covariate, include comparisons for treated and untreated units in terms of the mean and standard deviation. Report a test, for each covariate, of the hypothesis that the difference in means between treatment conditions is zero.

b. For each covariate, plot its distribution under treatment and control (either side-by-side using facet_grid or overlap).

c. Given your answers to part a and b, do the villagers seem similar in their pre-treatment covariates?

d. Regress the treatment on the pre-treatment covariates. What do you conclude?

e. Using the difference-in-means estimator, estimate the ATE and its standard error.

f. Using a simple regression of outcomes on treatment, estimate the ATE and its standard error. Compare your answer in (f) to (e).

g. Using the same regression from part (f), include pre-treatment covariates in your regression equation (additively and linearly). Report estimates of treatment effects and its standard error. Do you expect (g) to differ from (f) and (e)? Explain your answer.

Problem 3 (Coding Exercise)

We will be using a dataset that was simulated from real data. Oftentimes due to privacy concerns researchers will provide simulated data from the distribution of real data. The dataset you will be using are from a tutoring program focused on math for 7th graders. The dataset is called "tests_Rd.csv". The tutoring is the treatment variable, **treat**. Tutoring was given to students based on a pretest score, **pretest** thus the pretest score is the forcing variable. Students that received less than 215 were given a tutor. Our outcome of interest is the test score after tutoring, **posttest**.

We also have a series of control variables:

- age : age of student as of September 2010
- gender : 1 if student's gender is male
- frlunch : 1 if student is eligible a free lunch
- esol : 1 if student has english as a second language
- white : 1 if student's race/ethnicity is white

- asian : 1 if student's race/ethnicity is asian
- black : 1 if student's race/ethnicity is black
- hispanic: 1 if student's race/ethnicity is hispanic

We ask you to answer the following questions:

a. We want you to first plot the graph that justifies a sharp RD design. Plot the treatment as a function of the forcing variable. What do you see?

b. We now want you to plot the graph that justifies our forcing variable. Plot the outcome as a function of the forcing variable. What do you see?

c. Estimate the local average treatment effect (LATE) at the threshold using a linear model with common slopes for treated and control units (with no control variables). What are the additional assumptions required for this estimation strategy? Provide a plot of the post test scores (y-axis) and forcing variable (x-axis) in which you show the fitted curves and the underlying scatterplot of the data. Interpret your resulting estimate.

d. Re-do c., but use the control variables that are provided in the dataset. Interpret any differences you see.

e. Use the rdd package in R to estimate the LATE at the threshold using a local linear regression with a triangular kernel. Note that the function RDestimate automatically uses the Imbens-Kalyanamaran optimal bandwidth calculation. Report your estimate for the LATE and an estimate of uncertainty.

f. How do the estimates of the LATE at the threshold differ based on your results from parts (b) to (e)? In other words, how robust are the results to different specifications of the regression? What other types of robustness checks might be appropriate?

We are now going to do a series of robustness checks:

h. Plot the age variable as a function of the forcing variable. What should this graph look like for our RDD to be a valid design? What do you see? How does this relate to the covariate balance exercise we did in Problem 1?

i. One type of placebo test is to pick arbitrary cutoffs of your forcing variable and estimate LATE's for those cutoffs. Pick 10 cutoffs and report the average LATE across those cutoffs. Defining $\hat{\tau}$ as your average LATEs, what should the null hypothesis on the population counterpart of this estimator be for our design to be valid. Feel free to use which specification you want for estimating LATE, but please specify it.

k. An issue with RD designs is manipulation, or sorting around the cutoff point. To assess this, plot a histogram of the forcing variable, drawing a line at the cutoff point. What would sorting around the cutoff point look like? What do you see?