Problem Set 2

Professor Nathan Miller (Georgetown University)

Regression Discontinuity Design

The application features data on a tutoring program focused on math for seventh graders. The data are simulated from the distribution of real data in order to preserve the anonymity of the students. Tutoring was given to students based on a pre-test score, which therefore can be conceptually as the running variable. Students that received a pre-test score less than 215 were given a tutor. All students subsequently were scored on an exam. The file testsRd.csv contains observations on the following variables:

- *treat*: an indicator that equals 1 if the student received tutoring
- *pretest*: the score on the pre-test
- *posttest*: the exam score.
- *age*: the age of the student as of September 2010
- *gender*: equals 1 for males
- $frlunch$: an indicator that equals 1 if the student is eligible for free lunch
- *esol*: an indicator that equals 1 if the student has English as a second language
- *white*: an indicator that equals 1 if the student's race/ethnicity is white
- *asian*: an indicator that equals 1 if the student's race/ethnicity is Asian
- *black*: an indicator that equals 1 if the student's race/ethnicity is black
- *hispanic*: an indicator that equals 1 if the student's race/ethnicity is Hispanic

Familiarize yourself with the data and then answer the following questions:

1. Plot treatment as a function of the running variable. Does the graph justify sharp RDD?
2. Plot the exam score as a function of the pre-test score. What do you observe, and does this justify the use of the pre-test as a running variable?
3. Estimate the treatment effect at the threshold using a linear model with common slopes for treated and control units. Implement with an OLS regression of the exam score on the indicator for tutoring and the pre-test score. Under what assumptions does this estimation strategy obtain a consistent estimate of the causal effect? Provide a plot of the exam scores (y-axis) and the pre-test scores (x-axis) in which you show the regression fits and the underlying scatterplot of the data. Interpret your estimate.

4. Repeat the exercise from question 3, but this time include both the pre-test variable and the square of the pre-test variable in the regression. Does the estimate change much? And is that consistent with your expectation, given your answer to question 2?

5. Again repeat the exercise from question 3, but this time include the control variables that are provided in the dataset. Interpret any differences you see.

6. Use the rdd package in R to estimate the treatment effect 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 treatment effect and an estimate of uncertainty.

7. How do the estimates of the treatment effect differ across your results for questions 3-6? In other words, how robust are the results to different specifications of the regression?

8. Plot the age variable as a function of the running variable. What should this graph look like for RDD to be a valid research design? What do you see?

9. One issue with RDD is manipulation, i.e., sorting around the cutoff threshold in the running variable. Plot a histogram of the running variable, drawing a vertical line at the cutoff. What would sorting around this cutoff point look like? What do you see? Use the rdplotdensity function in R to evaluate the statistical significance of any changes.

Differences-in-Differences

The application features data from the Stevenson and Wolfers (2006) study of divorce law reform and suicide.[1] Especially in the 1970s, many states revised divorce laws to allow for unilateral divorce (i.e., so that a married individual can obtain a divorce without consent from their spouse). Stevenson and Wolfers examine whether this led to a decline in suicide and domestic violence. The file sw06.csv contains observations on the following variables:

- $stfips$: the FIPS code of the state
- $year$: the year
- nfd: the year in which the state revised its divorce laws
- $post$: and indicator that equals 1 if the state has revised its divorce laws
- $asmrs$: the number of suicides per 1 million people
- $pcinc$: per capita income
- $cases$: a measure of poverty based on the AFDC program
- $copop$: the population

[1] Betsey Stevenson and Justin Wolfers (2006). "Bargaining in the Shadow of the Law: Divorce Laws and Family Distress." *Quarterly Journal of Economics* 121 (1), pp. 267-288.

- *posttrend*: the number of years since the state revised its laws
- *pretrend*: the number of years before the state revises its laws

Familiarize yourself with the data and then answer the following questions:

10. Use OLS to regress suicides on the treatment indicator and the control variables (income, poverty, and population). Interpret your results. Under what conditions does this provide an unbiased estimate of the causal effect?

11. Estimate the treatment effect with differences-in-differences by regressing suicides on the treatment indicator, state fixed effects, and year fixed effects. In a second regression, also include the control variables. Interpret your results.

12. Re-estimate the model (without controls) excluding state-year observations for which the divorce law revisions have been in place longer than 2 years. This can be implemented by selecting observations with $posttrend \leq 2$. Repeat the analysis, cutting off the sample after 5, 10, and 15 years of treatment. What do you observe, and what does this imply for the validity of the estimates obtained for the previous question?

13. Now estimate an event study model. Regress suicides on state fixed effects, year fixed effects, and fixed effects for the different values of *posttrend*. The last set of fixed effects is the object of interest. Plot the values of these fixed effects, along with a confidence interval, on a graph where *posttrend* is on the horizontal axis. Interpret the graph.

14. To check the parallel trends assumption, reestimate the model from the previous question, this time also including fixed effects for the different value of *pretrend*. Interpret the results. Do they support the validity of the event study approach to estimation?