Harris School of Public Policy
PPHA 346 Program Evaluation
Winter 2018

Assignment 4

Due date: March 7, 2018, by 11:59pm via Canvas.

Format: For this assignment, provide a write-up in pdf or html format where you answer the questions below. Selectively cut and paste output and be concise in your write-up; excess wordiness will be penalized. Also submit your code script that includes commands for your entire analysis.

Data: You will find data birthweight.dta and MTO_DT_subset.dta on Canvas.

Non-Stata or -R Languages:
For this assignment, you are recommended to use either STATA or R because answer keys and script will be in STATA and R. When you get wrong answers, suggested solutions probably won't be helpful if you use other software. Please keep in mind:

- Your grade will be based entirely on whether you got the right numeric answer and whether your written explanations/interpretations are correct. STATA and R users who get wrong numeric answers due to a minor mistake in their code might still get partial credit if their reasoning and interpretation are right, but we might not be able to identify mistakes in code for other languages.
- Regardless of what language you use, you are always required to show your code after written explanations and before outputs, as indicated above.
- Regardless of what language you use, PLEASE submit both of your write-up and code script on Canvas.

Late Submissions:
Please turn in your assignment on time. Per syllabus policy, two percentage points of your grade (10 percentage points max) will be discounted due to late submission.

Academic Honesty

You can discuss this assignment with your classmates, but you should not share your code or your write-up. Academic dishonesty will not be tolerated. If you commit plagiarism, your homework will be penalized and it might result in further severe consequences.

Part I: Regression Discontinuity Design

Background: A key policy question of medical expenditures on treating at-risk newborns is whether the marginal returns are sufficient to justify the cost. Accurate measurement of these returns, however, has been difficult due to challenges measuring returns to medical expenditures as well as downward bias in estimates because households with less healthy newborns often have higher medical expenditures. In this assignment, we are going to use the method of Regression Discontinuity Design to analyze the benefit of medical care for at-risk newborns by comparing treatment outcomes on each side of the risk classification. Download the dataset birthweight.dta on Canvas.

Question 1. Assuming newborns weighing below 1500 grams receive the medical treatment while those weighing above 1500 don't, taking birth weight as our running variable, we focus on a commonly-used newborn risk classification: the "very low birth weight" (VLBW) classification at 1500 grams (under 3 pounds, 5 ounces). One assumption of RDD is that the probability of treatment receipt must be discontinuous at cutoff. In order to test this assumption, plot one-year mortality rates against birth weight. To do this, construct bins of one ounce around VLBW, compute the mean mortality rate on each bin and save them to a new variable. How many bins did you construct? Describe the relationship between birth weight and one-year mortality rate.

Question 2. Plot one-year mortality rates against birth weight using bin-width of three ounces around the same threshold. How many bins does the plot have? Discuss your finding with Question 1.

Question 3. A requirement of a valid RDD is that agents are not able to manipulate the running variable. Discuss if birth weight exhibits evidence of being manipulable.

Question 4. Another assumption of RDD is the "continuity restriction." Select two pre-treatment covariates and check that they are smooth around the VLBW threshold by plotting them near the threshold. Discuss your findings. How would discontinuities on other covariates affect your RDD estimate?

Question 5. Assuming a **sharp regression discontinuity**, use *OLS* to estimate the effect of VLBW on the one-year mortality rate. To do so, use the following model:

$$Y_i = \alpha_0 + \alpha_1 Treatment + \alpha_2 BirthWeight + \varepsilon$$

Report and interpret the coefficients on Treatment and BirthWeight.

Question 6. Now we want to estimate the discontinuity in treatment effect at the VLBW threshold. For one-year mortality, estimate the size of the discontinuity using a 50 grams caliper above and below the threshold. Use the following model:

$$Y_i = \alpha_0 + \alpha_1 Treatment + \alpha_2 (BirthWeight - 1500) + \alpha_3 Treatment\,(BirthWeight - 1500) + \alpha_4\,(1 - Treatment)\,(BirthWeight - 1500) + \varepsilon$$

Interpret the coefficients.

Question 7. Repeat question 5 using caliper sizes 30 and 100 grams. Report your regression tables and discuss your findings. Among calipers 30, 50, and 100, which one do you think is the best for the RDD analysis? Do you always want the caliper as small as possible?

Part II: Instrumental Variable

Background: For this part, we will again use the Moving to Opportunity data. We have simplified the dataset so that our subset does not reflect the full design of the MTO study. We want to test the impact on having diabetes of making a move through the Moving to Opportunity program. Due to the endogeneity of the move decision (the treatment, in this case), the OLS estimate of the effect of moving on diabetes does not provide a consistent estimate. In the following analysis, take the variable **diabetes**, an indicator of diabetes as identified by a blood test, as the outcome variable. A new variable **treated** indicates that an actual move occurred through the MTO program (that is, a member of the experimental group made a move). The dummy variable **assignment** is an instrument that indicates that an individual had a chance to receive the treatment.

(Note: some variables and some rows, including group_section8 and treated observations in the group_section8 group have been removed from the original data set. Please download the new dataset MTO_DT_subset.dta on Canvas instead of the MTO_data.dta.)

Question 1. The variable Treated is an indicator which equals one for individuals who made a move through the MTO program, and zero for those who did not participate in that program. Is assignment of treatment randomized in this program? Identify never-takers, defiers, compliers, always-takers.

Question 2. Start with the naive regression. Run an OLS regression of **diabetes** on **treated**. Interpret the coefficients. Can you interpret your OLS estimate as a causal effect? Explain your answer.

Question 3. State intuitively the required *instrument condition* for the variable **assignment** to be a valid instrument for the variable **treated.** In this case, is the condition met? Analyze the relationship between **assignment** and **treated**. Where relevant, show and interpret tabulations of the data.

Question 4. State intuitively the required *exclusion restriction* for the variable **assignment** to be a valid instrument for the variable **treated**. What are the potential violations of this requirement? As a test, run a regression of **diabetes** on the instrument variable **assignment**. Discuss what this reduced form regression tells you.

Question 5. Manually compute the Wald estimator for LATE of **treated** on **diabetes**. Discuss the difference between the Wald estimator you calculated and the OLS estimate.

Question 6. Estimate the LATE of **treated** on **diabetes** using a manual implementation of 2SLS with two covariates of your choice. (Explain why you chose those two covariates). In the first stage, explain if the instrument is strong. Describe what the estimate of the treatment effect indicates.

Question 7. Now, use ivregress() to estimate your 2SLS. Evaluate your estimates. Are they different from the results you got in Question 7? Explain why they are similar/different. (Note: ivregress() should be available in R/stata/sql/Python.)