14.310x Data Analysis for Social Scientists
Practical Issues in Running Regressions and Omitted Variable Bias

Welcome to your ninth homework assignment! You will have one week to work through the assignment. We have provided this PDF copy of the assignment so that you can print and work through the assignment offline.

Good luck!

Please find a glossary of R terms that will be useful for this week's homework here and additional resources here.

Difference in Differences (DiD): Questions 1 – 10

Difference in differences is a statistical tool broadly used by empirical economists. In this problem, we are going to replicate the results of David Card and Alan Krueger's "*Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*." The accompanying data (fastfood.csv) was used to study the effects of an increase in the minimum wage on unemployment. Here is the abstract of the study:

> On April 1, 1992, New Jersey's minimum wage rose from $4.25 to $5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above $5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment.

The data set contains the following variables:

Note: NJ refers to New Jersey and PA refers to Pennsylvania

- **chain**: 1 = Burger King, 2 = KFC, 3 = Roy Rogers, 4 = Wendy's
- **state**: 1 if NJ; 0 if PA
- **southj**: 1 if in southern NJ
- **centralj**: 1 if in central NJ
- **northj**: 1 if in northern NJ
- **shore**: 1 if on NJ shore
- **pa1**: 1 if in PA, northeast suburbs of Philadelphia
- **pa2**: 1 if in PA, all other areas besides the northeast suburbs of Philadelphia
- **empft**: number of full-time employees before the change in the minimum wage
- **emppt**: number of part-time employees before the change in the minimum wage

- **wage_st**: starting wage in the local (per hour) before the change in the minimum wage
- **empft2**: number of full-time employees after the change in the minimum wage
- **emppt2**: number of part-time employees after the change in the minimum wage
- **wage_st2**: starting wage in the local (per hour) after the change in the minimum wage

Load the data into R and run a linear model in which you compare whether there are differences between fast-food restaurants located in NJ and Pennsylvania prior to the change in the minimum wage in terms of the number of full-time employees and the starting wage.

Question 1
What is the average difference between the fast-food restaurants located in NJ and Pennsylvania in terms of the number of full-time employees (before the change in minimum wage)?

Note: This is not an absolute difference. Please include signs in your answer and round to the nearest three decimal places.

Question 2
Is the difference statistically significant at the 1% level?
- Yes
- No

Question 3
Now, let's look at the starting wage. According to the model that you've run, what is the average wage in Pennsylvania prior to the change?

Please round your answer to the nearest three decimal places.

Question 4
What is the average starting wage in New Jersey?

Please round your answer to the nearest three decimal places.

Question 5
Can we reject the null hypothesis that the average starting wage is the same in NJ and Pennsylvania prior to the change in the minimum wage?
- Yes
- No

Now assume that someone pointed out that the northeast suburbs of Philadelphia are very different from the rest of Pennsylvania. This person claims that the model that should be used to estimate the differences between fast food restaurants located in NJ and PA prior to the change is as follows:

$$full\ time\ employment = \beta_0 + \beta_1 state + \beta_2 pa1 + \beta_3\, pa2 + \varepsilon$$

Question 6
According to this model above, what would the average difference in full time employment between the restaurants located in the northeast suburbs of Philadelphia and the rest of Pennsylvania?

- $(\beta_2 + \beta_3) - \beta_1$
- $\beta_2 - \beta_3$
- $\beta_3 - \beta_2$
- β_1
- It is not possible to tell from this model

Now, let's run the difference in differences model to see whether the change created a difference in the employment. According to what we saw in the lecture, the model to estimate should be the following:

$$empft_{it} = \beta_0 + \beta_1 state_i + \beta_2 post_t + \beta_3 state_i\ x\ post_t \quad \text{(equation 1)}$$

where $post_t$ is a dummy variable that takes the value of 1 after the change takes place.

Question 7
What is the parameter that captures the difference between New Jersey and Pennsylvania prior to the implementation of the change?

- $\beta_3 + \beta_1 - \beta_2$
- β_2
- β_3
- β_1

Even though we aim to estimate equation (1), the data we have is "wide", meaning that we have one observation per fast food restaurant, and the different yearly observations are in different variables. Model 1 would require the data to be in the "long" format, with the different years stacked. We could change the structure of the data to solve this problem.

Alternatively, someone has suggested that instead we could run the following model:

$$empft_{i2} - empft_{i1} = \alpha_0 + \alpha_1 state_i + v_i \quad \text{(equation 2)}$$

And that our estimate for α_1 in equation (2) would be equivalent to β_3 in equation (1).

Question 8
Is this statement correct? In other words, is it true that α_1 in equation (2) is equivalent to β_3 in equation (1)?

- Yes
- No

Question 9
Now estimate the model in equation (2) in R. What value do you obtain for the DiD estimate?

Do not round. Please input the answer exactly as it appears in the summary output in R.

Question 10
Assuming that we can interpret the estimate for α_1 as causal and that the minimum wage for these fast-food restaurants is binding, can you conclude that the NJ increase in the minimum wage had a negative effect on full-time employment?

- Yes
- No

Regression Discontinuity: Questions 11 – 18

In this part of the homework we are going to replicate the results of David S. Lee's paper and we have kindly been provided his data to the Mostly Harmless Econometrics Data Archive.

Lee (2008) studies the effect of party incumbency on reelection probabilities. In general, Lee is interested in whether a Democratic candidate for a seat in the U.S. House of Representatives has an advantage if his party won the seat last time. Here is the abstract of the working paper version of "The Electoral Advantage to Incumbency and Voters' Valuation of Politicians Experience: A Regression Discontinuity Analysis of Elections to the U.S. Houses."

> Using data on elections to the United States House of Representatives (1946-1998), this paper exploits a quasi-experiment generated by the electoral system in order to determine if political incumbency provides an electoral advantage – an implicit first-order prediction of principal-agent theories of politicians and voter behavior. Candidates who just barely won an election (barely became the incumbent) are likely to be ex ante comparable in all other ways to candidates who barely lost, and so their differential electoral outcomes in the next election should represent a true incumbency advantage. The regression discontinuity analysis provides striking evidence that incumbency has a significant *causal* effect of raising the probability of subsequent electoral success – by about 0.4 to 0.45. Simulations – using estimates from a structural model of individual voting behavior – imply that about two-thirds of the apparent electoral success of incumbents can be attributed to voters' valuation of politicians' experience. The quasi-experimental analysis also suggest that heuristic "fixed effects" and instrumental variable" modeling approaches would have led to misleading inferences in this context.

We have provided you with the data set individ_final.csv. It contains the following variables:

- **yearel**: election year
- **myoutcomenext**: a dummy variable indicating whether the candidate of the incumbent party was elected
- **difshare**: a normalized running variable: *proportion of votes of the party in the previous election* – 0.5. If *difshare* > 0 then the candidate runs for the same party as the incumbent.

Load this data into R and install the package `rdd` to answer the following questions.

Question 11

Based on the information provided, create a variable for whether the party of the candidate is the same party as the incumbent. What is the proportion of these cases in your data set?

IPlease round your answer to the second decimal place, i.e. if your answer is 0.8982, round to 0.90 and if it is 0.8922, round to 0.89.

One of the main assumptiosn in RD designs is that there are no jumps in the density of the running variable around the cutoff. The package in R `rdd` has a command `DCdensity`. Run the command in R using *difshare* as the running variable. Refer to the documentation for the command if you have any questions.

Question 12
What is the difference in the log estimate in heights at the cutpoint? *Note: this should be a negative number.*

Please round your answer to the fourth decimal place, i.e. if your answer is 1.03456, please round to 1.0346, and it it is 1.03451, round to 1.0345.

Question 13
Can you reject the null hypothesis that this difference is equal to zero?

- Yes
- No

Now, keep only the observations within 50 percentage points of the cutoff (the absolute value of **difshare** is less thab or equal to 0.5). Also, create in R the required variables to run the following models:

$$y_i = \beta_0 + \beta_1 1_{difshare \geq 0,i} + \varepsilon_i \ (model\ 1)$$

$$y_i = \beta_0 + \beta_1 1_{difshare \geq 0,i} + \gamma_1 difshare_i + \varepsilon_i \ (model\ 2)$$

$$y_i = \beta_0 + \beta_1 1_{difshare \geq 0,i} + \gamma_1 difshare_i + \delta_1 difshare_i x 1_{difshare \geq 0,i} + \varepsilon_i \ (model\ 3)$$

$$y_i = \beta_0 + \beta_1 1_{difshare \geq 0,i} + \gamma_1 difshare_i + \gamma_2 difshare_i^2 + \varepsilon_i \ (model\ 4)$$

$$y_i = \beta_0 + \beta_1 1_{difshare \geq 0,i} + \gamma_1 difshare_i + \gamma_2 difshare_i^2 + \delta_1 difshare_i x 1_{difshare \geq 0,i} + \delta_2 difshare_i^2 x 1_{difshare \geq 0,i} + \varepsilon_i \ (model\ 5)$$

$$y_i = \beta_0 + \beta_1 1_{difshare \geq 0,i} + \gamma_1 difshare_i + \gamma_2 difshare_i^2 + +\gamma_3 difshare_i^3 + \varepsilon_i \ (model\ 6)$$

$$y_i = \beta_0 + \beta_1 1_{difshare \geq 0,i} + \gamma_1 difshare_i + \gamma_2 difshare_i^2 + \gamma_3 difshare_i^3 + \delta_1 difshare_i x 1_{difshare \geq 0,i} + \delta_2 difshare_i^2 x 1_{difshare \geq 0,i} + \delta_3 difshare_i^3 x 1_{difshare \geq 0,i} + \varepsilon_i \ (model\ 7)$$

where y_i corresponds to the **myoutcomenext** variable in the data set, and $1_{difshare \geq 0}$ to a dummy variable that indicates whether the party of the candidate won in the previous election.

Question 14
For which of the models do you find that the effects of party incumbency over re-election is greater than 0.6? Select all that apply.

- □□ Model 1
- □□ Model 2
- □□ Model 3
- □□ Model 4
- □□ Model 5
- □□ Model 6
- □□ Model 7

Question 15
For which of the models can you reject the null hypothesis that the incumbent party has no advantage over re-election outcomes with a significance level of 99%? Select all that apply.

- □□ Model 1
- □□ Model 2
- □□ Model 3
- □□ Model 4
- □□ Model 5
- □□ Model 6
- □□ Model 7

Question 16
Now use the `RDestimate` command in R to estimate the effect non-parametrically. What is the point estimate that you obtain using this command?

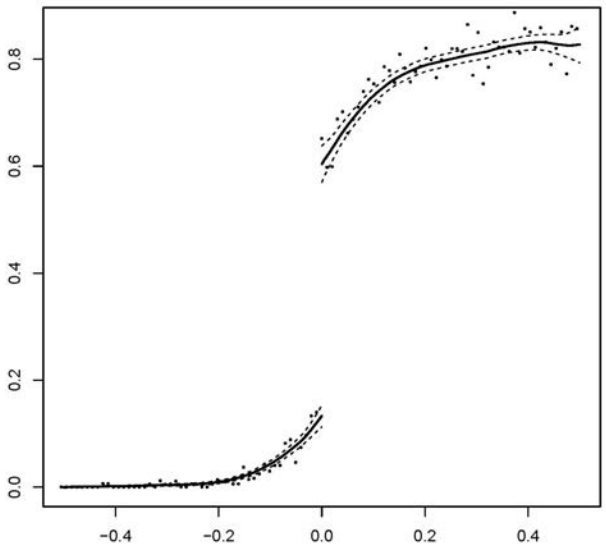
Do not round. Please input the answer exactly as it appears in your R output.

Question 17
The command also returns the estimate with half and double of the optimal bandwidth. Which one of the following values corresponds to the point estimate with **half of the bandwidth?**

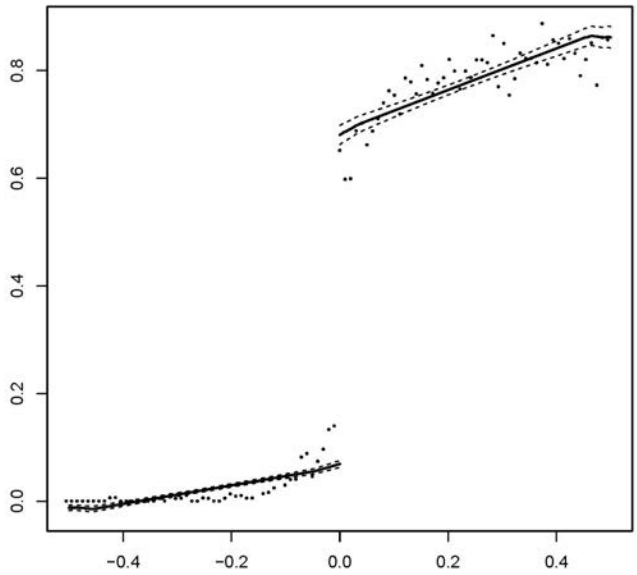
- •□ 0.5118950
- •□ 0.4510954
- •□ 0.4707463

Now take a look at the following plots:

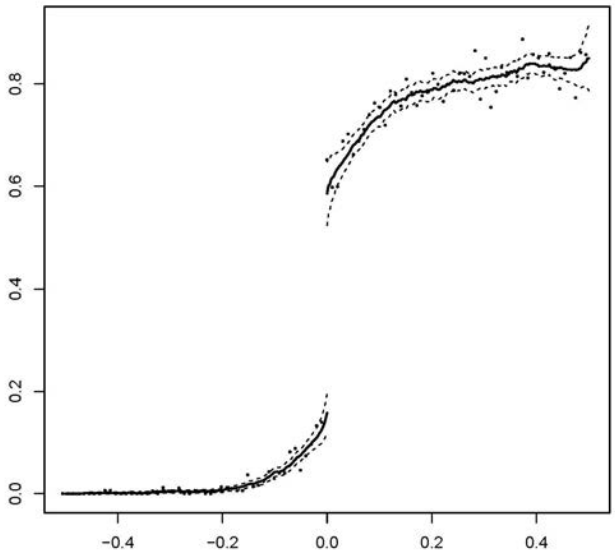
Plot A:



Plot B:



Plot C:



Question 18

One of them was done with the optimal bandwidth, another one with three times the optimal bandwidth and a third one with one-third of the optimal bandwidth. Rank them based on the size of the bandwidth (from largest to smallest). What is the ranking?

- B, C, A
- B, A, C
- C, A, B
- A, B, C

14.310x Data Analysis for Social Scientists
Spring 2023