PROBLEM SET 3

**Due on Monday, April 1, 2024**

I - INSTRUCTIONS

To successfully complete this problem set, please follow these steps:

1. Download this Word document file into your computer and download the datasets into a data subfolder in your problem set-specific RStudio or Stata Project directory.
2. Insert your answers into this document and organize your code in a Stata or R script. You can also insert non-Word objects such as handwritten work or screenshots in your answers.
3. Once your document is complete, please save it as a PDF.
4. Please submit an electronic copy of the **PDF** and your **replicable Stata or R script** to the Canvas assignment page.

II - IDENTIFICATION

1. Your information

|  |  |
| --- | --- |
| Your Last Name: | *Boochever* |
| Your First Name: | *Oscar* |

(2) Group Members (please list the classmates you worked with on this problem set):

|  |
| --- |
|  |

1. Compliance with Harvard Kennedy School Academic Code[[1]](#footnote-1) (mark with an X below)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **Yes** | **No** | | I certify that my work in this problem set complies with the Harvard Kennedy School Academic Code |  |  | |

For this problem set, we will be examining the methods used in the following paper:

Ozier, Owen. (2018). “The Impact of Secondary Schooling in Kenya: A Regression Discontinuity Analysis,” Journal of Human Resources, 53(1), 157-188. https://doi.org/10.3368/jhr.53.1.0915-7407.

# Conceptual Questions (32 points)

|  |
| --- |
| **Instructions:** Please keep your answers *concise*. Most questions can be answered in 1-2 sentences. Bolding or italicizing keywords also help grading. |

1. Clearly state the primary research question that the author is trying to answer. Does this research question have any policy implications? Explain these implications in 1-2 sentences. (2 points)

|  |
| --- |
|  |

1. In 3-5 sentences, explain the main finding of the paper using non-technical jargon, as if you were writing a brief policy memo. (2 points)

|  |
| --- |
|  |

For the following questions, consider Section III of the paper (Empirical Strategy).

1. The author used a regression discontinuity design because he believed a simple OLS specification would be insufficient. Consider the effect of attending any secondary school on one of the outcomes of interest (educational attainment, low-skill self-employment rate, fertility). What are two possible confounders (omitted variables) that would bias the results from a simple OLS specification? Explain the mechanism of the omitted variable and use the omitted variable bias formula to argue whether it would lead to an understatement or overstatement of the true effect. (3 points)

|  |
| --- |
|  |
|  |

1. Describe how the discontinuity the author exploits corrects for the type of omitted variable bias you explored in the previous question, and consequently achieves a causal explanation of the relationship of interest. (2 points)

|  |
| --- |
|  |

1. Why is it important to test for continuity of pre-treatment observable characteristics across the test score cutoff? (2 points)

|  |
| --- |
|  |

1. Explain the purpose of Figure 4. How does this compare to Figure 6? Explain how both figures are constructed. (3 points)

|  |
| --- |
|  |

1. Explain why the manipulation of the cutoff is a concern in an RD design. Explain what it would mean in this context, and how the author addresses this concern. (3 points)

|  |
| --- |
|  |

1. Consider Table 2.
   1. Interpret columns 2, 5, and 8. (3 points)

|  |
| --- |
|  |

* 1. Pick one column of columns 2, 5, and 8, and evaluate whether the result is statistically and economically significant. (1 point)

|  |
| --- |
|  |

1. Consider the difference between a sharp and a fuzzy RD design.
   1. What design does the author use? Why is it appropriate in this context? (1 point)
   2. How is the other design different? Explain how it would be constructed. (1 point)
   3. If the author had used the other design, what difference would it have made? Consider which group is induced to receive treatment in each context and how this affects the interpretation of the estimates. (2 points)
   4. In the context of a fuzzy RD design, how are the ITT (intent to treat) and LATE (local average treatment effect) related? Why would policymakers care more about the ITT in certain contexts? (2 points)

|  |
| --- |
|  |
|  |

1. Explain in your own words what bandwidth refers to in the context of an RD design and in this particular context. Generally, do larger bandwidths lead to more or less bias? Discuss what tradeoffs are involved in choosing between larger and smaller bandwidths. (3 points)

|  |
| --- |
|  |

1. List potential threats to either the internal or the external validity in this study. Explain what the potential threat is, and whether it should be a major concern for policymakers. (2 points)

|  |
| --- |
|  |

# Data Analysis Questions (20 points)

|  |
| --- |
| **Instructions for Stata or R code:** Follow the guidelines when starting your Stata or R script.   * + - 1. Do not leave package installation commands in your script.       2. Do leave package loading commands at the top of your script.       3. *Only* load packages that you actually need for the *particular* script.   These guidelines have been mentioned before, but this new [screencast](https://vimeo.com/399959368) consolidates them and explains the reason behind each. Please take a look. Also, use relative paths in a project, instead of hard-coded absolute paths, for input/output. |

In the following, we will replicate some of the results of Ozier’s paper.

Download the dataset available in the course website. Here are the main variables of interest:

* + **kcpe\_self\_or\_matched\_recent:** Most recent, self-reported KCPE score, corrected if administrative data is available (500 scale)
  + **finishsecondary:** Indicator for completing at least 12 years of schooling (secondary school)
  + **has\_score\_2016:** Indicator for reporting a (valid possible) KCPE score
  + **rkcpe:** KCPE score (first attempt, admin data confirmed), rescaled and recentered at the relevant cutoffs ((KCPE – cutoff)/100)
  + **passrkcpe:** Indicator for whether the KCPE score (first attempt, admin data confirmed, recentered on cutoffs) exceeds the relevant cutoff
  + **int\_pass\_rkcpe:** Interaction between **rkcpe** and **passrkcpe**
  + **female:** Indicator for whether respondent is female
  + **ravens\_plus\_vocab\_standardized:** Standardized sum of standardized scores on the Ravens B and Vocabulary tests

1. Create summary statistics for **kcpe\_self\_or\_matched\_recent, finishsecondary,** and **two other** variables from Table 1 of your choice. To match Table 1, restrict to observations with a valid KCPE score (**has\_score\_2016** == 1), and for outcome variables, additionally restrict to the 80-point bandwidth around the score cutoff (the absolute value of **rkcpe** < 0.8). Note that Panels D and E require further restrictions. (4 points)

|  |
| --- |
|  |

1. Create a table illustrating the first-stage effect of the test score cutoff (**passrkcpe**) on the probability of completing secondary school (**finishsecondary**). Replicate the coefficients and standard errors from Columns 1, 4, and 7 of Table 2, the first-stage estimates for the three different samples (Pooled, Male, and Female) without controls. Use a linear model, a uniform kernel, and an 80-point bandwidth around the score cutoff (set the absolute value of **rkcpe** < 0.8). Please cluster at the test score level (**rkcpe**), but note that it is up for debate when standard errors should be clustered in regression discontinuity designs. (4 points)

|  |
| --- |
|  |

1. Create a table replicating the coefficients and standard errors from Column 3 of Table 3, representing the effect of completing secondary school (**finishsecondary**) on the vocabulary and non-verbal reasoning tests (**ravens\_plus\_vocab\_standardized**). Use a linear model, uniform kernel weights, and an 80-point bandwidth. Please cluster at the test score level (**rkcpe**). Note that you will need to instrument for secondary school completion with the test score cutoff. (3 points)

|  |
| --- |
|  |

1. Create a regression discontinuity plot using a linear polynomial approximation, to illustrate the effect of scoring above the cutoff on cognitive scores in adulthood (**ravens\_plus\_vocab\_standardized**). To replicate panel B of Figure 6, use data within an 80-point bandwidth of the score cutoff, and use evenly-spaced spaced bins containing 10 points.
   1. Report the local linear estimates of the average treatment effects around the cutoff, and the 95% robust confidence intervals and robust p-values. To replicate Column 3 of Table 3, control for gender, and use uniform kernel weights and an 80-point bandwidth. Please cluster at the test score level (**rkcpe**). *(Hint: follow Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). A practical introduction to regression discontinuity designs: Foundations. Cambridge University Press. Use the rdrobust and rdplot packages in R and Stata. Note that standard errors will differ from Ozier’s.)* (4 points)
   2. Explain in plain English the advantages and disadvantages behind the methods Ozier chose (uniform kernel weights, bandwidth selection, and choice for the number of bins), as described in Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). (2 points)

|  |
| --- |
|  |

1. Implement a manipulation test based on density discontinuity (following Cattaneo et al., 2020) to assess whether there is manipulation of the running variable (**rkcpe**) at the cutoff or not in the optimal bandwidth selected. Interpret the results. (3 points)
   1. Use the default settings for the functions in Stata or R. Note that this will be similar to but not perfectly match Figure 1, since Cattaneo et al. (2020) is an updated version of the McCrary sorting test that Ozier uses.
   2. *Hint: Use the “rddensity” and the “rdplotdensity” of the rddensity package in R for the manipulation test based on the density. In Stata, you will need to install rddensity and lpdensity.*

|  |
| --- |
|  |

**Reminder: please include your replicable script in your submission following the package loading guidelines.**

# RDs in Your Own Work (8 points)

1. Think about a social relationship that would be best studied using an RD design. Briefly state the research question and the main variables of interest in non-technical terms. (4 points)

|  |
| --- |
|  |

1. Write out the empirical specification you would use and explain the equation. (2 points)

|  |
| --- |
|  |

1. What could be a potential threat to the validity of your RD design? (2 points)

|  |
| --- |
|  |

# References

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698-714.

Cattaneo, M. D., Jansson, M., & Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, *115*(531), 1449-1455.

1. We abide by the Harvard Kennedy School Academic [code](https://www.hks.harvard.edu/educational-programs/academic-calendars-policies/student-handbook/general-regulations-and-1) for all aspects of the course. In terms of problem sets, unless explicitly written otherwise, the norms are the following: You are free (and encouraged) to discuss problem sets with your classmates. However, you must hand in your own unique written work and code in all cases. Any copy/paste of another’s work is plagiarism. In other words, you can work with your classmate(s), sitting side-by-side and going through the problem set question-by-question, but you must each type your own answers and your own code. For more details, please see syllabus. [↑](#footnote-ref-1)