PROBLEM SET 3

**Due on Tuesday April 4, 2023**

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 Project directory.
2. Insert your answers into this document and organize your code in a 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 R script** to the Canvas assignment page.

II - IDENTIFICATION

1. Your information

|  |  |
| --- | --- |
| Your Last Name: |  |
| Your First Name: |  |

(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:

Ludwig, J., and Miller, D. L. (2007), “Does Head Start Improve Children’s Life Chances? Evidence from a Regression Discontinuity Design,” Quarterly Journal of Economics, 122, 159–208

# Conceptual Questions

|  |
| --- |
| **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.

|  |
| --- |
|  |

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

|  |
| --- |
|  |

For the following questions, consider Section II of the paper (Research Design).

1. The authors used a regression discontinuity design because they believed a simple OLS specification would be insufficient. Consider the effect of Head Start on human capital outcomes (education and/or health). 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.

|  |
| --- |
|  |

1. Describe how the discontinuity the authors exploit helps correct the type of omitted variable bias you explored in the previous question, and consequently achieve a causal explanation of the relationship of interest. Use your own words adapted to the context of this case.

|  |
| --- |
|  |

1. Why is it important to test for continuity of pre-treatment observable characteristics across the program eligibility threshold?

|  |
| --- |
|  |

1. Explain the purpose of Table I, and how it is constructed.

|  |
| --- |
|  |

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’s argument addresses this concern.

|  |
| --- |
|  |

1. Consider Figures 1-3 and Table II.
   1. Interpret the three ‘Nonparametric’ columns of Table II.

|  |
| --- |
|  |

* 1. Pick one dependent variable and judge whether the results in these three columns are statistically and economically significant.

|  |
| --- |
|  |

* 1. Overall, do Figures 1-3provide evidence in favor of or against using the RD design?

|  |
| --- |
|  |

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?
   2. How is the other design different? Explain how it would be constructed.
   3. If the author had used the other design, what difference would it have made?
   4. In the context of a fuzzy RD design, how are the ITT and LATE related? Why would policymakers care more about the ITT in certain contexts?

|  |
| --- |
|  |
|  |

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

|  |
| --- |
|  |

1. Answer the following questions, each in a single sentence.
   1. Explain why the author includes Table III.

*Hint: see section VII.*

|  |
| --- |
|  |

* 1. Why do the authors also look at mortality from injuries?

|  |
| --- |
|  |

* 1. Explain why the author includes Table IV.

|  |
| --- |
|  |

1. Consider different estimation methods of the RD design. What is the difference between a parametric and a non-parametric method in this context? Which form does the author use? What role do kernels play?

|  |
| --- |
|  |

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 trying to understand this evidence.

|  |
| --- |
|  |
|  |

1. Now consider Section IX.B of the paper (Specification Tests).
   1. What is the author trying to show in this section?

|  |
| --- |
|  |

* 1. What is the key logic of the “pseudo-cutoff” identification strategy in this context?

|  |
| --- |
|  |

* 1. Do you find it convincing?

|  |
| --- |
|  |

# Bunching

1. Suppose we are interested in the effect of an income tax reform. Before the tax reform, there is a constant tax rate of *t0*, i.e., everyone pays *t0* percent of their income in taxes. The tax reform is as follows. If you report to earn below some threshold *z\*,* you still pay *t0.* If you report to earn above *z\*,* you have to pay *t1* percent of your income in taxes (which is higher than *t0*).

* With *z* being pre-tax income, call *h0(z)* the smooth density of income before the tax reform.
* Call *h(z)* is density of income after the tax reform.

We can expect people to change their income around *z\** because of the tax reform (bunching).

* 1. Explain intuitively how you estimate excess bunching in this setting.   
     *Hint: if it helps, consider the following figure, taken from Saez (2010).* A close up of a map

     Description automatically generated

|  |
| --- |
|  |

* 1. What is the tradeoff in choosing the width of the income bands surrounding *z\**? ( in the figure)

|  |
| --- |
|  |

* 1. Now suppose you had the income distribution before the tax reform *h0(z)* and after the tax reform *h(z).* How would the additional data help you better estimate the bunching due to the reform?

|  |
| --- |
|  |

# Data Analysis Questions

|  |
| --- |
| **Instructions for R code:** Follow the guidelines when starting your 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.  As always, the appendix contains specific explanations for each question and the videos linked in question 16 cover the basics of RD analysis in R. |

In the following, we will replicate some of the results of Ludwig and Miller’s paper.

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

* + **oldcode:** ID in Ludwig-Miller dataset
  + **povrate:** Poverty rate in 1960 relative to 300th poorest county (which had poverty rate 59.1984)
  + **mort\_age59\_related\_postHS:** Average Mortality rate per 100,000 for children aged 5-9 over 1973–83 due to causes addressed as part of Head Start's health services
  + **mort\_age59\_injury\_postHS:** Average Mortality rate per 100,000 for children aged 5-9 over 1973–83 due to injury
  + **census1960\_pop:** County population (1960 census)
  + **census1960\_pctsch1417:** Percent attending school, ages 14-17 (1960 census)
  + **census1960\_pctsch534:** Percent attending school, ages 5-34 (1960 census)
  + **census1960\_pop1417:** Population aged 14-17 (1960 census)
  + **census1960\_pop534:** Population aged 5-34 (1960 census)
  + **census1960\_pop25:** Population aged 25+ (1960 census)
  + **census1960\_pcturban:** Percent urban (1960 census)
  + **census1960\_pctblack:** Percent black (1960 census)

1. Create summary statistics for **povrate1960, mort\_age59\_related\_postHS** and **two other** variables of your choice.

|  |
| --- |
|  |

1. Create two balance tables for the 1960 and 1990 census variables (similar to Table I in the QJE article) comparing the county characteristics for counties above and below the poverty rate cutoff of 59.1984%. Add a column showing the estimated difference between the two sets of counties, and the p-value that this difference is statistically significant.

|  |
| --- |
|  |

1. Replicate panels A, B, C, and D of Figure IV using the most recent standards of discontinuity plots. Use triangular kernel weights, optimal MSE bandwidth selection, and optimal data-driven methods for choosing the number of bins to plot above and below the cutoff. Plot a linear polynomial approximation with their respective confidence of intervals. Report the local linear estimates of the average treatment effects around the cutoff, and the 95% robust confidence intervals and robust p-values. (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). Explain in plain English the reason for using the methods (triangular kernel weights, bandwidth selection, and choice for number of bins) described by Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019).

|  |
| --- |
|  |

1. Implement a McCrary (2008) sorting test and a manipulation test based on density discontinuity (following Cattaneo et al., 2020) to assess whether there is manipulation of the running variable at the cutoff or not in the optimal bandwidth selected in the replication of Panel A of figure IV. Interpret the results. *Hint: use the function“DCdensity”of the rdd package in R for the McCrary test. Use the “rddensity” and the “rdplotdensity” of the rddensity package in R for the manipulation test based on the density.*

|  |
| --- |
|  |

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

# RDs in Your Own Work

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.

|  |
| --- |
|  |

1. Write out the empirical specification you would use and explain the equation.

|  |
| --- |
|  |

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

|  |
| --- |
|  |

# 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)