PROBLEM SET 3

Due on Monday April 3, 2023, 10:00 am

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:	*Chaturvedi*
Your First Name:	*Shreya*

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

Manisha Jha, Neha Verma

(3) Compliance with Harvard Kennedy School Academic Code[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	X	

[1] We abide by the Harvard Kennedy School Academic code 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.

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 (30 points + 8 extra 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)

The paper studies if Head Start funding has long-term effects on health and education outcomes using a regression discontinuity (RD) design. This study has policy implications for early childhood education and health programs for disadvantaged populations, since the positive results could lead to increased funding and expansion. The study also looks at sustained impacts across gender and race, providing insights into the effectiveness of federal policies for marginalized populations.

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

The paper found that the Head Start funding had a significant impact on child mortality rates for causes affected by the program, with a 33-50% reduction in eligible counties. The impact was consistent across racial groups. The program also had a positive effect on educational outcomes for eligible cohorts, with higher rates of high school completion and college attendance.

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

3. 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. (2 points)

The simple OLS model would result in biased results because of omitted variables. Two possible confounding variables that could cause this bias are:
- The economic status of counties: The economic status of counties could lead to an overstatement of the effect of Head Start because more economically prosperous counties could have better educational outcomes
-The socio-economic characteristics of local service providers: Service providers that cater to economically disadvantaged communities could lead to an understatement of the effect because children from such institutions tend to have worse educational outcomes.

4. 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. (2 points)

To achieve a causal explanation of the relationship of interest, the authors use a discontinuity in Head Start's program funding, tied to grant-writing assistance for the 300 poorest US counties just above a certain threshold.
The assistance led to 50-100% higher program participation and funding for treated counties, creating treatment and counterfactual control groups similar at baseline. This discontinuity serves as a source of exogenous variation in the treatment, eliminating omitted variable bias and estimating causal impact, which would have been afflicted in a simple OLS specification.

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

Testing for continuity of pre-treatment observable characteristics is important to ensure unbiasedness of the treatment effect in a regression discontinuity design. Discontinuity in observable characteristics at the threshold may indicate a selection bias or manipulation of eligibility criteria. Continuity in pre-treatment variables acts as a balance check, providing evidence that the selected treatment and control groups are comparable, increasing the credibility of the design. Large differences in pre-treatment characteristics would suggest a non-random occurrence, biasing the estimated treatment effect. It is important to ensure that any difference in outcomes between the *two groups can be attributed to the program itself, rather than other factors that might have affected eligibility or the outcomes of interest.*

6. Explain the purpose of Table I, and how it is constructed. (1 points)

   Table 1 establishes the relationship between eligibility criteria and treatment by comparing Head Start spending for treatment and control counties. It also provides summary statistics to assess the balance of observable characteristics between the two groups. The table is constructed by computing means and standard deviations of pre-treatment observables for counties just above and just below the eligibility threshold (10pp poverty), such as rural/urban shares, race, per capita income, baseline participation and spending rates.

7. 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. (3 points)

   Manipulation of the cutoff in an RD design creates bias in the causal estimate of program impact. The authors address this concern by using a predetermined variable (poverty rates from five years prior) as the cutoff, which eliminates the possibility of manipulation. There was also no incentive for counties to manipulate the cutoff, as the OEO had excess funding supply rather than demand. The authors' approach mitigates selection bias and ensures the validity of the results.

8. Consider Figures 1-3 and Table II.
   a. Interpret the three 'Nonparametric' columns of Table II. (3 points)

      Table II reports the results of the program on Head Start spending and participation, with two dependent variables and three bandwidths of 9, 18, and 36. The columns show the number of counties with nonzero weight, the point estimate, standard error, and p-value for each variable. The nonparametric columns present the results from nonparametric RD specifications, with varying bandwidths of 43, 96, and 288 counties. The results show an average increase in probability of Head Start participation and spending per child, with the point estimates declining as the bandwidth increases. For example, the point estimate for Head Start participation in the first year follow-up sample declines from 0.24 to 0.15 as the bandwidth increases from 9 to 36, and the p-value declines from 0.19 to 0.09.

   b. Pick one dependent variable and judge whether the results in these three columns are statistically and economically significant. (1 points)

      Focusing on the Head Start participation for the first year follow up sample, the results show that the point estimate is not statistically significant for the first *column with a bandwidth of 9, but is statistically significant at the 10% level for the second and third columns with bandwidths of 18 and 36, respectively. These estimates are also economically significant as they represent an increase from the average increase in probability of participation, suggesting a long-term impact of the program. Therefore, the results suggest that the Head Start program has a positive impact on participation, particularly when analyzed with larger bandwidths.*

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

Figures 1-3 provide strong evidence in favor of using the RD design, as they clearly demonstrate a discontinuity in both Head Start participation and funding at the poverty rate cutoff. This indicates that targeted poor counties, which received technical assistance, were more likely to participate and receive funds. Additionally, Figure 3 shows that there is no statistically significant discontinuity in other forms of federal social spending around the poverty rate cutoff, suggesting that the only salient difference in spending across groups relates to the provision of Head Start-related technical assistance for eligible counties. The authors note that their estimates are robust to different bandwidths, which further supports the use of the RD design. Therefore, overall, Figures 1-3 provide strong evidence in favor of using the RD design.

9. Consider the difference between a sharp and a fuzzy RD design.
   a. What design does the author use? Why is it appropriate in this context?
   b. How is the other design different? Explain how it would be constructed.
   c. If the author had used the other design, what difference would it have made?
   d. 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? (4 points)

The authors used a sharp RD design, which is appropriate in this context as the Head Start eligibility criterion creates a clear cutoff separating eligible children from ineligible ones. The fuzzy design is appropriate for situations where crossing the threshold does not ensure treatment and instead significantly increases the likelihood of treatment. This can be constructed using an IV method, where the relevant instrument is a dummy for being above/below the threshold. If the correct instrumental variable were used, there would have been no difference between the choice of sharp or fuzzy design.

Under a fuzzy RD design, the ITT effect is the estimated average treatment effect for individuals assigned treatment, regardless of whether or not they received the treatment. The LATE estimates the average treatment effect for those whose treatment assignment was affected by being just above or below the cutoff. The LATE is thus a *subset of the broader ITT. In general, policymakers would prefer the ITT in cases where eligibility does not guarantee treatment since there might be non-random reasons why individuals did not receive treatment even if they were assigned to receive treatment.*

10. 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. (3 points)

Bandwidth in the context of an RD design refers to the range around the threshold that determines the sample size included in the analysis. In this study, it refers to the range of poverty rates above and below the cutoff, with bandwidths of 9, 18, and 36 percentage points. Generally, larger bandwidths lead to more bias, but more precise estimates because they include more observations. However, including samples further away from the threshold may introduce omitted variable bias and specification errors, creating a tradeoff between precision and bias that must be balanced when choosing the bandwidth.

11. **[Optional]** Answer the following questions, each in a single sentence.
    a. Explain why the author includes Table III. (+ points)
    Hint: see section VII.

    The author includes Table III to demonstrate the relationship between increased Head Start participation at the OEO cutoff and decreased mortality rates for Head Start-related causes in children aged 5 to 9.

    b. Why do the authors also look at mortality from injuries? (1 point)

    The authors include mortality rates from injuries to validate that there is no significant discontinuity at the cutoff, which supports the idea that the difference in mortality rates for Head Start-related causes are due to the program and not other baseline characteristics.

    c. Explain why the author includes Table IV. (2 points)

    Table IV is included by the authors to estimate the long-term effects of Head Start on educational attainment, specifically high school and college completion, in order to determine if there is a more significant causal impact on educational attainment for those cohorts who were exposed to the program *during their childhood and reached adulthood after 1990, compared to birth cohorts.*

12. 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? (3 points)

    The difference between parametric and non-parametric methods in the RD design is that the former assumes a specific functional form while the latter does not. The authors of this study use both methods, but their preferred estimates are from the non-parametric method as it allows them to control for unobserved variables that vary with county poverty rates. Kernels play a role in weighting observations based on their distance from the cutoff. The authors use a locally weighted kernel regression method with a triangle kernel to estimate treatment effects. For the parametric RD, the authors estimate treatment effects using different polynomial functions of the poverty rate, calculated using counties near the OEO cutoff. On the other hand, for the non-parametric RD, the authors use local linear regressions to estimate the left and right limits of the discontinuity, and the difference between the two is the estimated treatment impact.

13. 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. (2 points)

    Internal validity threat: Technical assistance from OEO in grant writing may overstate treatment effect, as counties receiving assistance could be more motivated to implement Head Start programs well.

    External validity threats: Analysis includes mostly southern US counties, possibly limiting generalizability to other regions or in other time periods wherein the technical assistance was not given.

14. **[Optional]** Now consider Section IX.B of the paper (Specification Tests).
    a. What is the author trying to show in this section? (1 point)

       The author is trying to show that non-random migration patterns are not a threat to the internal validity of the treatment effects estimated.

b. What is the key logic of the “pseudo-cutoff” identification strategy in this context? (1 point)

The "pseudo-cutoff" identification strategy is used in this study to ensure that discontinuities only occur at the OEO cutoff and not at any other cutoffs. This helps validate that the treatment effects are due to changes in outcomes and not the choice of functional form. By demonstrating that the cut-off is not observed in other random cutoffs, this strategy strengthens the validity of the study.

c. Do you find it convincing? (1 point)

I found it largely convincing as I thought that the authors sufficiently made a case against the main threats to the validity of their estimates.

Data Analysis Questions (22 points)

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 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 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)

15. Create summary statistics for **povrate1960, mort_age59_related_postHS** and **two other** variables of your choice. (4 points)

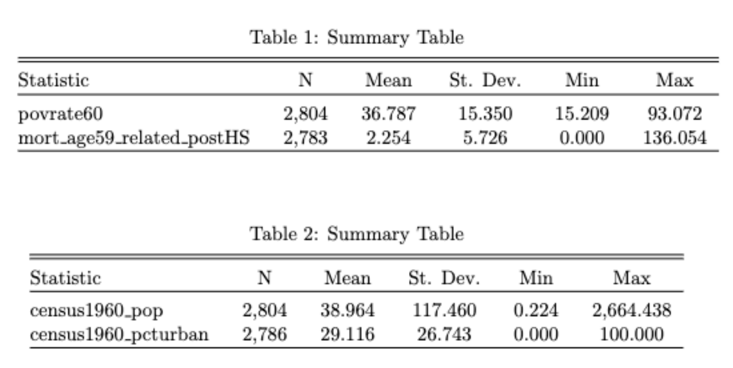


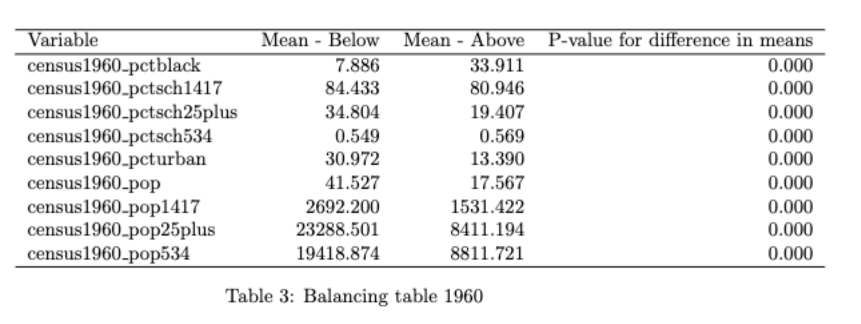
Table 1: Summary Table

Statistic	N	Mean	St. Dev.	Min	Max
povrate60	2,804	36.787	15.350	15.209	93.072
mort_age59_related_postHS	2,783	2.254	5.726	0.000	136.054

Table 2: Summary Table

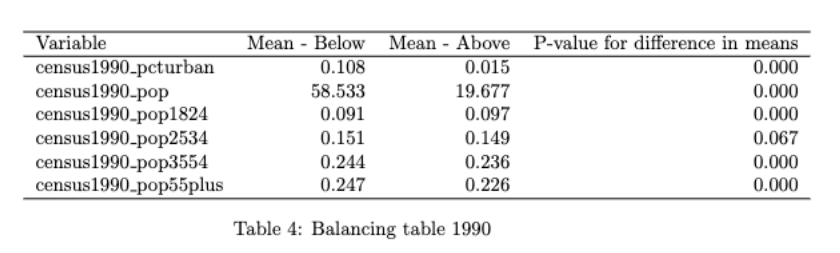
Statistic	N	Mean	St. Dev.	Min	Max
census1960_pop	2,804	38.964	117.460	0.224	2,664.438
census1960_pcturban	2,786	29.116	26.743	0.000	100.000

16. 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. (6 points)



Variable	Mean - Below	Mean - Above	P-value for difference in means
census1960_pctblack	7.886	33.911	0.000
census1960_pctsch1417	84.433	80.946	0.000
census1960_pctsch25plus	34.804	19.407	0.000
census1960_pctsch534	0.549	0.569	0.000
census1960_pcturban	30.972	13.390	0.000
census1960_pop	41.527	17.567	0.000
census1960_pop1417	2692.200	1531.422	0.000
census1960_pop25plus	23288.501	8411.194	0.000
census1960_pop534	19418.874	8811.721	0.000

Table 3: Balancing table 1960



Variable	Mean - Below	Mean - Above	P-value for difference in means
census1990_pcturban	0.108	0.015	0.000
census1990_pop	58.533	19.677	0.000
census1990_pop1824	0.091	0.097	0.000
census1990_pop2534	0.151	0.149	0.067
census1990_pop3554	0.244	0.236	0.000
census1990_pop55plus	0.247	0.226	0.000

Table 4: Balancing table 1990

17. 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. (8 points) *(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).*

18. Reminder: please include your replicable script in your submission following the package loading guidelines. (4 points)

RDs in Your Own Work (8 points)

19. 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. (2 points)

One potential research question I have previously worked on using an RD design is whether the implementation of the Panchayat Raj election system in Bihar has led to more inclusive and representative local governance. The main variables of interest were the characteristics of the elected representatives (primarily gender and caste) and the impact of these characteristics on the policies and decisions made by the local government, as well as the social network structure of the village.

By comparing the outcomes of the election at the cutoff point where the Panchayat Raj system was introduced, we tried to estimate the impact of this system on the network structure and local policies in the villages in Bihar.

20. Write out the empirical specification you would use and explain the equation. (4 points)

Yi = β0 + β1Xi + β2Zi + β3XiZi + ui

where:

Yi represents the outcome variable of interest, such as the proportion of inter-caste links in a village
Xi is a binary variable that equals 1 for villages that were exposed to the Panchayat Raj system and 0 for villages that were not exposed
Zi is a vector of controls that includes demographic characteristics of the village, such as population size, caste composition, and gender composition
β0 is the intercept
β1 measures the overall effect of the Panchayat Raj system on the outcome variable
β2 represents the effect of the control variables on the outcome variable
β3 captures the differential effect of the Panchayat Raj system on the outcome variable across villages with different characteristics
ui is the error term.

The coefficient β1 represents the causal effect of the Panchayat Raj system on inter-caste links formed in a village, after controlling for other relevant factors. A statistically significant and positive β1 would suggest that the introduction of the Panchayat Raj system has led to more inter-caste links in the village, while a negative or insignificant β1 would indicate that the system has had no impact or a negative impact on inter-caste links. By examining the interaction term β3, we can also investigate whether the effect of the Panchayat Raj system varies depending on the caste and gender characteristics of the village

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

One potential threat to the validity of an RD design in this context could be manipulation of the election results around the cutoff point. For example, if political parties or candidates strategically targeted specific areas or demographics to win the election, this could bias the results of the RD analysis.