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):

N/A

(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:

> 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)

> ***What are the impacts of secondary schooling in Kenya on human capital, occupational choice, and fertility among young adults, and what are the policy implications of these impacts?*** *This research question has significant policy implications as understanding the effects of secondary schooling can inform education policies aimed at improving labor market outcomes and reproductive health for young adults in Kenya, potentially contributing to broader development goals.*

2. 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)

> *Completing secondary school in Kenya has several positive impacts on young adults. For men, it leads to* ***better job opportunities****, with a decrease in low-skill self-employment and potential increases in formal employment. For women, completing secondary school is associated with a* ***reduced likelihood of teenage pregnancy****. These findings suggest that* ***investing in secondary education*** *could not only improve economic prospects but also contribute to better reproductive health outcomes for young people in Kenya, aligning with broader development goals.*

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

3. 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)

Outcome variable: Fertility/Pregnancy

*1. **Income** of households receiving the "treatment" (students going to secondary school). Households with higher income typically have better access to resources that promote safer sexual health choices, reducing the likelihood of teen pregnancy, and are more likely to have children go to secondary school.*

Pregancy = a0 + a1Secondary_i + vi

Pregnancy = b0 + b1Secondary_i + b2Income_i + ei

Income_i = l + dSecondary_i = n_i

*Bias: a1 = b1 + b2*d*

*Secondary education decreases teen pregnancy likelihood so b1: **negative***

*Income decreases teen pregnancy likelihood so b2: **negative***

*Wealthier households' children **more likely** to get secondary schooling so **d: positive***

*Since the effect of schooling on teen pregnancy likelihood would be negative, and bias is negative, failing to control for household income would lead to an **overstatement** of the true effect.*

2. Education level of parents in households receiving the "treatment" (students going to secondary school). Parents with higher education levels typically have better access to resources for, and better knowledge of, safe sexual health behavior, reducing the likelihood of teen pregnancy among their children, and they are more likely to prioritize their children's education by sending them to secondary school.

Pregnancy = a0 + a1Secondary_i + vi

Pregnancy = b0 + b1Secondary_i + b2Parental_Education_i + ei

Parental_Education_i = l + dSecondary_i = n_i

*Bias: a1 = b1 + b2 * d*

Secondary education decreases teen pregnancy likelihood so b1: negative

Higher parental education decreases teen pregnancy likelihood so b2: negative

Parents with higher education levels more likely to send their children to secondary school so d: positive

Similarly, since the effect of schooling on teen pregnancy likelihood would be negative, and bias is negative, failing to control for parental education would lead to an ***overstatement*** *of the true effect.*

4. 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)

By focusing on the threshold of admission to secondary school based on standardized test scores, the author ensures that individuals on either side of the threshold are similar in observable characteristics but differ in their likelihood of attending secondary school. This discontinuity in admission provides an exogenous source of variation in secondary schooling attendance, mitigating the bias that could arise from unobserved factors like parental education. Consequently, the author achieves a causal explanation of the relationship between secondary schooling attendance and outcomes such as teen pregnancy by leveraging this quasi-random variation, for individuals around the cutoff (a LATE). .

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

If there are discontinuities in observable characteristics at the threshold, it suggests that the assignment to treatment (attending secondary school) may not be truly exogenous, potentially biasing the estimated treatment effect. By verifying continuity, it strengthens the credibility of an RD design / causal interpretation by confirming that the assignment to treatment is not driven by systematic differences in observable characteristics.

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

The purpose of Figure 4 is to assess the validity of the smoothness assumption in the regression discontinuity design by examining the continuity of pre-treatment observable characteristics across the test score cutoff. It compares the characteristics such as gender, age, and parental education levels on either side of the cutoff point. Figure 4 is

constructed by plotting local quadratic regressions of these covariates against KCPE scores, ensuring that any differences are neither large enough to be important nor statistically significant.

Figure 6, on the other hand, serves to illustrate the first stage and reduced forms of the regression discontinuity design. It demonstrates the relationship between the KCPE scores (normalized so that the cutoff equals 0) and various outcomes, such as the probability of completing secondary school, cognitive performance, probability of self-employment, and probability of pregnancy by age 18. The figure is constructed by plotting these outcomes against the KCPE scores.

Figure 6 helps to assess the validity of the identifying assumption of the regression discontinuity design, which assumes that the treatment assignment (in this case, attending secondary school) is determined solely by the value of the forcing variable (here, the KCPE scores) and is otherwise unrelated to potential outcomes.

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 addresses this concern. (3 points)

In a regression discontinuity (RD) design, the manipulation of the cutoff is a concern because it would introduce bias into the estimated treatment effect. In the context of this study, manipulation of the cutoff could occur if individuals or institutions deliberately alter students' standardized test scores to influence their admission to secondary school, thus affecting the treatment assignment.

To address this concern, the author compares density of self-reported and confirmed KCPE scores around the cutoff. By demonstrating similar and continuous distributions around the cutoff (which appear normally distributed), the author provides evidence against manipulation of the cutoff.

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

In columns 2, 5, and 8 of Table 2, the author presents the estimated discontinuity in the probability of completing secondary school for different groups based on gender. Specifically:
Column 2 (Pooled): *The estimated discontinuity indicates a 15 percentage point change in the probability of completing secondary school when individuals are pooled together, regardless of gender. This means that there is a substantial increase in the likelihood of completing secondary school for individuals just above the cutoff compared to those just below it.*

Column 5 (Men): *For men, the estimated discontinuity suggests a 16 percentage point change in the probability of completing secondary school. This*

indicates a significant increase in the likelihood of completing secondary school for men just above the cutoff compared to those just below it.

Column 8 (Women): *Similarly, for women, the estimated discontinuity indicates a 13 percentage point change in the probability of completing secondary school. This suggests a significant increase in the likelihood of completing secondary school for women just above the cutoff compared to those just below it.*

The results provide evidence of a clear discontinuity in the probability of completing secondary school around the cutoff point, but suggests results are more pronounced for male students.

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

Table 2 Column 2: Pooled

Statistical significance: *All coefficients, except for the interaction term, are statistically significant at the 1% level.*

Economic significance: *The coefficient on "KCPE >= cutoff" indicates a 15 percentage point change in the probability of completing secondary school, while the coefficient on "KCPE centered at cutoff" suggests a larger effect, indicating a 27 percentage point change.*

Overall, the results are both statistically and economically significant, with substantial changes in the probability of completing secondary school associated with KCPE scores near the cutoff point.

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? (1 point)
   b. How is the other design different? Explain how it would be constructed. (1 point)
   c. 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)
   d. 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)

a. The author uses a ***fuzzy regression discontinuity design****. This design is appropriate because the admission to secondary schools is not solely determined by the KCPE*

score; there are additional factors influencing admission, such as available spaces and affordability of fees.

*b. The other design is a **sharp regression discontinuity design**. In this design, admission to secondary school would be solely determined by the KCPE score, with a clear cutoff point separating those who are admitted and those who are not.*

c. If the author had used the sharp regression discontinuity design instead, the estimates would reflect the treatment effect for individuals precisely at the cutoff point. This could lead to an overestimation of the treatment effect, as individuals just above the cutoff might differ in unobservable characteristics from those just below.

*d. In a fuzzy RD design, the **ITT (intent to treat)** effect reflects the average treatment effect for all individuals who are induced to receive treatment due to being close to the cutoff, regardless of whether they actually receive treatment. The **LATE (local average treatment effect)**, on the other hand, reflects the average treatment effect for compliers, i.e., those individuals who are induced to receive treatment and actually receive it. Policymakers might care more about the ITT in certain contexts because it provides information about the overall impact of a policy or intervention, including its effectiveness in inducing individuals to seek treatment. However, the LATE provides insight specifically into the effect of treatment on those who comply with it, which might be more relevant in understanding the effectiveness of the treatment itself.*

10. 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)

In the context of an RD design, the bandwidth refers to the range of observations around the cutoff point that are used for estimation. A larger bandwidth includes more observations both near and far from the cutoff, leading to smoother estimates but potentially higher bias due to reliance on extrapolation and functional form assumptions. Conversely, a smaller bandwidth focuses only on observations close to c, reducing bias but increasing variance due to fewer data points. Therefore, choosing the appropriate bandwidth involves a bias-variance tradeoff: larger bandwidths reduce bias but increase variance, while smaller bandwidths reduce bias but increase the risk of estimation error due to fewer observations. The goal is to balance this bias-variance tradeoff.

11. 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)

For external validity: since the RD analysis focuses on a specific cutoff score, the estimated treatment effects may not generalize the effect of secondary schooling for students at other levels of academic performance or to regions with different educational systems/cutoffs/enrollment policies. This could be a concern for policymakers if they wanted to implement lower or higher (relative) acceptance score measures in their respective jurisdictions' testing/enrollment criteria.

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 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

12. 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)

Variable	Min	Quartile_1	Median	Mean	Quartile_3	Max	SD
kcpe_self_or_matched_recent	90.00	220.00	255.00	254.50	290.00	435.00	52.23
finishsecondary	0.00	0.00	0.00	0.37	1.00	1.00	0.48
ravens_plus_vocab_standardized	-1.88	-0.05	0.58	0.49	1.07	1.73	0.76

13. 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)

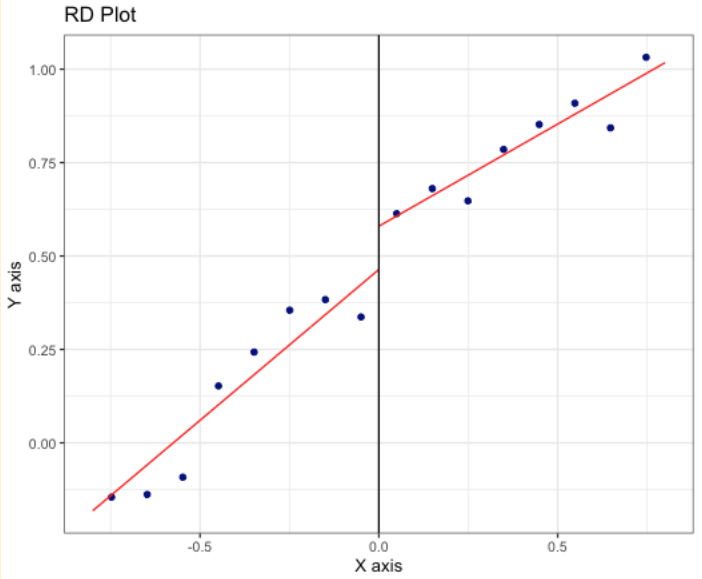
term	estimate_pooled	std.error_pooled	p.value_pooled	estimate_male	std.error_male	p.value_male	estimate_female	std.error_female	p.value_female
passrkcpe	0.16	0.04	0.000	0.17	0.05	0.001	0.16	0.06	0.006
rkcpe	0.27	0.06	0.000	0.30	0.09	0.001	0.24	0.08	0.002
passrkcpe:rkcpe	0.02	0.09	0.841	-0.02	0.11	0.871	-0.01	0.14	0.965
(Intercept)	0.33	0.02	0.000	0.39	0.04	0.000	0.27	0.04	0.000

(sorry for the small numbers above!)

14. 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)

	Coefficient	Standard.Error
(Intercept)	0.334	0.141
finishsecondary	0.670	0.283
rkcpe	0.637	0.168
female	-0.183	0.042
rkcpe:passrkcpe	-0.311	0.127

15. 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.
    a. 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)
    b. 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)



Method	Coef.	Std. Err.	z	P>|z|	[95% C.I.]
Conventional	0.110	0.023	4.818	0.000	[0.065 , 0.155]
Robust	-	-	6.301	0.000	[0.141 , 0.268]

Uniform Kernel Weights
Advantages: Uniform kernel weights assign equal weight to all observations within the bandwidth, simplifying the estimation process. They are easy to implement and interpret.

Disadvantages: Uniform kernel weights may not effectively capture the variation in data density around the cutoff point. They give the same importance to all observations within the bandwidth, which could lead to inefficient estimation, especially if the data density varies significantly.

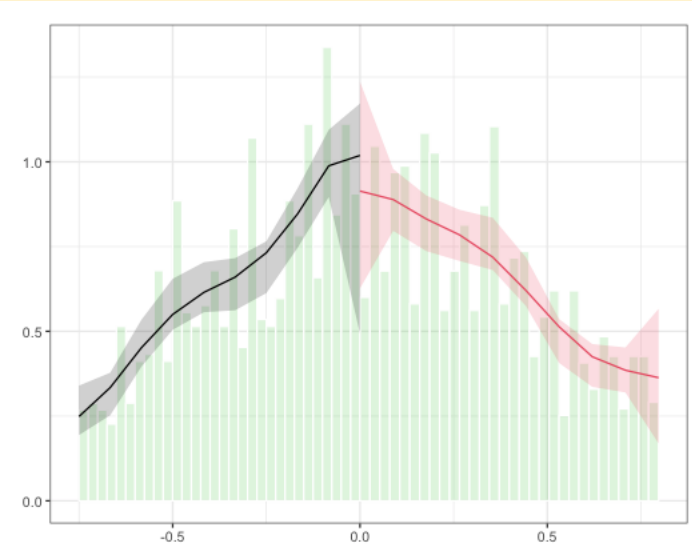
Bandwidth Selection
Advantages: A larger bandwidth increases the number of observations used in estimation, which can improve the generalizability of the estimates. On the other hand, a smaller bandwidth reduces bias but may increase variance.

Disadvantages: With this bias-variance tradeoff, the results implicitly do not account for scores outside of the bandwidth. There is an argument to be made that 0.8 does not sufficiently capture what happens outside of the bandwidth with regards to effect of scoring above the cutoff.

Number of Bins
Advantages: Dividing the data into evenly spaced bins allows for a clear visualization of the relationship between the running variable and the outcome variable. It simplifies the interpretation of the RD plot by summarizing the data within each bin, making it easier to identify patterns and trends.

Disadvantages: The choice of the number of bins can influence the visual appearance of the RD plot and the precision of the estimates – it conceals data density in that way. Using too few bins may oversimplify the data, while using too many bins may result in excessive noise or variability.

16. 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)
    a. 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.
    b. *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)

17. 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)

Research Question: *Does participation in a summer jobs program reduce the likelihood of recidivism for formerly juvenile-incarcerated high schoolers, where eligibility is determined by GPA, with a cutoff at 3.00?*

Main Variables of Interest:

- ***Treatment Variable:*** *Participation in the summer jobs program.*
- ***Outcome Variable:*** *Recidivism rates (e.g., percentage of individuals re-offending) among formerly juvenile-incarcerated high schoolers.*
- ***Running Variable:*** *GPA of the individual, used to determine eligibility for the summer jobs program.*

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

$Yi=\beta 0+\beta 1Ti+\beta 2Xi+\epsilon i$

Where:

- *Yi is the outcome variable (recidivism rates) for individual i.*
- *Ti is the treatment variable, indicating participation in the summer jobs program (1 for participants, 0 for non-participants).*
- *Xi is the running variable (GPA) of individual i, used to determine eligibility for the summer jobs program.*
- *$\beta 0$ represents the intercept, capturing the baseline level of recidivism rates for formerly juvenile-incarcerated high schoolers.*
- *$\beta 1$ represents the treatment effect, indicating the change in recidivism rates associated with participation in the summer jobs program.*
- *$\beta 2$ represents the effect of GPA on the outcome (recidivism rates).*
- *ϵi is the error term.*

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

One potential threat would be manipulation of GPA near the eligibility cutoff (3.00 in this case). Individuals might strategically adjust their GPA to become eligible for the program, or teachers might slightly curve scores near the cutoff, leading to a violation of the "as good as random" assumption underlying the RD design. To address this, we'd examine any signs of density manipulation around the cutoff.

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.