PROBLEM SET 3

Jacob Jameson

IDENTIFICATION

(1) Your information

Jacob Jameson

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

Bohan Li, Jenna Rogers, Nolan Kavanaugh, Becca Duflano

(3) Compliance with Harvard Kennedy School Academic Code

I certify that my work in this problem set complies with the Harvard Kennedy School Academic Code

Conceptual Questions

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.

The primary research question is: What is the effect of an increase in Head Start funding on the mortality and educational attainment of low-income children? This clearly has policy implications, since a desired effect on either or both of the outcomes would promote investment into Head Start and potentially other programs like it.

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.

This paper found that raising Head Start funding by 50-100 percent caused a sizable drop in five- to nine-year-old mortality rates due to health risks that Head Start is equipped to combat. The decrease seen ranged from 33 to 50 percent below the average risk for five- to nine-year-olds living in areas without federal Head Start funding. This decrease was enough to lower mortality due to the specific set of health risks in this low-income population down to the national average.

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.

Since Head Start is associated with increased educational attainment, we know that $\alpha_1 > 0$.

Two possible confounding factors are “parental motivation” and “family SES”.

If “parental motivation” is a confounder, then families who are more motivated to enroll their children in Head Start are likely to encourage their children to pursue more formal education in the long run, resulting in $\beta_2 > 0$ and a positive association between motivation and enrollment rates, $\gamma_1 > 0$. In this case, $\beta_2\gamma_1 > 0$, which leads to an overestimation of the true effect of Head Start on educational attainment.

If “family SES” is a confounder, then wealthier families, who are generally better connected to educational opportunities, are more likely to have the resources and know-how to secure a seat in a Head Start program, resulting in $\gamma_1 > 0$. Additionally, higher family SES is also associated with greater lifetime education, $\beta_2 > 0$. In this case, $\beta_2\gamma_1 > 0$, which again leads to an overestimation of the true effect of Head Start on educational attainment.

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.

The authors leverage the discontinuity created by the allocation of extra support to certain counties in the Head Start application process. As the financial characteristics of counties on either side of the threshold are similar, which side of the threshold a county ends up on can be considered as quasi-randomly assigned. By analyzing the difference in trends of outcomes just before and just after the discontinuity, this quasi-random assignment can ideally account for observable and unobservable characteristics that might lead to omitted variable bias in a causal estimate produced through OLS.

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

In order to have an internally valid estimate, counties on either side of the threshold must only vary in the treatment assignment, i.e. federal support for Head Start. Any discontinuity in pre-treatment characteristics would suggest manipulation in the receipt of federal support, driven by an unobserved omitted variable. This omitted variable would render non-causal any estimate of the program’s effect.

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

Table 1 aims to assess the similarity of counties on either side of the threshold in their observable characteristics. By showing that the two groups are similar, it bolsters the case that treatment assignment at the threshold was as good as random. The table is constructed by dividing counties with poverty rates within either +10% or −10% of the threshold and generating summary statistics for those two groups. We can see a large difference in federal spending for Head Start (as expected) but otherwise similar pre-treatment demographic and socioeconomic characteristics.

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.

Manipulating the cutoff in an RD design can cause a problem called bunching. This happens when more individuals or entities are pushed towards one side of the cutoff due to active manipulation. It violates the assumption of a smooth relationship between the cutoff and potential outcomes, which is necessary for estimating a counterfactual causal difference. In the case of grant-writing assistance, it could mean that more assistance was given to schools just above the cutoff or that the lowest income counties received more assistance, leaving those near the cutoff with little help. To address this issue, the authors argue that the cutoff was not manipulated because it was based on predetermined data and intentionally created by the grant-writing assistance administrators.

8. Consider Figures 1-3 and Table II.

a. Interpret the three 'Nonparametric' columns of Table II.

Each column represents an RD design done with a different nonparametric specification. The authors varied the nonparametric specifications by using three different bandwidths, shown in the first row for each column. The columns show dependent outcomes using a nonparametric RD specification with bandwidths of 9, 18, and 36 to test for potential fit and the sensitivity of the model. Using all three specifications, the authors are checking to see if receiving extra assistance applying for Head Start actually increased the funding received in those counties. Overall, these columns indicate that this relationship holds and that the regression discontinuity is valid.

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

According to the results from two of the nonparametric specifications, the likelihood of Head Start participation after a year experiences a significant increase at a 0.1 level of significance. All three coefficients demonstrate a minimum of 14% rise in the likelihood of Head Start participation. Even though the statistical significance is not at its maximum, in accordance with the study, I believe that these findings are still of considerable economic importance, indicating that the changes in Head Start participation resulting from funding assistance could lead to significantly different outcomes of mortality.

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

Figures 1-3 support the use of the RD design, despite the large standard errors and weak statistical significance observed in Table II. The visual evidence presented in Figure I and II using both nonparametric and flexible quadratic specifications show a jump in trends, suggesting an increase in Head Start participation and funding for counties that received the extra application support. Similarly, Figure III reinforces this by demonstrating consistency in trends for other social spending, where there is no visible gap before and after the threshold.

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?

a. The authors employ a sharp RD design to investigate the impact of Head Start assistance on mortality and education. This design is appropriate because they are interested in an ITT, which is more policy-relevant than the impact per dollar of funding. A fuzzy RD approach would be more suitable if they were interested in the effect of Head Start.

b. A fuzzy RD design assumes that being below a cutoff increases the probability of treatment. In the context of this study, treatment refers to the intention to assist with Head Start funding. A fuzzy RD would be constructed using the threshold as an instrumental variable and 2-stage least squares to find the causal effect.

c. Given that the threshold is a true sharp RD, using a fuzzy RD design would not alter the results. The qualitative evidence suggests that the probability of getting grant-writing assistance goes from 0 to 1, which is already captured in the sharp RD design.

d. In a fuzzy RD, the ITT is the impact of being offered grant-writing assistance, while the LATE is the impact on those who would not have applied for Head Start without assistance. Policymakers are more interested in the ITT since they cannot control who signs up for Head Start, but they can control if they fund extra assistance. The mortality and education outcomes speak to the returns on investment that may be important to policymakers before allocating funds towards grant-writing assistance.

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.

Bandwidth is the metric used to define the cutoff for scores included in the RD specification. Larger bandwidths include scores further away from the threshold, leading to more bias as the boundary of quasi-randomization is pushed. However, larger bandwidths also decrease the variance of the model. Using smaller bandwidths incorporates fewer data points, increasing the variance of possible results.

11. Answer the following questions, each in a single sentence.

a. Explain why the author includes Table III. Hint: see section VII.

Table III is important to include because it depicts the main findings of the paper using several different specifications for validity.

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

Mortality from injuries is important to examine because if the causal pathway for the RD holds, there should be no significant difference in something like injuries where Head Start funding is not put to use.

c. Explain why the author includes Table IV.

Table IV provides results from the rest of the initial specified analysis, which are causal estimates of the impact of Head Start on educational outcomes.

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?

The choice of a parametric or non-parametric method determines the equation used to model pre- and post-trends in the data. In this study, parametric methods involve a researcher-defined structure with equal data weighting, while non-parametric methods use a specified bandwidth and kernel to fit the model. The author utilized three non-parametric specifications with different bandwidths and the same triangle kernel, as well as flexible linear and quadratic parametric specifications, but they preferred the nonparametric estimates. The kernel specification is crucial in an RD design because it allows for a larger bandwidth, improving validity, while downweighting the influence of data points farther away from the threshold to reduce bias.

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.

One threat to external validity is threshold bias, where the effects seen at a specific threshold may not hold up when applied to other populations. Population differences, such as income level, politics, and social services, can also impact external validity. However, these concerns are only relevant if Head Start does not consider which applications are accepted and where funds are distributed. Another threat is scalability, which refers to a potential drop in the intervention's quality when applied nationwide. Policymakers may be particularly concerned about this threat, as it could affect the validity of the ITT estimates.

14. Now consider Section IX.B of the paper (Specification Tests).

  a. What is the author trying to show in this section?

In this section, the author is trying to clarify and validate their argument that the internal validity of their regression discontinuity holds up by referencing various tests of causal specificity.

  b. What is the key logic of the "pseudo-cutoff" identification strategy in this context?

The pseudo-cutoff identification strategy logic is that if the results found at the specified cutoff are invalid due to functional form misspecification, that similar results would be able to be found at other fake cutoffs.

  c. Do you find it convincing?

I think it is generally convincing against this one form of internal validation. There are certainly other scenarios in which significant effects are found using the specified cutoff and not at any pseudo-cutoffs and results are still internally questionable, such as the existence of another significant causal pathway.

15.

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

we could use the observed distribution in tax filers, h(z), with the tax threshold defined as $z*$. We can then define two bands of tax filers as $H*- = (z*-2\delta, z*-\delta) and H*+ = (z*+\delta, z*+2\delta)$, where δ is defined such that it does not enclose the tax files who are bunched near the threshold, $z*$. We estimate the excess bunching, then, as the sum density of those two bands, $H*$, minus the difference between them, $(H*\_ - H*+)$. Since those two bands should sum to the density in the same range of expected distribution, h0(z), any displacement will correspond to the tax filers who have shifted.

(b) What is the tradeoff in choosing the width of the income bands surrounding z*? (δ in the figure)

If the width is too narrow, not all of the bunching will be captured but what is captured is very likely to be attributable to bunching. With a wider band, you can capture all of the bunching but may include differences between the model and the counterfactual that are not attributable to bunching.

(c) 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?

To estimate excess bunching in this situation, h0(z) should serve as the counterfactual distribution and be compared to h(z). The excess density surrounding the threshold z* can be calculated as the shaded mass in pink, which is where h(z) greatly differs from h0(z) and there is distinctly excess mass between the two functions. Intuitively, integrating h0(z) – h(z) within some specified range centered at z* will get a bunching estimate of that region..

Data Questions

1. Create summary statistics for povrate60, mort_age59_related_postHS and two other variables of your choice.

```
data <- read_csv('headstart.csv')


sumtable(data,
         vars = c('povrate60', 'mort_age59_related_postHS',
                  'census1960_pop', 'census1990_pctblack'),
         out='latex')
```

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
povrate60	2804	37	15	15	24	47	93
mort_age59_related_postHS	2785	2.3	5.7	0	0	2.8	136
census1960_pop	2804	38964	117460	224	9133	33417	2664438
census1990_pctblack	2806	0.091	0.15	0	0.0015	0.11	0.86

2. 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.

```
data <- data[!is.na(data$povrate60),]

threshold <- 59.1984
data$indicator <- ifelse(data$povrate60 > threshold,1,0)

treated_data <- subset(data, data$indicator==1)
untreated_data <- subset(data, data$indicator==0)

no_obs_untreat <- nrow(untreated_data)

mean_pop60_u <- round(mean(untreated_data$census1960_pop, na.rm = TRUE), 3)
mean_pctsch534_u <- round(mean(untreated_data$census1960_pctsch534, na.rm = TRUE), 3)
mean_pctsch1417_u <- round(mean(untreated_data$census1960_pctsch1417, na.rm = TRUE), 3)
mean_pop1417_u <- round(mean(untreated_data$census1960_pop1417, na.rm = TRUE), 3)
mean_pop534_u <- round(mean(untreated_data$census1960_pop534, na.rm = TRUE), 3)
mean_pop25plus_u <- round(mean(untreated_data$census1960_pop25plus, na.rm = TRUE), 3)
mean_black_u <- round(mean(untreated_data$census1960_pctblack, na.rm = TRUE), 3)
mean_urban_u <- round(mean(untreated_data$census1960_pcturban, na.rm = TRUE), 3)

###sd
sd_pop60_u <- round(sd(untreated_data$census1960_pop, na.rm =T), 3)
sd_pctsch534_u <- round(sd(untreated_data$census1960_pctsch534, na.rm =T), 3)
sd_pctsch1417_u <- round(sd(untreated_data$census1960_pctsch1417, na.rm =T), 3)
sd_pop1417_u <- round(sd(untreated_data$census1960_pop1417, na.rm = T), 3)
sd_pop534_u <- round(sd(untreated_data$census1960_pop534, na.rm = T), 3)
```

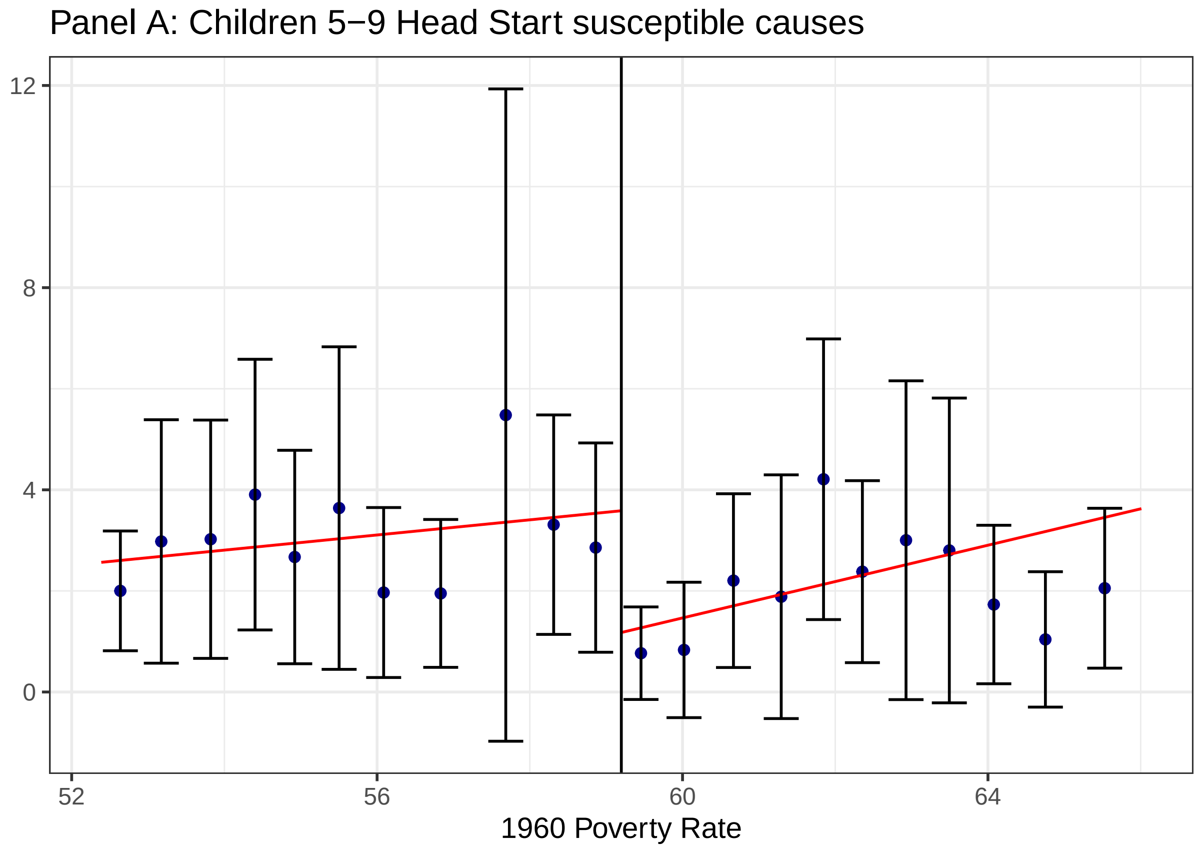
```r
sd_pop25plus_u <- round(sd(untreated_data$census1960_pop25plus, na.rm = T), 3)
sd_black_u <- round(sd(untreated_data$census1960_pctblack, na.rm =T), 3)
sd_urban_u <- round(sd(untreated_data$census1960_pcturban, na.rm =T), 3)

no_obs_treat <- nrow(treated_data)
mean_pop60_t <- round(mean(treated_data$census1960_pop, na.rm = TRUE), 3)
mean_pctsch534_t <- round(mean(treated_data$census1960_pctsch534, na.rm = TRUE), 3)
mean_pctsch1417_t <- round(mean(treated_data$census1960_pctsch1417, na.rm = TRUE), 3)
mean_pop1417_t <- round(mean(treated_data$census1960_pop1417, na.rm = TRUE), 3)
mean_pop534_t <- round(mean(treated_data$census1960_pop534, na.rm = TRUE), 3)
mean_pop25plus_t <- round(mean(treated_data$census1960_pop25plus, na.rm = TRUE), 3)
mean_black_t <- round(mean(treated_data$census1960_pctblack, na.rm = TRUE), 3)
mean_urban_t <- round(mean(treated_data$census1960_pcturban, na.rm = TRUE), 3)


sd_pop60_t <- round(sd(treated_data$census1960_pop, na.rm =T), 3)
sd_pctsch534_t <- round(sd(treated_data$census1960_pctsch534, na.rm =T), 3)
sd_pctsch1417_t <-round(sd(treated_data$census1960_pctsch1417, na.rm =T), 3)
sd_pop1417_t <-round(sd(treated_data$census1960_pop1417, na.rm =T), 3)
sd_pop534_t <-round(sd(treated_data$census1960_pop534, na.rm =T), 3)
sd_pop25plus_t <-round(sd(treated_data$census1960_pop25plus, na.rm =T), 3)
sd_black_t <-round(sd(treated_data$census1960_pctblack, na.rm =T), 3)
sd_urban_t <-round(sd(treated_data$census1960_pcturban, na.rm =T), 3)

table2_60 <- data.frame(
  c(no_obs_untreat, mean_pop60_u, mean_pctsch534_u, mean_pctsch1417_u, mean_pop534_u,
    mean_pop1417_u, mean_pop25plus_u, mean_black_u, mean_urban_u),
  c("", sd_pop60_u, sd_pctsch534_u, sd_pctsch1417_u, sd_pop534_u,
    sd_pop1417_u, sd_pop25plus_u, sd_black_u, sd_urban_u),
  c(no_obs_treat, mean_pop60_t, mean_pctsch534_t, mean_pctsch1417_t, mean_pop534_t,
    mean_pop1417_t, mean_pop25plus_t, mean_black_t, mean_urban_t),
  c("", sd_pop60_t, sd_pctsch534_t, sd_pctsch1417_t, sd_pop534_t,
    sd_pop1417_t, sd_pop25plus_t, sd_black_t, sd_urban_t)
)

rownames(table2_60) <- c("# of observations",
                         "county population",
                         "% attending school, ages 5-34",
                         "% attending school, ages 14-17",
                         "Population aged 5-34", "Population aged 14-17",
                         "Population aged 25+", "% black", "% urban" )

colnames(table2_60) <- c("Mean - < 59.1984%",
                         "Std. Error - < 59.1984%",
                         "Mean - > 59.1984%",
                         "Std. Error - > 59.1984%")


kable(table2_60) %>%
  column_spec(1, width = "5em")
```

	Mean - < 59.1984%	Std. Error - < 59.1984%	Mean - > 59.1984%	Std. Error - > 59.1984%
# of observations	2504.000		300.000	
county population	41527.335	123923.452	17566.553	16390.402
% attending school, ages 5-34	0.549	0.059	0.569	0.047
% attending school, ages 14-17	84.433	16.852	80.946	10.853
Population aged 5-34	19418.874	55386.663	8811.721	8620.457
Population aged 14-17	2692.200	6989.71	1531.422	1358.707
Population aged 25+	23288.501	74131.161	8411.194	7239.242
% black	7.886	12.855	33.911	26.374
% urban	30.972	26.991	13.390	18.087

3. 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).

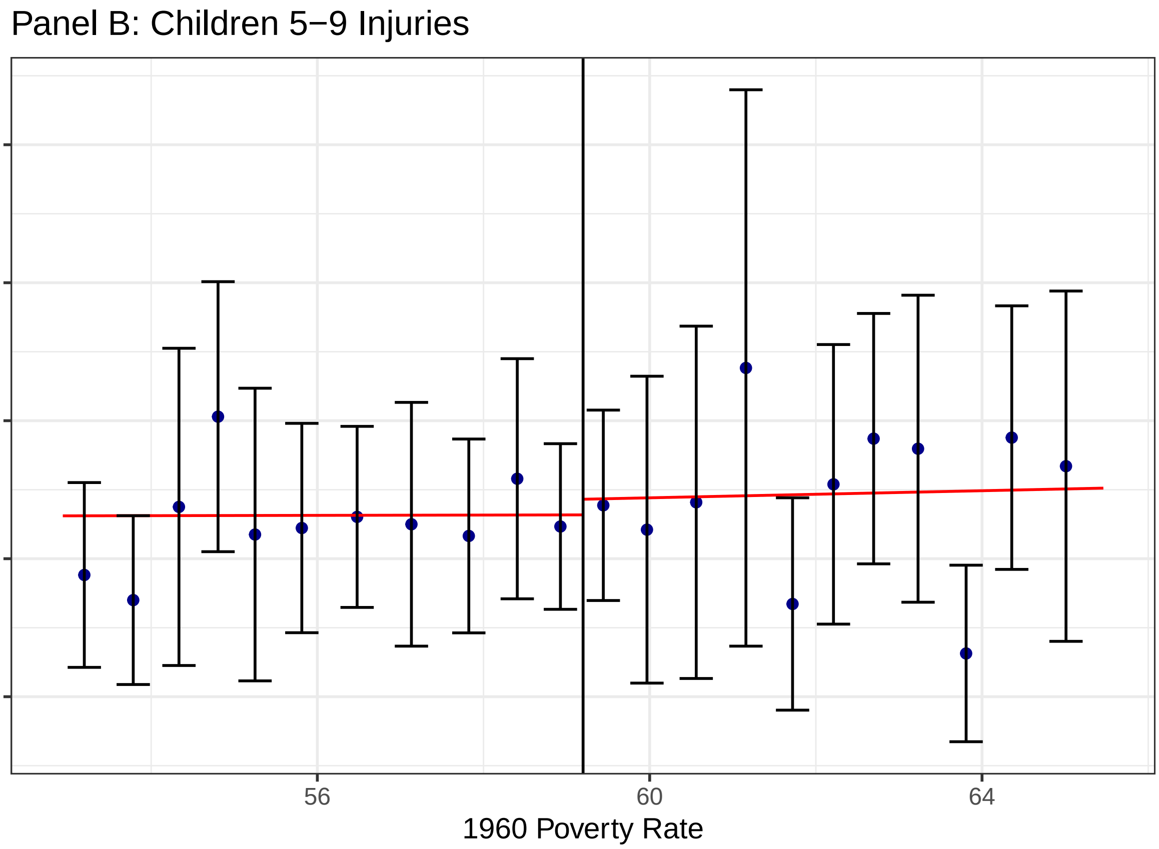
```
library(rdrobust)

helper <- function(data, title, mort_col) {
  bandwidth <- rdbwselect(data[[mort_col]], data$povrate60, c=threshold, p = 1,
                          kernel = "triangular", bwselect = "mserd") #MSE-optimal bandwidth
  h <- bandwidth$bws[1]
  panel_a <- rdplot(data[[mort_col]], data$povrate60,
                    subset = data$povrate60>=threshold-h & data$povrate60<=threshold+h,
                    c=threshold, p = 1,
                    binselect = "qsmv", kernel = "triangular",
                    h=h, ci=95, x.label = "1960 Poverty Rate", y.label = "",
                    title = title)

  estimate_a <- rdrobust(data[[mort_col]],
                         data$povrate60, c=threshold, p = 1,
                         kernel = "triangular", h=h)

  return(list(panel_a = panel_a, estimate_a = summary(estimate_a)))
}
```

```
helper(data, "Panel A: Children 5-9 Head Start susceptible causes",
                      "mort_age59_related_postHS")
```



```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                 2783
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 2489          294
## Eff. Number of Obs.             234          180
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   6.811        6.811
## BW bias (b)                   6.811        6.811
## rho (h/b)                     1.000        1.000
## Unique Obs.                    2489          294
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    -2.409     1.206    -1.998     0.046    [-4.772 , -0.046]
##         Robust         -         -    -2.760     0.006    [-6.412 , -1.087]
## =============================================================================
## $panel_a
## Call: rdplot
##
## Number of Obs.                  414
## Kernel                   Triangular
##
## Number of Obs.                  234             180
## Eff. Number of Obs.             234             180
## Order poly. fit (p)               1               1
## BW poly. fit (h)              6.811           6.811
## Number of bins scale          1.000           1.000
##
##
## $estimate_a
## NULL
```

```
helper(data, "Panel B: Children 5-9 Injuries",
                      "mort_age59_injury_postHS")
```

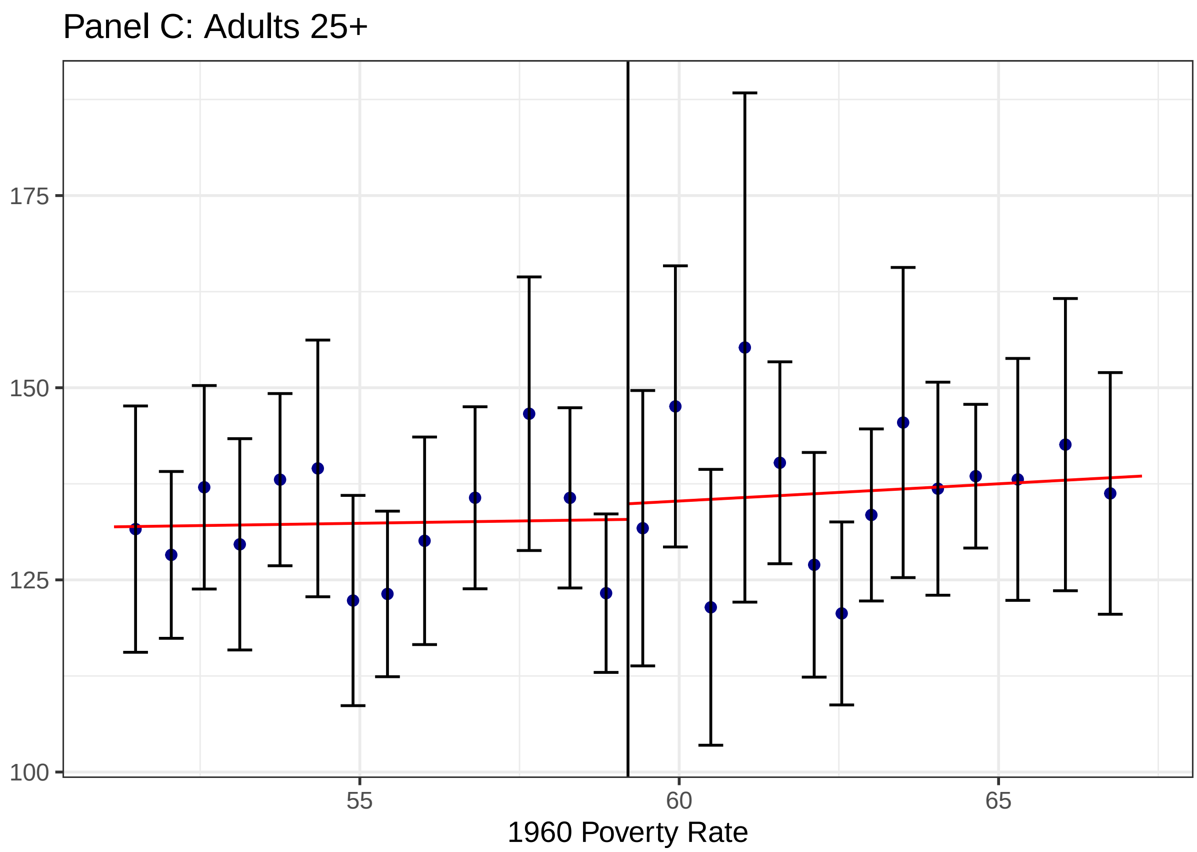


```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                 2783
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
```

```
##
## Number of Obs.                 2489          294
## Eff. Number of Obs.             211          169
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   6.262        6.262
## BW bias (b)                   6.262        6.262
## rho (h/b)                     1.000        1.000
## Unique Obs.                    2489          294
##
## ==============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ==============================================================================
##   Conventional     1.133     3.775      0.300     0.764    [-6.265 , 8.531]
##         Robust         -         -      0.134     0.894    [-8.949 , 10.260]
## ==============================================================================
```

```
## $panel_a
## Call: rdplot
##
## Number of Obs.                  380
## Kernel                   Triangular
##
## Number of Obs.                  211              169
## Eff. Number of Obs.             211              169
## Order poly. fit (p)               1                1
## BW poly. fit (h)              6.262            6.262
## Number of bins scale          1.000            1.000
##
##
## $estimate_a
## NULL
```

```
helper(data, "Panel C: Adults 25+",
                      "mort_age25plus_related_postHS")
```

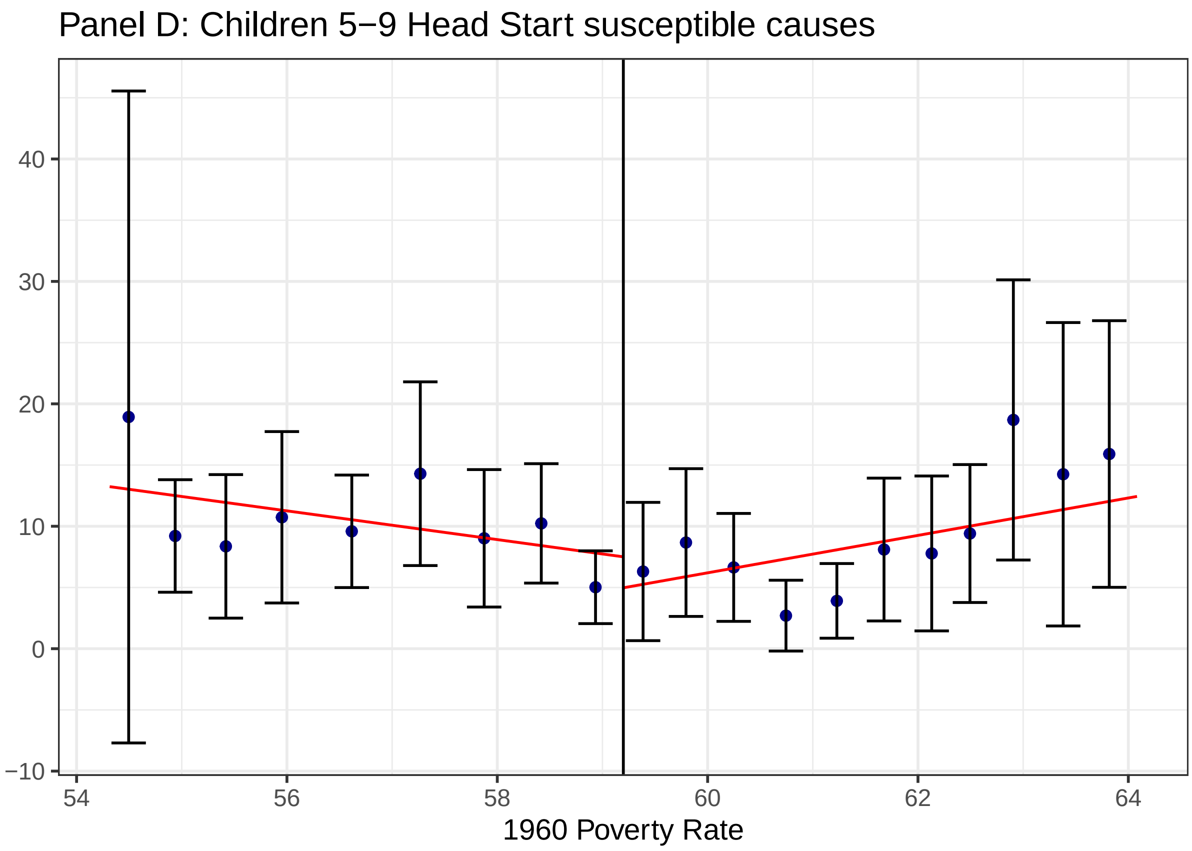


```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                 2783
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 2489          294
## Eff. Number of Obs.             281          203
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                    8.047        8.047
## BW bias (b)                    8.047        8.047
## rho (h/b)                      1.000        1.000
## Unique Obs.                    2489          294
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional     2.033     5.972     0.341     0.733    [-9.671 , 13.738]
##         Robust         -         -     0.547     0.585   [-11.110 , 19.703]
## =============================================================================
```

```
## $panel_a
## Call: rdplot
```

```
##
## Number of Obs.                     484
## Kernel                      Triangular
##
## Number of Obs.                     281             203
## Eff. Number of Obs.                281             203
## Order poly. fit (p)                  1               1
## BW poly. fit (h)                 8.047           8.047
## Number of bins scale             1.000           1.000
##
##
## $estimate_a
## NULL
```

```
helper(data, "Panel D: Children 5-9 Head Start susceptible causes",
                      "mort_age59_related_preHS")
```



```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                    2804
## BW type                         Manual
## Kernel                      Triangular
## VCE method                          NN
##
## Number of Obs.                    2504          300
## Eff. Number of Obs.             164          138
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   4.884        4.884
## BW bias (b)                   4.884        4.884
## rho (h/b)                     1.000        1.000
## Unique Obs.                    2504          300
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    -2.533     2.205    -1.149     0.251    [-6.854 , 1.788]
##         Robust         -         -     1.288     0.198    [-2.046 , 9.885]
## =============================================================================
```

```
## $panel_a
## Call: rdplot
##
## Number of Obs.                  302
## Kernel                   Triangular
##
## Number of Obs.                  164             138
## Eff. Number of Obs.             164             138
## Order poly. fit (p)               1               1
## BW poly. fit (h)              4.884           4.884
## Number of bins scale          1.000           1.000
##
##
## $estimate_a
## NULL
```

Methodological choices:

- A triangular kernel (versus a rectangular kernel) weights values near the threshold more highly than those farther away to reduce the risk of boundary bias. In practice, however, we may want to verify that our results are not sensitive to the kernel.
- A data-driven approach to bandwidth selection reduces the MSE and optimizes the bias-variance tradeoff; that is, it balances the precision of larger bandwidths (i.e. variance) and the risk of specification error of narrower ones (i.e. bias).
- The number of bins only matters for the visual presentation of the discontinuity; however, quantile-spaced bins ensure that each bin has an equal N, which makes it easy to visually compare the bins since they are equally precisely estimated.

4. 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.

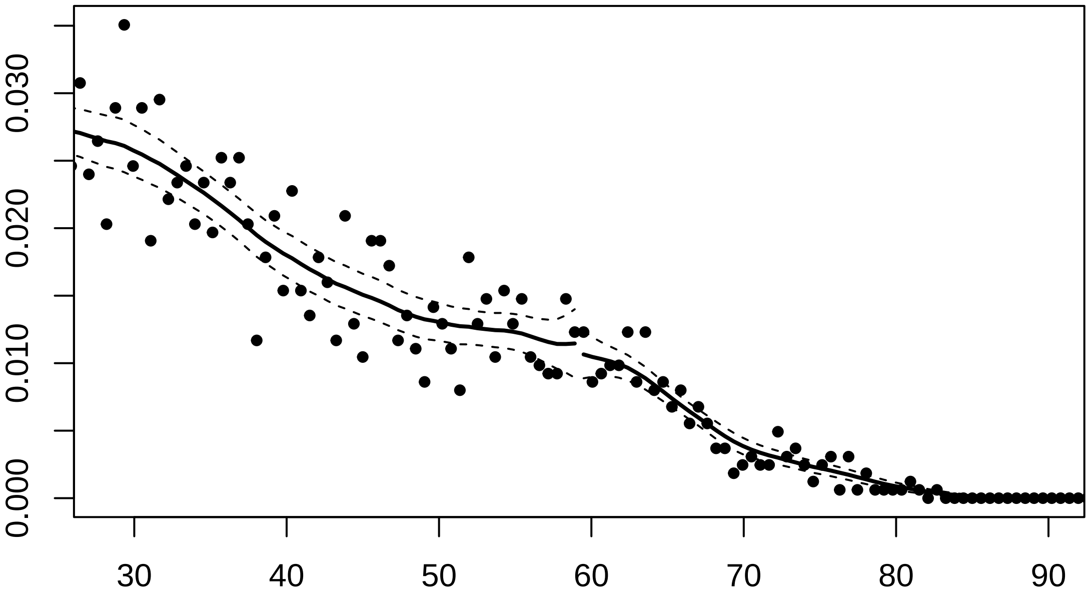
```
library(rdd)
```

```
## Loading required package: Formula
```

```
library(rddensity)
bandwidth <- rdbwselect(data$mort_age59_related_postHS,
                        data$povrate60, c=threshold, p = 1,
                        kernel = "triangular", bwselect = "mserd") #MSE-optimal bandwidth
summary(bandwidth)
```

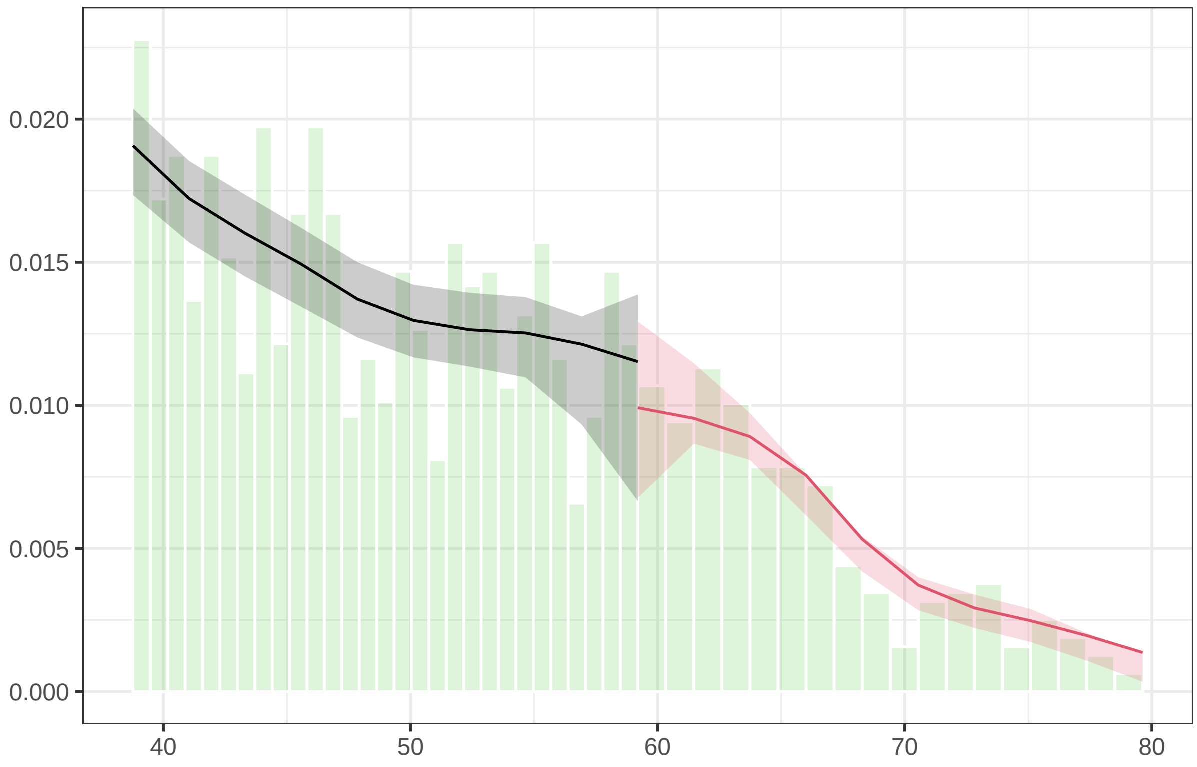
```
## Call: rdbwselect
## 
## Number of Obs.                 2783
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
## 
## Number of Obs.                 2489          294
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## Unique Obs.                    2488          294
## 
## =======================================================
##                   BW est. (h)    BW bias (b)
##             Left of c Right of c  Left of c Right of c
## =======================================================
##      mserd     6.811      6.811     10.726     10.726
## =======================================================
```

```
h <- bandwidth$bws[1]

#McCrary test
DCdensity(data$povrate60, threshold, bw = h, htest =TRUE)
```



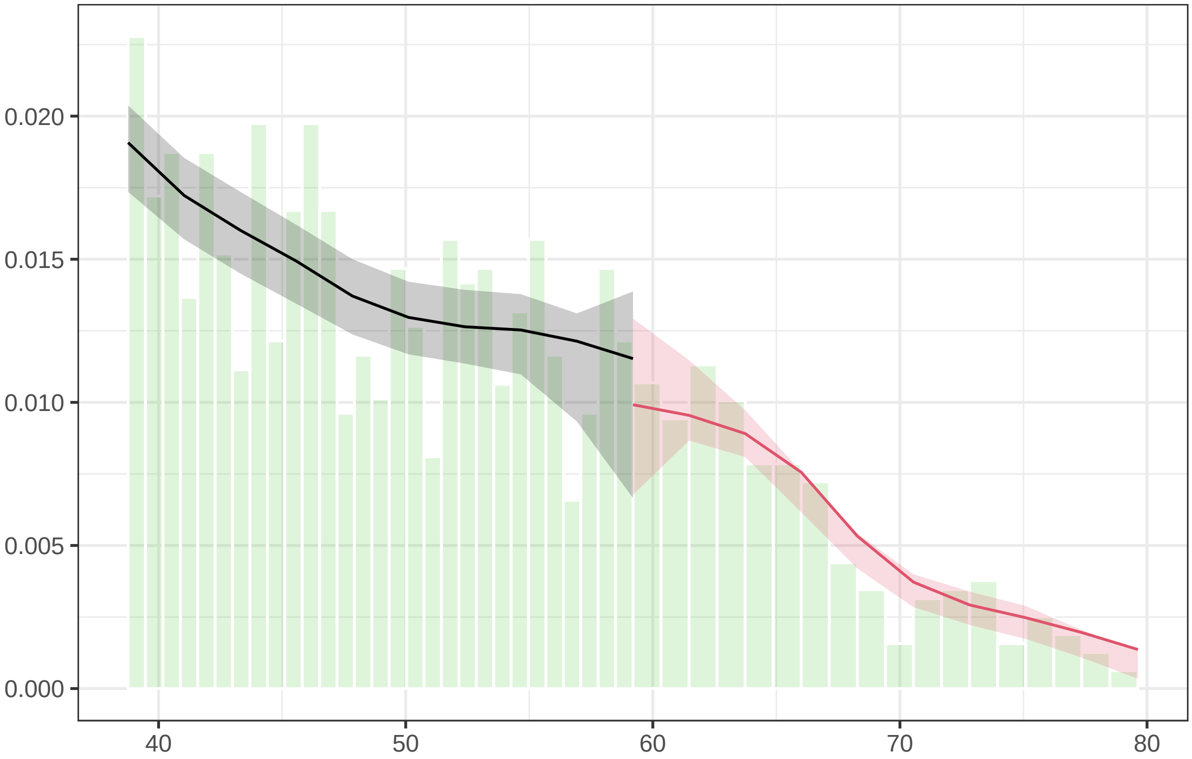
```
##
##  McCrary (2008) sorting test
##
## data:
## z = -0.32633, binwidth = 0.57976, bandwidth = 6.81072, cutpoint =
## 59.19840, p-value = 0.7442
## alternative hypothesis: no apparent sorting
```

```
#Manipulation test
rdd<-rddensity(data$povrate60,c = threshold, p =1, kernel = "triangular", h = h)
rdplotdensity(rdd = rdd, data$povrate60)
```



```
## $Estl
## Call: lpdensity
##
## Sample size                                      2504
## Polynomial order for point estimation    (p=)    1
## Order of derivative estimated            (v=)    1
## Polynomial order for confidence interval (q=)    2
## Kernel function                                  triangular
## Scaling factor                                   0.892971815911523
## Bandwidth method                                 user provided
##
## Use summary(...) to show estimates.
##
## $Estr
## Call: lpdensity
##
## Sample size                                      300
## Polynomial order for point estimation    (p=)    1
## Order of derivative estimated            (v=)    1
## Polynomial order for confidence interval (q=)    2
## Kernel function                                  triangular
## Scaling factor                                   0.106671423474848
## Bandwidth method                                 user provided
##
## Use summary(...) to show estimates.
##
```

```
## $Estplot
```



There doesn't appear to be bunching near the threshold.

Overall, these two tests suggest minimal manipulation in the running variable.

22. 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.

Question: What is the effect of a warning letter on opioid prescribing for physicians future opioid prescribing patterns.

I think that the main variables of interest would be patient level characteristics around their condition and provider history of opioid prescribing.

We could specify an RD where providers who are prescribing at a rate that exceeds some threshold value receive a letter. The outcome is future rate of opioid prescriptions.

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

$$Y_i = \alpha + \beta_1 T_i + \beta_2 X_i + \gamma Z_i + \epsilon_i \tag{1}$$

where Y_i is the outcome variable (future rate of opioid prescriptions), T_i is a binary variable indicating whether the physician received a warning letter or not, X_i is a vector of patient-level characteristics and physician history of opioid prescribing, Z_i is a vector of control variables, and α, β_1, β_2, γ, and ϵ_i are the intercept, treatment effect, coefficient vector for X_i, coefficient vector for Z_i, and error term, respectively.

In this RD design, physicians who are prescribing at a rate that exceeds some threshold value are assigned to the treatment group (those who receive the warning letter) and those who prescribe at a rate below the threshold are assigned to the control group (those who do not receive the warning letter). The treatment effect is then estimated by comparing the future rate of opioid prescriptions between the treatment and control groups, while controlling for patient-level characteristics, physician history of opioid prescribing, and other relevant control variables.

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

A potential threat to the validity of the RD design is the manipulation of the threshold. This could occur if individuals or institutions are able to strategically manipulate their scores to be just above or below the threshold in order to receive or avoid treatment. If this occurs, it violates the assumption of randomization around the threshold and undermines the causal interpretation of the RD design.