Problem Set 7

Due on October 31, 2023, 10am

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

1. Download the contents of the assignment Dropbox link in a folder (directory) on your computer designated for this problem set only. Follow the directory structure, e.g., putting the `data` folder containing the datasets as a subdirectory of the project directory.

2. Insert your answers in the yellow boxes using Microsoft Word, and prepare a single `.R` script for what you produce. Save the word document as a `.PDF`.

3. Please submit the PDF to the designated `PS-XX: pdf` link and your R Script to the `PS-XX: R` link.

(1) Your name:

Nikhilla Bhuvana Sundar

(2) Group members, if any:

(3) Compliance with the Academic Code on problem set[1] (sign with an X below)

X

[1] You may use the same code from classmates, Ed Discussion Board, instructors, and generative AI. However, you must hand in your own unique written work and code in all cases. Any wholesale 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, or use generative AI to provide potential code for you, but you must each type your own answers and your own code.

This problem set continues with the following article:

> Alix-Garcia, Jennifer M., Katharine RE Sims, Victor H. Orozco-Olvera, Laura E. Costica, Jorge David Fernández Medina, and Sofía Romo Monroy. "Payments for environmental services supported social capital while increasing land management." *Proceedings of the National Academy of Sciences* (2018).

W1e will discuss this article on Wednesday, October 25th. There is no pre-class exercise, but please come having read the results and methods of the article. Finally, *if you have relevant experience in the sort of environmental programs discussed in this article*, please email me asap before Wednesday class so I can see if I can incorporate your experience in the class.

To review basic RD implementation, you might also want to look at the screencasts below.

- Part 1, Visualizing RD: https://vimeo.com/400826628
- Part 2, Regression RD: https://vimeo.com/400826660

Code style tip: Add comments to make the script easier to organize. Per the style guide, each comment starts with one or multiple `#` marks and there should be a space after each mark. e.g., `# Problem 1 -----` instead of `#Problem 1`. Adding more than four or more dashes (`----`) makes RStudio recognize the comment as a header and makes them available in the "Outline View".

Gradescope will have the following packages installed:

```
library(tidyverse)
library(haven)
library(fixest)
library(modelsummary)
library(patchwork)
library(rdd)
library(broom)

library(glue)
library(gt)
```

Problem 1

The data for this paper is a cross-sectional dataset of 862 localities in 12 Mexican states. This is named `data/ mexico_localities.dta` and includes the variables:

Variable name	Description
id	Unique identifier for the applicant locality
cohort	One of two cohort in which the localities entered. See paper's appendix for discussion of municipalities that entered twice.
state_id	A numeric code identifying the state,
points	Applicant score determining eligibility, i.e. the running/forcing variable. Score is already re-centered so that 0 indicates the state's cutoff point. States above or at the cutoff were offered the PES. Note that in this paper, points of *exactly 0* **were** admitted to the PES.
bin	Binned points, relative to the cutoff, recentered so that 0 indicates the smallest set of scores that were admitted
bin_raw	Binned version of points before centering.
uptake	Whether or not the locality accepted the PES treatment
mng_index_raw	Management index (outcome). Not re-centered to state mean
lnIMLoc2010_EJawm	log(poverty index in 2010)
lndist_any_road	log(km to road)
lndist_maj_city	log(km to major city)
mean_canopy	log(Mean canopy cover 2000)

A full codebook of variable descriptions is attached in the Dropbox.

Problem 1

We will first visualize the data. You will want to create a binary variable that indicates whether the locality made the cutoff and was offered the PES.

1.1

Make two well-labeled RD plots with the running variable on the x-axis, shown side by side:

1. In one, the y-axis should be the log poverty index measured in 2010 (Fig. S2).
2. In the other, the y-axis should be the management index **re-centered** to its state average (a simpler version of Fig. 2A)

Each RD plot should show the all the data points as a scatter plot, and also show a regression fit on top of it. There should be one line to the left and one to the right of the cutoff. These lines should be a polynomial of degree 1, i.e., a regular OLS with no squared terms. To avoid clutter, we recommend making the individual points small.

Finally, add a brief sentence or two on what the reader should take away from both figures.

The treatment effect can be visually understood as the difference or the distance between the lines at the cut off point.
Based on this from the figure we can interpret that PES (treatment) seems to have a greater impact on land management index as compared to poverty index.

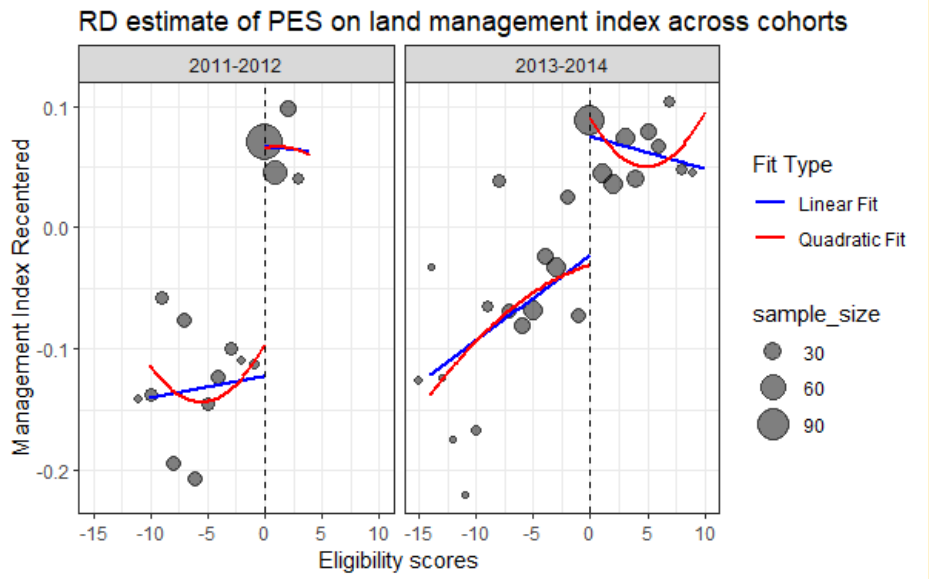
1.2

We will now turn to replicating the main finding, similar to Fig. 2A of the paper. Your figure should look similar to the figure below, and satisfy the following:

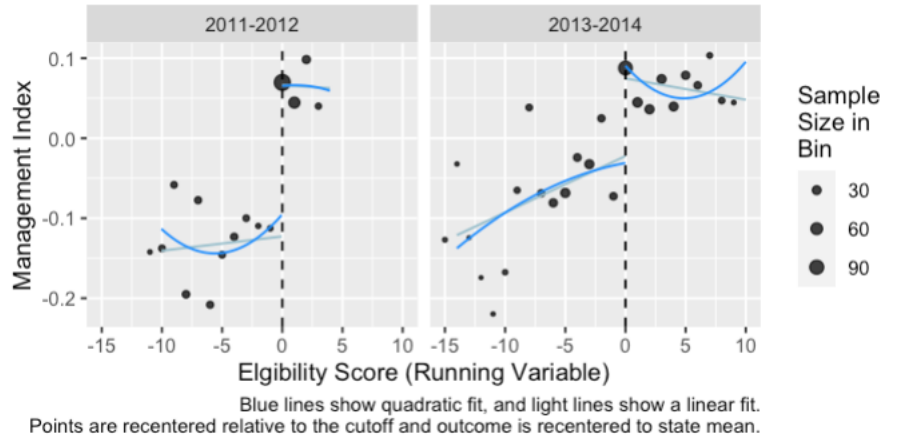
- Facet the data by the two cohorts using `facet_wrap()`,
- Use the re-centered management index outcomes as in 1.1
- Show a quadratic specification (the preferred method in the published paper) and another specification of your choice. These two specifications should be in different colors, with a legend or caption identifying each. Similar to the figure you made in 1.1, there should be separate lines below and above the cutoff.
- Show the sample average of the (re-centered) outcome in each bin (`bin`) as points, scaled by the sample size as a ggplot aesthetic. Show these points instead of showing the proportion of participation as in the published figure.

In other words, the regression lines should be fit from the locality-level data, but the scatter plots should represent bin-level averages. This type of plot is called a binned scatter plot and is common in RD.

Finally, add a brief sentence or two what the reader should take away from the figure.



- PES seems to have a greater impact on land management index in 2011- 12 than 2013-14
- A quadratic fit specification seems to better fit the data and capture the nonlinear relationship between scores and management index better than linear fit.
- Binned scatterplot reveals this nonlinear pattern without assuming a specific relationship form. It helps us look at average outcomes over eligibility score ranges scaled by sample size.



1.3

Recall that the regression that represents the linear equation in the figure you showed was of the form

$$Y_i = \alpha + \beta D_i + \gamma X_i + \delta X_i D_i + \varepsilon_i$$

where Y is the outcome, D is the eligibility, and X is the running variable.

Consider what we are assuming *if* we did not include the term $X_i D_i$ in the regression. That is, what we are implying about the structure of the potential outcomes so that the coefficient on D_i estimates the valid treatment effect? In your answer, you should mention the condition for a RD estimate to be valid.

When we do not include the term $X_i D_i$ in the regression, we are assuming a constant slope on either side of the cut off point.

For the estimates to represent a valid treatment effect, by this assumption we are implying that the potential outcomes around the cut off scores (which is where we assume continuity of potential outcomes for a RD estimate to be valid) also trend by the same slope.

Problem 2

A local linear regression is more flexible than fitting a single function in the entire half of the data. In this problem, we will document how the results are sensitive to the bandwidth of that local linear function.

In the subsequent, we will only use the data from the 2013-2014 cohort. You can drop the earlier cohort from your analysis. We will also only focus on one outcome: the uncentered management index (`mng_index_raw`).

2.1

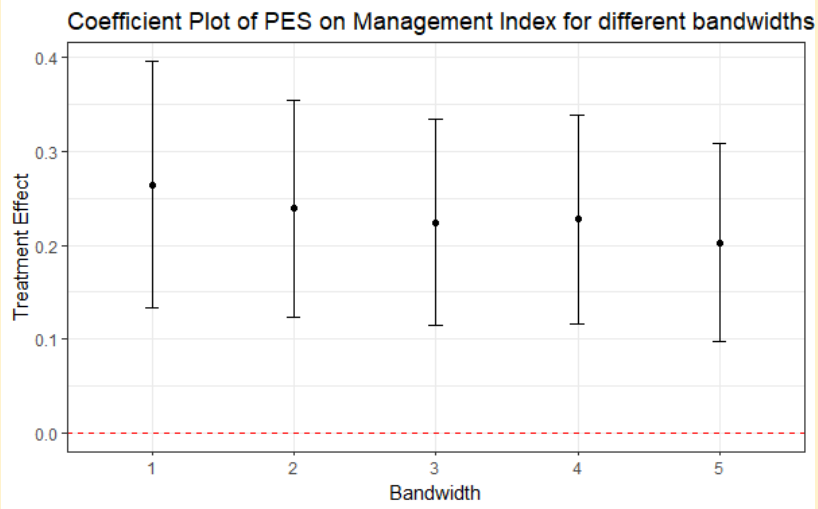
Estimate the treatment effects by considering bandwidths of $X \in [-h, h]$ (inclusive) for $h = \{1,2,3,4,5\}$. Within this local specified by the bandwidth, estimate the RD effect by a quadratic regression (polynomial degree 2) with varying slopes on both sides. In this subquestion, do not use a kernel; i.e. simply weigh all points within the bandwidth equally.

Produce a **regression table** with each column representing a different bandwidth, and a **coefficient plot** with the same coefficients with their 95 percent confidence intervals.

- The regression table should be simplified so only the coefficient and standard error of the *treatment variable of interest* is displayed. In the bottom set of rows, show only the sample size.
- Make sure the coefficient plot is well-labeled. Include a reference line for an estimate of 0, so that it is easy to tell whether the confidence interval crosses 0.

Finally, add a brief sentence what the reader should take away from the patterns of the coefficient estimate and *standard errors* from the figure and table.

	1	2	3	4	5
PES_offeredTRUE	0.265	0.239	0.224	0.228	0.203
	(0.066)	(0.058)	(0.055)	(0.055)	(0.053)
Num.Obs.	103	156	219	271	333



We can observe that as we increase or widen bandwidth, both coefficient on treatment variable and standard errors tend to decrease (although not by a great amount). From our understanding of bias-variance trade off we know that as we widen bandwidth, we have low variance and higher bias and vice versa for smaller bandwidth

2.2

We now move onto using the `rdd` package on CRAN created by Drew Dimmery originally in 2012. This package defaults to estimating the curves by a local linear regression with a kernel within a bandwidth, as suggested by Imbens and Kalyanaraman (IK, 2012).

IK also recommend picking the bandwidth with a procedure that they argue minimizes the Mean Square Error of applying a local linear regression. They also recommended a triangular kernel to downweight points far away from the cutpoint, whereas our estimates in 2.1 used a rectangular kernel that weights the points within the bandwidth equally.

In this problem, follow the package instructions to estimate the RD estimate on the uncentered land management index with local linear regression **using the IK optimal bandwidth selection procedure**. Estimate two specifications: the triangular and rectangular kernel. Report the bandwidth and its associated coefficient estimate (with standard error) in bullet points.

- Triangular kernel
  - Bandwidth: 1.7896
  - Coefficient estimate: 0.2389
  - Standard error: 0.04916

- Rectangular kernel
  - Bandwidth: 2.813
  - Coefficient estimate: 0.2102
  - Standard error: 0.04732

2.3

Interpret the first RD estimate (triangular kernel) in **2.2 substantively**, attempting to put the magnitude of the size into context by standard deviation and percentage terms. In addition to the usual requirement for describing the units of the outcome, please specify:

- Whether the effect is statistically distinguishable from 0 given its standard error
- The magnitude of the effect in terms of the standard deviations of the outcome among all 218 localities that were not eligible for the program
- The magnitude of the effect in percent terms, relative to the mean outcome among all 218 localities that were not eligible for the program

There is 23.89% points increase in land management index for groups that were accepted into PES program when compared to groups that were rejected.

In addition,

- The effect is statistically distinguishable from 0 as Z score is 4.859 which is greater than critical value of 1.96.
- The treatment effect is 1.132 standard deviations away from the mean of the control group (Standard deviation of control group is 0.211 and since the estimate of 0.2389 is already the difference means, the RD estimate is divided by standard deviation)
- There is a ~82% increase in effect of treatment relative to the mean outcome of those who were not eligible for the program

Problem 3

The paper reports intent to treatment effects (ITT), even though some localities that were eligible did not end up taking up the PES program. Conduct a Fuzzy RD estimation and **report the effect of *taking up the PES*** instead of the ITT. As noted in the variables table, the indicator for taking up the PES is `uptake`.

For this problem, use a linear regression on the full sample of the 2013-2014 cohort (i.e., a similar specification as Problem 2.1 but not setting a bandwidth, and not using squared terms).

Finally, add a brief sentence or two on what the reader should take away from the results.

There is a 0.092 unit increase in management index as a result of uptake of treatment or among compliers

Appendix

The screencast noted at the top of the document provides a basic tutorial of the RD visualization and regression.

1.1

- Use `geom_smooth()` `stat_smooth()` to add a regression fit. Use the `group` aesthetic to allow the lines to vary below and above the cutoff. You can remove the SEs.
- To reduce the visual clutter of points, consider fixing their `size` to be a smaller number or making them translucent with `alpha`.
- You can use the `patchwork` package to combine two ggplot figures. Once the package is loaded, two ggplot objects can be loaded simply with a `+` symbol (https://patchwork.data-imaginist.com/).
- To recenter the state mean, consider using `mutate()` combined with `group_by()`. You can also create a tibble of means with `summarize()` and then join those means to the entire dataset. The former is slightly more compact.
  - In recent versions of `dplyr`, there is a slightly more compact way of implementing `group_by()`. You can add a `.by =` argument in other dplyr verbs like `summarize`, `mutate`, or even `filter` and it will conduct the operation group by group. This is the same as adding a `group_by()` function in the line before, but it is sometimes faster and "ungroups" the groups immediately after. See https://dplyr.tidyverse.org/reference/dplyr_by.html.

1.2

- See the `geom_smooth()` function help page to control the specification of the regression fit. This problem asks you to construct polynomials of degree 1 and 2. You can either specify the generic formula as in the screencasts, or use the poly(x, 2) shorthand as shown in the help page.

- The local polynomial used in the published figure can be approximated with the LOESS specification in ggplot (`method = "loess"`). If you are using a LOESS plot, you should try a few alternative values of the `span` parameter in `geom_smooth`, which controls the bandwidth that LOESS smoothes over.

- Technically, Stata's local polynomial regression is a kernel regression with an Epanechnikov kernel, which is not an option in the standard `stat_smooth()` layer. You can still overlay a kernel regression fit onto a ggplot after estimating it with a different pacakge, but this is somewhat tedious.

- The smoothed curve and binned points come from different datasets: one from the original dataset and another from a grouped average. Layers using a different dataset can be added to a ggplot, by specifying the `data` argument with a different dataset (this argument is often left unspecified to inherit the original dataset specified in `ggplot()`).

- You can control the size of the points with the `scale_size_area()` layer just like the `scale_color_manual()` layer. `max_size` controls the size of the largest point.
- `labs()` specifies the axis labels, captions, and legend labels.

2.1

- You can rely on the idea in PS-04 to loop across values with `map()` or with a for loop. `set_names()` will assign the values of a vector as its names. This feature is useful in this problem set when the values of h can be the names of your items.
- You can rely on the technique in the solutions of PS-06 to present a coefficient plot with confidence intervals.

- The screencasts (which are from 2020) use `stargazer` as a regression table generator, but in the year 2023 I would recommending switching to `modelsummary`

- To enter square terms of a variable `X` in a regression formula, you must wrap the squared term with an identity function `I`, as in `y ~ X + I(X^2)`. It may be easier to create a single variable that is the square of the variable in the dataset beforehand.

2.2

- Note that formula specification in the `rdd` package is slightly different and more simple than your previous hand-coded versions of linear regression. You will need to read the package help page.
- This package was created in 2012 so it is more minimal in its documentation, but it still works well. Getting used to working with a variety of new packages is an useful programming skill. The R code to load the table of contents for a package that does not have a website is through `help(package = rdd)` in this case. This is identical content to the PDF reference manual found in https://cran.r-project.org/web/packages/rdd/index.html

3

Fuzzy RD is a instrumental variables regression where uptake is instrumented by the offer of treatment. Even the offer of treatment appears in two different parts of the regression (the non-interacted treatment and the interacted treatment), we only instrument for the coefficient of interest.