Problem Set 2

Regression discontinuity

August 2023

General objectives and description

The goal of this problem set is to practice common steps in regression discontinuity analysis and to deepen your understanding of it.

1 Hypothetical

Consider the following scenario, similar to Suri et al. (2021): Customers, indexed by i, are assigned a credit score, s_i, by the only active bank in the region. Scores range from -10 to +10. Individuals with a positive credit score ($s_i > 0$) are granted loans by that bank if they apply for one. Other households cannot get a loan from a bank. To estimate the effect of loans on entrepreneurship, data on expenditures (in USD) for productive assets, y_i, is collected in a cross-sectional survey, alongside the credit score and information on demographics, for a representative sample of bank customers.

i) Researchers estimate a regression equation:

$$\log y_i = \alpha + \beta \mathbb{1}_{(s_i>0)} + \gamma_1 s_i + \gamma_2 s_i \cdot \mathbb{1}_{(s_i>0)} + \varepsilon_i,$$

where $\mathbb{1}_{(s_i>0)}$ is a dummy variable indicating if $s_i > 0$. The resulting estimates are:

Table 1: Regression estimates

	(1) $\log y$
score, s	-0.2 ***
	(0.05)
score above zero, $\mathbb{1}_{(s_i>0)}$	1.9 *
	(1.0)
interaction, $s \cdot \mathbb{1}_{(s_i>0)}$	0.2 **
	(0.09)
constant	2.9 ***
	(0.6)
N	321

Standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

a) Draw a stylized graph that has the running variable (s_i) on the x-axis, the outcome on the y-axis, and illustrates the treatment effect implied by the regression results above. Highlight all parameter estimates in the graph.

b) What treatment effect is estimated by this regression? Make a concise statement about the interpretation and magnitude of the effect.

c) Assume you run the same regression but use age_i as dependent variable. You find that the estimate for β is statistically highly significant and negative. Why could that pose a problem for causal identification?

d) Your colleague suspects a bank officer manually increased scores of young people who were just below the threshold to ensure they became eligible for loans. Suggest a method to plot the data that would allow you to graphically examine if that might be the case. Briefly explain ($<$ 100 words) how this method would be informative about your colleague's hypothesis and state whether it would constitute a problem for analysis if the hypothesis is true.

2 Applied

Imagine a scenario where a vaccine against shingles is recommended only for people aged 80 (not younger, not older). Only people who are 80 are eligible. The vaccine was introduced on the 1st of September 2010. Hence, everyone born before the 1st of September 1930 is too old and will never become eligible. Everyone else becomes eligible on their 80th birthday and stays eligible for one year. You can take it as given that most people choose to take the vaccine when eligible.

Our goal is to study if the vaccine reduced the incidence of shingles. Note that the data is entirely made up and does not represent actual health data or actual treatment effects. The dataset is a cross-section of patients and has three relevant variables:

- `dob` (date of birth) indicates how many weeks away from the 1st of September 1930 a person's birthday was. Negative values indicate earlier birthdays (i.e., older people who were eligible). Positive values indicate later birthdays, so those people were eligible for the vaccine for one year following their 80th birthday.
- `shingles` is a binary indicator for whether or not the person was diagnosed with shingles at least once between September 2010 and September 2020.
- `vaccine` is a binary indicator for whether or not the person has taken the vaccine against shingles.

Descriptives

i) Describe the distribution of variables in the dataset.

a) What range of birthdays is present in the data? What does this imply for the age of people in the sample in 2010?

b) What fraction of people have had shingles between 2010 and 2020? What is this fraction for people with the with `dob` less than -50? What is this fraction for people with the with `dob` over 50?

Visualization and analyses

ii) Use the command `rdplot` from the package `rdrobust` (you may have to install it) to draw a plot that illustrates the treatment effect of eligibility on having received the shingles vaccine. Interpret your results.

iii) Use the command `rdplot` to draw a plot illustrating the treatment effect of eligibility on having had shingles. Create it once using a linear polynomial and once using a quadratic polynomial. Interpret your results.

iv) Use the command `rdrobust` to estimate the effect of being eligible for the shingles vaccine on getting shingles.

v) Estimate the regression

$$shingles = \alpha + \beta \mathbb{1}_{(wob_i \geq 0)} + \gamma_1 wob_i + \gamma_2 wob_i \cdot \mathbb{1}_{(wob_i \geq 0)} + \varepsilon_i.$$

Note: you might have to first create a dummy variable for $wob_i \geq 0$. The estimate for β estimates the treatment effect of eligiblity on shingles. Compare it against the estimate you got using rdrobust. What are the reasons for the difference between the two estimates?

Bonus: Fuzzy RD

vi) Use the command `rdrobust` to implement a fuzzy RD analysis to estimate the effect of taking the vaccine (as opposed to the effect of being eligible for the vaccine) on getting shingles. (Consult the documentation of the `rdrobust` package to find out how to implement a fuzzy RD).

Hint

Potentially useful Stata commands: `rdrobust`, `rdplot` `ssc install rdrobust` Potentially useful R packages and commands: `rdrobust`, `rdplot`