MY457: Problem Set 4 - Regression Discontinuity Designs

Wed/19/Mar

This problem set is formative and will not contribute to your final grade. However, it is designed to build and deepen your conceptual understanding of the material and to practice applying the material in R. Using tools and resources such as ChatGPT and Stack Overflow is allowed and encouraged. Bear in mind that you are ultimately responsible for the work you submit, and that these tools often provide incorrect solutions. Make sure that however you use these tools it aligns with your best interests, and enhances your learning in this class.

This problem set must be submitted on Moodle by 5pm on Thu/27/Mar. You must also use the provided `.Rmd` template to produce a .pdf with your answers. If your submission is late, is not a .pdf, or is not appropriately formatted, you will not receive feedback on your work.

1 Concepts

This question reviews some of the concepts covered in class. Mathematical notation can be a useful tool to explain concepts, but it's important that you understand and can explain the concepts clearly and concisely. If you want to support your explanations with mathematical notation, this page provides a tutorial on including mathematical notation in Rmarkdown.

Consider a setting with $i \in [1, ..., N]$ units. Treatment D_i is assigned based on values of a variable X_i, such that $D = \mathbf{1}[X_i > c]$. We are interested in the effect of D_i on Y_i.

1.1 What is the key identifying assumption in this setting, from the local randomization perspective?

1.2 What is the key assumption from the continuity perspective?

1.3 Explain the difference between sharp regression discontinuity (RD) designs and fuzzy RD designs. Make sure you explain the difference in the design, assumptions, and estimands targeted by these two approaches.

1.4 Explain the connection between fuzzy RD and instrumental variables.

2 Simulations

In this question we will use simulated data to test some of our intuitions about instrumental variables. The advantage of using a simulated dataset is that we have explicit control over the data generating process, and know the 'true' answer to any question we pose.

2.1. Explain the code below.

```r
n <- 1000

set.seed(123)
X <- rnorm(n, mean = 0, sd = 15)
c <- 0

Y0 <- 100 + 0.45*X - 0.15*(X^2) + 0.001*(X^3) + rnorm(n, mean = 0, sd = 25)
Y1 <- 175 + 0.25*X + 0.15*(X^2) - 0.001*(X^3) + rnorm(n, mean = 0, sd = 25)

D <- ifelse(X > c, 1, 0)
```

```r
Y <- ifelse(D == 1, Y1, Y0)

dat <- tibble(
  X,
  D,
  Y,
  Y0,
  Y1
)
```

2.2 In this simulation, what is the true ATE and, what is the true LATE (at the threshold)? What do you conclude?

2.3 In a clear, well formatted, and compelling figure, visualise *both* potential outcomes and how they relate to the cutpoint.

2.4 In a clear, well formatted, and compelling figure, visualise only the observed `Y`.

2.5 Use a global (do not focus on a narrow window) linear regression with common slopes to estimate the LATE. What do you find? How does this estimator perform?

2.6 Use a global (do not focus on a narrow window) polynomial regression with varying slopes to estimate the LATE. You may pick the polynomial order. What do you find? How does this estimator perform? How might you expect this estimator to perform outside of a simulated setting?

2.7 Let's now analyse our data with the `rdrobust` package. Estimate the LATE with local polynomial approximation. You may leave the settings at their defaults, or change settings if you so choose. What do you find?

2.8 Use the `rdplot` function within the package to create a regression discontinuity plot.

3 Replication

In this section, we will use real-world data to reinforce what we have learned. We will analyse the dataset employed in *Rural Roads and Local Economic Development*. Physical connectivity through paved roads may be a catapult for economic development. The authors of this paper study the economic impact of a $40 billion national programme designed to build roads across rural India.

3.1 Explain the authors' research design, and how they approach the identification strategy.

3.2 Read the data into `R` and using the `{rdrobust}` package, visualise any discontinuity in the variable `r2012` as a function of a cutoff or threshold in population (`v_pop`). What is the point of this analysis, and what do you find?

3.3 A common concern in RD designs using population thresholds is manipulation at the threshold (see Eggers et al.). Using the `{rddensity}` package, test for sorting in around the cutoff. What do you find?

3.4 Select any two outcomes of interest and estimate the Fuzzy RDD. For each dependent variable, generate a single regression discontinuity plot, and overlay on the plot the key statistical results (point estimate of interest and 95% confidence interval). Use a polynomial of order 1 (`p = 1`), a triangular kernel, and the MSE-optimal bandwidth, as is done in the paper. Explain carefully what the quantity of interest you are estimating is, and in what ways it is `local`. What do you find?