LPO 7870 Research Design and Data Analysis II, 2024
Lecture 10 Sample Exercises

Problem 1 – RD using simulated data (1)

1) Open the simulated dataset in Stata using the command below (or use the sample code in the accompanying do-file).

```
use https://github.com/spcorcor18/LPO-7870/raw/main/data/RDsim1.dta
```

This is a dataset with n=1,000 observations, where *x* is the running variable and *y* is the outcome. The nice thing about using simulated data is that the analyst *knows* how the data were generated. In this case, there is a "treatment" that occurs for all observations where *x*>=1. The treatment effect is known to be +0.5. We can use this known benchmark to see how close we can get when estimating an RD model.

2) Re-center the running variable at *x*=1. Call the new variable *xc*.

3) Create a scatterplot showing the relationship between *y* and *x* (or between *y* and *xc*, the re-centered version). Do you see evidence of a discontinuity?

4) As an alternative to the scatterplot in part (3), try using the user-written command called binscatter (ssc install binscatter). Note you can choose the number of bins used with the option nq(#). Do you see any evidence of a discontinuity?

5) There is another user-written command called rdplot (ssc install rdrobust). This command will provide a binned scatterplot and fitted lines on either side of the cutpoint. Use the option c(#) to specify the cutpoint.

6) Now fit four RD models using OLS (note the treatment variable *t* has already been created for you):

    a. Linear regression – same slope on either side of the cutpoint
    b. Linear regression – different slopes on either side of the cutpoint
    c. Quadratic regression – same slope on either side of the cutpoint
    d. Quadratic regression – different slopes on either side of the cutpoint

    Interpret the estimated coefficients from each (but especially b and d). Is the "discontinuity" in the outcome statistically significant?

    The aim of RD is to accurately estimate the "jump" (if any) at the cutpoint. How well did the above regressions do?

7) Sometimes analysts can do better by narrowing in on a smaller bandwidth around the cutpoint. One way to do this is just to choose a bandwidth. Repeat (6d) above, but only use observations that are within +/- 0.5 of the cutpoint.

8) Finally, researchers have developed an "optimal" bandwidth that trades off bias (where a narrower bandwidth is preferred) and precision (where a wider bandwidth is preferred, using more of the data). This can be implemented using the rdrobust command as shown below:

   rdrobust y xc, c(0) p(2) bwselect(mserd) kernel(triangular)

Problem 2: RD using simulated data (2) – effect of participating in a G&T program

In this problem—also using simulated data—students are tested in grade 3 for eligibility to the gifted and talented (G&T) program in 4th grade. Students who score a 56 or higher on the 3rd grade test are eligible for G&T.

1) The data can be read using the command below (or using the sample code in the accompanying do-file):

   ```
   use https://github.com/spcorcor18/LPO-7870/raw/main/data/RDsim2.dta
   ```

2) Apply steps #2-6 and #8 from Problem 2 to this dataset. Note: be sure to use the correct cutpoint for this problem. Assume everyone complies with their treatment assignment (i.e., all eligible for G&T actually participate).

3) What assumptions are required in order to interpret the estimated discontinuity as the causal effect of participating in G&T?

4) Suppose that compliance is not perfect—that is, not everyone *offered* participation in the G&T program actually participates. There is another variable in this dataset called *GTpart* that =1 if the student actually participated. The other variable *grade4testfuz* is the grade 4 test score that reflects this imperfect compliance. What percent of students offered G&T actually took it up?

5) When compliance is imperfect, we assume that the observed *intent-to-treat effect* (the estimated discontinuity assuming perfect compliance is a weighted average of the *treatment-effect-for the treated* and zero (for those who didn't actually take up the treatment):

   $$ITT = p^* TOT + 0$$

   Which implies that $TOT = ITT/p$. (p is the proportion of those offered who "took up" the treatment—at the cutpoint). The rdrobust command can calculate this using the "fuzzy" option. Try this.

Problem 3: Regression interpretation – charter school effects on long-run earnings

Suppose you wish to examine whether attending a charter high school impacts adult earnings. You have data on 10,000 students that were randomly assigned (via a school lottery) to either attend a charter school or a regular public school in California, Oregon and Washington. Using that data you specify the following model:

$$lwage_i = \beta_0 + \beta_1 Charter_i + \beta_2 Score_i + \beta_3 Female_i + u_i$$

Where *lwage*$_i$ denotes the natural log of individual *i*'s wages, *Charter*$_i$ is a dummy variable that equals one if individual *i* attended a charter school and zero if they attended a regular public school, *Score*$_i$ is an individual's 8th grade math score (prior to attending high school) measured in points from 1 to 100, and *Female*$_i$ is a dummy variable equal to for females. Results are reported below with standard errors in parentheses.

$$\widehat{lwage_i} = \underset{(2.5)}{2} + \underset{(0.02)}{0.06}\, Charter_i + \underset{(0.003)}{0.01}\, Score_i - \underset{(0.03)}{0.09}\, Fe_i$$

1) Interpret the coefficient on: *Charter* and *Score*

2) Are these statistically significant? Please show all your work.

3) How would you test *jointly* that the coefficients of *Charter* and *Female* are equal to zero?

4) Create a graph that illustrates the relationship between the log of wages and 8th grade test scores for *men* that attended regular public schools and those that attended charter schools. Use a single diagram and label your diagram.

5) Suppose you wished to examine whether attending a charter school had a larger impact on the wages of women than the wages of men. Write out the model you would estimate and the hypothesis test you would conduct.

6) You run the following model:

$$\widehat{lwage_i} = \underset{(2.6)}{1.2} + \underset{(0.01)}{0.08}\, Charter_i + \underset{(0.004)}{0.02}\, Score_i - \underset{(0.03)}{0.04}\, Female_i + \underset{(0.02)}{0.12}\, Female_i * Charter_i$$

7) Please interpret the coefficient on *Charter, Female and Female*Charter*

.

Problem 4: Effects of the Kalamazoo Promise Scholarship using Difference-in-Differences

Bartik, T. J., Hershbein, B., & Lachowska, M. (2021). The Effects of the Kalamazoo Promise Scholarship on College Enrollment and Completion. *Journal of Human Resources, 56*(1), 269-310. Retrieved from http://jhr.uwpress.org/content/56/1/269.abstract

Bartik et al (2021) used a differences-in-differences research design to estimate the effects of the Kalamazoo Promise, a place-based college scholarship, on postsecondary education outcomes (e.g., college attendance and degree completion). This program was announced in November 2005 and pays up to 100% of tuition and fees at public postsecondary institutions for graduates of the Kalamazoo (Michigan) Public Schools. To be eligible for the scholarship, students must have been enrolled in KPS from at least 9th grade. Data come from the school district and the National Student Clearinghouse.

1) The first post-Promise graduating cohort was in 2006. Students who had been in KPS since at least 9th grade were eligible for the scholarship. Those who had not were ineligible. What would

be the disadvantage of comparing the post-secondary outcomes of eligible and ineligible students in the 2006 cohort, perhaps using a regression with control variables (a "cross-sectional" design)?

2) Bartik et al. have several pre-Promise graduating cohorts (2003-2005). Describe what an "interrupted time series" design would look like with these data, and the chief disadvantage of this approach.

3) The authors conduct several difference-in-differences analyses. The first one looks like this:

$$(1) \quad y_{ist} = \alpha + \delta_1 \widetilde{\text{Elig}}_{ist} + \delta_2 (\text{After} \times \widetilde{\text{Elig}})_{ist} + \gamma_{st} + \mathbf{x}_{ist}\boldsymbol{\beta} + u_{ist},$$

Where *Elig*=1 if the student met the eligibility criteria and *After=1* if the student was in a post-Promise cohort (2006+). The subscript refers to peson *i* in high school *s* in year *t.* Explain how to interpret the δ_1 and δ_2 coefficients and why this is a "difference-in-differences." How does this improve on the previous two approaches?

4) What assumptions are required in order to interpret the estimates from equation (1) as causal effects?

5) The authors write that the introduction of the Kalamazoo Promise scholarship was a "surprise" to local residents. Why is this useful, from a research design standpoint?

6) See Tables 3-6 for the main results—focus on the OLS column. What do they find?

7) One of the authors' robustness checks is a difference-in-differences analysis that adds cohort of graduates from non-KPS districts. Here, treatment is defined as graduating from KPS in 2006 or later. Write down how you think this regression is specified. See their Table 7 for the results.