Regression II
Vanderbilt University
Prof. Sean P. Corcoran November 11, 2023

Lecture 8 In-Class Examples

Example 1. This example is taken from *Mastering Metrics* chapter 4 and is based on Carpenter & Dobkin (2009). The analysis revisits the question of whether legal access to alcohol is associated with higher mortality rates. The dataset referenced below includes death rates by age, in which age (from 19-23) is divided into 50 equal-width cells. Individuals are considered "treated" when they reach 21 and can legally drink alcohol. The discontinuity in treatment is sharp, since *all* persons who reach this age are treated. If legal access to alcohol is associated with higher mortality, we would expect to see a discontinuity in death rates at 21. Read the file *AEJfigs.dta* into Stata using:

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/AEJfigs.dta
```

(a) Create a new age variable *age* that is centered at 21, and a "treatment" variable *over21* that equals 1 if $age > 0$ (and 0 otherwise).

(b) Create a scatterplot of mortality rates from all cases (*all*) against *age*. Is a discontinuity evident? Try a few RD plots using different bin methods (e.g., ES or QS, IMSE or MV) and different polynomial orders (e.g., quartic (the default), quadratic, linear). Note the RD plot does not have a lot of added value here, since the data are already binned.

(c) We will eventually use `rdrobust` for estimation. But for now use OLS regression to estimate the RD under different scenarios:

- Linear model, assuming the same slope on both sides
- Linear model, assuming different slopes on each side
- Quadratic model, assuming the same slope terms on both sides
- Quadratic model, assuming different slope terms on each side

How are the coefficients interpreted in these models?

Get predicted values and overlay these predicted values on the scatterplot. (Note: this is what `rdplot` can do for you, but this is for illustration). You can easily reproduce Figure 4.2 in *Mastering Metrics* (linear model, same slopes).

(d) Repeat the linear and quadratic models for two other measures of mortality: deaths by motor vehicle accidents and deaths by internal causes. One might expect the former to be most affected by access to alcohol. The latter could be considered a placebo test. You can easily reproduce Figure 4.5 in *Mastering Metrics* that combines these.

(e) Try part (c) using `rdrobust`. For comparability, use a uniform kernel and a bandwidth of 2. Compare the point estimates and standard errors to (c).

(f) Use `rdbwselect` to find the optimal MSE bandwidth under different choices of polynomial (linear or quadratic) and kernel (uniform or triangular). Use these in `robust` and compare your point estimates and standard errors. Normally, one could pass through the optimal bandwidth to `rdplot`, but this is problematic with these data (they are already binned).

Example 2. This example will generate data with a known discontinuity in y at a threshold level of x, and then estimate a RD model. It illustrates various RD commands and results, and simulates manipulation of the running variable. (Adapted from Dale Ballou).

(a) First produce simulated data using the syntax below. Notice that x is the running variable. What is the functional relationship between the outcome y and the running variable? What is the cut score? What is the treatment effect? Is this a sharp or fuzzy regression discontinuity? Create a new variable xc that is x centered at the cut score.

```
clear
set seed 1234
drawnorm x w e u, n(1000)
gen y = 3 + 3*x + 0.5*x^2 + w + u
gen t = (x >= 1)
replace y = y + 0.5*t
```

(b) Produce a scatterplot of y against xc. Do you see evidence of a discontinuity? Try using `binscatter` and `rdplot`, the latter implementing a quartic $p = 4$ and then quadratic $q = 2$. Do you see a discontinuity in these plots? Which shows it best?

(c) Estimate several (global, parameter) RD models using OLS, below. Do not use a kernel and use the full range of data. How close do these get to estimating the true treatment effect? Which model performs best, and why?

- Linear model, assuming the same slope on both sides
- Linear model, assuming different slopes on each side
- Quadratic model, assuming the same slope terms on both sides
- Quadratic model, assuming different slope terms on each side

(d) Obtain local, nonparametric RD estimates using `rdrobust`. Use a linear fit, the optimal MSE bandwidth selection, and triangular kernel. Pass through the optimal bandwidth to `rdplot` to get a local RD plot. Note: the accompanying do-file also shows how to use the older command `rd`. See the help menu for that command for options.

(e) Check for manipulation in the running variable xc in two ways: by inspection using `histogram`, and using `rddensity`. Do you expect manipulation here? What does the test conclude?

(f) Now modify the data a bit to introduce manipulation in x. Try the syntax below and explain in words what the first line is doing. (Create a new centered x variable xcm) Then, re-do the `rddensity` test. Does it detect the manipulation?

```
gen xm = x
replace xm = xm + .4 if xm < 1 & xm > .65 & e > 0
```

(g) Now that we know there is manipulation, try estimating the nonparametric RD using `rdrobust` (use same specs as part d). What does this yield?

Example 3. This example, based on an example created by Celeste Carruthers, also uses simulated data to estimate the effect of participation in a gifted and talented (G&T) program. This provides a simple illustration of both sharp and fuzzy RD.

(a) Generate 10,000 student observations. The data will include a measure of students' "true ability," $trueability \sim N(50, 4)$, and their 3rd grade test score, which is a noisy measure of their true ability $grade3test = trueability + u$ where $u \sim N(0, 1)$. To add a bit of realism, we will round test scores to the nearest 0.25 to create a discrete scale.

```
clear
set seed 195423
set obs 10000
gen id=_n
gen trueability = 50 + 4*rnormal()
gen grade3test  = trueability + rnormal()
replace grade3test = round(grade3test, 0.25)
```

(b) Suppose 3rd graders scoring at or above 56 are eligible for the G&T program. Create a treatment assignment variable re-centered at zero, and a "gap" variable that contains the distance between the running variable and the cut score.

```
gen above56 = (grade3test>=56)
gen gap = grade3test-56
```

(c) Assume perfect compliance. Create an indicator variable for G&T participation $inGT$, that equals one for treated students and zero otherwise. What proportion of students are treated? Try estimating an OLS regression for G&T participation where $inGT$ is regressed on the *gap* and the threshold indicator *above56*. What happens and why?

(d) Create an outcome variable (*grade 4* test score) such that G&T participation has a positive treatment effect of +3 points. Assume that test growth from 3rd to 4th grade would be 5 points in the absence of treatment. As before, we will include some random noise, and round the test scale to the nearest 0.25.

```
gen grade4test = round(trueability + 5 + rnormal() + (3*inGT), 0.25)
```

(e) Estimate an RD model for 4th grade test scores, assuming a linear relationship with the running variable (3rd grade test scores). Use the full range of data and no kernel. First do this assuming the same slope on either side of the cut score. Then, allow the slope to vary on either side. Is there evidence of a change in slope beyond the cut score? Does this finding make sense to you? Repeat using `rdrobust` with the default options.

(f) Drop the existing *inGT* and *grade4test* variables and re-create them assuming a "fuzzy" GT treatment that increases smoothly with grade 3 test scores and then jumps discontinuously (by about 70 percentage points) at the cut score. This might arise if G&T placement is dependent on the grade 3 test score as well as other factors (e.g., parental input, teacher recommendation). Use the syntax below. Now what proportion of students are treated, overall? Below the cutoff? Above? Use `binscatter` to visualize the relationship between treatment at the grade 3 score.

```
drop inGT grade4test
gen inGT=round(-.77+.007*grade3test+0.7*above56+runiform())
gen grade4test = round(trueability + 5 + rnormal() + (3*inGT), 0.25)
```

(g) As in (c), estimate a regression for G&T placement where *inGT* is regressed on the *gap* and the threshold indicator *above56*. Interpret your results. (Try estimating this in two ways: first assuming the slope is constant on either side of the cutoff, and then allowing the slope to change). For later use, use the `predict` command to get predicted values for treatment (placement in G&T) given the 3rd grade score. Call this variable *predGT*.

(h) Re-estimate the RD models from part (e) assuming a linear relationship with the running variable. Assume the discontinuity is "sharp," even though we know otherwise. How does the estimated treatment effect differ from the known treatment effect of 3 points?

(i) We will now estimate the treatment effect using RD but allowing for non-compliance (fuzzy RD). Do this two ways:

- Use `rdrobust` with the default options but add the option `fuzzy(inGT)`.
- Using two stage least squares. One way to implement this is to regress *grade4test* on the predicted *inGT* from (g). `ivregress` is a little trickier in this context—you'll need to manually create "gapabove" and "gapbelow" variables if you want the slope to vary above and below *c*. See the syntax below.

```
gen gapabove = gap*above56
gen gapbelow = gap*(1-above56)
ivregress 2sls grade4test (inGT=above56) gapbelow gapabove , first
```

Note the `rdrobust` and manual estimates are not directly comparable since the former uses the optimal bandwidth. How close do these get to the "true" treatment effect of 3 points?

(j) Finally, do a manipulation test using `rddensity`. What does it find? Would you expect to see evidence of manipulation here?