Regression II
Vanderbilt University
Prof. Sean P. Corcoran

November 19, 2021

Lecture 11 In-Class Exercise

Exercise 1. This exercise will generate data with a known discontinuity in y at a threshold level of x, and then estimate a RD model. It also illustrates the McCrary test. (Adapted from Ballou).

1. 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 strict or fuzzy regression discontinuity?

```
clear
set seed 1234
drawnorm x w e u, n(1000)
gen y = 3 + 3*x + .5*x^2 + w + u
gen t = (x > 1)
replace y = y + .5*t
```

2. Produce a scatterplot of y against x. Do you see evidence of a discontinuity? Try using `binscatter`. Do you see a discontinuity?

3. Now estimate a parametric RD model assuming a linear relationship with the running variable (with the same slope on either side of the cut score). How close does it get to estimating the true treatment effect? Provide an intuitive explanation for your finding.

4. Repeat but using a quadratic function of the running variable. Does this help?

5. Obtain some nonparametric estimates using the Stata command `rd`. `rd` estimates treatment effects using local linear or kernel regression models on both sides of the cut score. Note the option `z0( )` provides Stata the known cutoff value. The option `strineq` (strict inequality) tells Stata that treatment is assigned *above* the cut score—observations *at* the cut score are not treated. (The default assumes treatment begins *at* the cut score). The option `bwidth( )` allows you to select a bandwidth for local linear regression. There are also lots of options for `rd` that produce graphs.

```
rd y x, z0(1) strineq bwidth(.4)
```

   The estimate `lwald` (Local Wald) is your baseline result. The additional estimates `lwald50` and `lwald200` are robustness checks at other bandwidths (50% and 200% of your specified bandwidth).

6. Check for manipulation in the running variable in two ways: by inspection using `histogram`, and using McCrary's `DCdensity` command. The option `breakpoint` tells Stata the known cutoff value. What does the latter test conclude?

```
histogram x, kdens
DCdensity x, breakpoint(1) gen(Xj Yj r0 fhat se_fhat)
```

7. 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. Then, re-do the McCrary test.

```
replace x = x + .4 if x < 1 & x > .65 & e > 0
drop Xj Yj r0 fhat se_fhat
DCdensity x, breakpoint(1) gen(Xj Yj r0 fhat se_fhat)
```

8. Now that we know there is manipulation, try estimating the parametric and non-parametric RD models in (3) and (5). How do the estimates compare?

Exercise 2. This exercise, 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.

1. 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)
```

2. 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
```

3. Assume perfect compliance. Create an indicator variable for G&T participation $inGT$, that equals one for treated students and zero otherwise. What fraction of students participate in G&T? Try estimating a regression for G&T participation where $inGT$ is regressed on the *gap* and the threshold indicator *above56*. What happens?

4. Create the 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)
```

5. Estimate a parametric RD model assuming a linear relationship with the running variable. 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?

6. 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. What fraction of students are treated, overall? Below the cutoff? Above?

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

7. As in (3), 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). With a discrete running variable, it is usually advisable that you adjust the standard errors for clustering by that variable. For later reference, use the `predict` command to get predicted values for treatment (placement in G&T) given the 3rd grade score. Call this variable *hat_trt*.

8. Re-estimate the parametric RD model assuming a linear relationship with the running variable. Assume the discontinuity is "sharp," even though we know otherwise. Again, cluster the standard errors by the grade 3 score. How does the estimated treatment effect differ from the known treatment effect of 3 points? Repeat using the non-parametric `rd`. How does the point estimate compare?

9. We will now estimate the treatment effect using `rd` but allowing for non-compliance. Because of the fuzzy RD, you need to modify the `rd` command to include the treatment variable ($inGT$), otherwise it assumes $inGT = 0$ below the cut score and $inGT = 1$ above it. Notice the running variable goes last in the list of variables.

```
rd grade4test inGT grade3test, z0(56) graph noscatter
```

Note that `rd` gives you more estimates when the treatment assignment is fuzzy. The `numer` line here is nearly identical to the sharp RD result from (8). This is the "reduced form." The `denom` line is roughly equivalent to the effect of exceeding the threshold on treatment from (7). This is the "first stage." `lwald` is the reduced form divided by the first stage.

10. To connect to the lecture on IV, try the syntax below. (We cannot use the Stata factor variables for the two *gap*= slopes, since *above56* cannot be included in the list of regressors—it is the exogenous instrument). Compare the first stage and final point estimates to (9).

```
gen gapabove = gap*above56
gen gapbelow = gap*(1-above56)
ivregress 2sls grade4test (inGT=above56) gapbelow gapabove , first cluster(grade3test)
```