//PART 0

cls

clear all

\*COPY AND PASTE YOUR PATH

//cd "ADD YOUR PATH"

cd "C:\Users\Jinhyun\Documents\GitHub\stata\Financial econometrics\Week 3"

//Part 1

//1

use "final4.dta"

append using "final5.dta"

// We have appended the grade 5 data from below to the grade 4 data. All columns were the same.

// Another usefull command is merge, which matches data per unit rather than appending from below.

//2

gen cohpctboy = 100\* c\_boys / ( c\_boys + c\_girls)

// Here the cohpctboy variable is generated, which captures the share of boys in each cohort (in percentages)

ren (c\_size tipuach) (enroll cohpctda)

// c\_size and tipuach variables are renamed

global vlist schlcode classid grade enroll cohpctboy cohpctda classize avgmath avgverb

// vlist is created as a global variable

keep $vlist // drop variables not listed in vlist

order $vlist

// sort schlcode grade classid

// sort observations

//3 Checking data

sum

sum enroll classize avgverb avgmath, det

keep if classize<45 & enroll>5

// Drop schools that have very few enrolled students or classes that are unlikely large, given the Maimonides rule of 40

keep if inrange(avgverb,20,100) & inrange(avgmath,20,100)

// The test scores are on a scale from 1 to 100. Class averages should therefore also be within those bounds.

save work, replace

// 4 FS whole range

global grti ytitle(Class size) xtitle(Enrollment count)

scatter classize enroll, $grti

//Note: you should install the packages below if you have not done so already. You only have to install packages once, not every time you run this do-file

// net install rdrobust, from(https://raw.githubusercontent.com/rdpackages/rdrobust/master/stata) replace

// net install lpdensity, from(https://raw.githubusercontent.com/nppackages/lpdensity/master/stata) replace

// net install rddensity, from(https://raw.githubusercontent.com/rdpackages/rddensity/master/stata) replace

rdplot classize enroll, c(41) graph\_options(name(gr4a) $grti)

// Instead of a scatterplot with a dot for every single observation, we now plot the average of classize for different values of enrollment.

rdplot classize enroll, c(81) graph\_options(name(gr4b) $grti)

// By varying the option c(..), we can plot a vertical line for different values of the threshold.

rdplot classize enroll, c(121) graph\_options(name(gr4c) $grti)

rdplot classize enroll, c(161) graph\_options(name(gr4d) $grti)

// 5 FS at first cutoff only

keep if inrange(enroll, 0, 80)

rdplot classize enroll, c(41) graph\_options(name(gr5) $grti)

// Class size seems to be discontinuous at the cutoff of 41 enrolled students. Schools with 41 students have significant smaller classes than schools with 40 enrolled students.

// 6 FS Bin experiment

rdplot classize enroll, c(41) nbins(20 20) graph\_options(name(gr6a) $grti)

// With wider bins, we have less points in the plot. Each point represents schools with 2 different values of enrollment (e.g. schools with 39 and 40 students are represented by the point to the left of the cutoff)

rdplot classize enroll, c(41) nbins(10 10) graph\_options(name(gr6b) $grti)

// Here we have even fewer bins, with each bin representing 4 different values of enrollment

// 7 FS Maimonides and create your own plot

ereturn list

global f0 = subinstr(e(eq\_l), "-", "-41", .)

global f1 = subinstr(e(eq\_r), "-", "-41", .)

rdplot classize enroll, c(41) nbins(20 20) hide genvars

twoway (scatter rdplot\_mean\_y rdplot\_mean\_bin, msize(medium) ms(o) mc(gs10)) ///

(function $f0, ra(0 41) lc(black)) (function $f1, ra(41 80) lc(black)) ///

(function y = x /(int((x-1)/40)+1), ra(0 80) lp(dash) lc(gs7)), name(gr7) $grti ///

legend(order(1 "Sample average within bin" 2 "Polynomial fit order 4" 4 "Maimonides Rule"))

drop rdplot\*

\*Apart from the technicalities, this is useful to understand (i) how to customize a RDD plot (ii) understand the differences between the ideal implementation of the rule with respect to the actual one

// 8 RF

global grti ytitle(Average math score) xtitle(Enrollment count) ylabel(50(10)90)

rdplot avgmath enroll, c(41) nbins(10 10) graph\_options(name(gr8a) $grti)

// The average test score seems to increase a bit at the threshold, but a 4th order polynomial (which is used by default) could be overfitting the data

rdplot avgmath enroll, c(41) nbins(10 10) graph\_options(name(gr8b) $grti) p(1)

// Here we re-do the rdplot but use a 1st order polynomial instead. We do still see a jump in average scores at the cutoff

global grti ytitle(Average reading score) xtitle(Enrollment count) ylabel(50(10)90)

rdplot avgverb enroll, c(41) nbins(10 10) graph\_options(name(gr8c) $grti)

rdplot avgverb enroll, c(41) nbins(10 10) graph\_options(name(gr8d) $grti) p(1)

// Also for the verbal test there seems to be a jump in average scores at the cutoff, indicating that schools with 41 enrolled students have better academic outcomes than schools with 40

\*Visually, it seems we have a good FS and the RF is SUGGESTIVE of a positive effect of class size (smaller classes => better performance!)

// 9 IV

gen w = enroll - 41

gen z = (w>=0)

egen clid = group(schlcode grade)

global cov cohpctboy cohpctda

qui foreach y of varlist avgmath avgverb {

local c

forv i = 0 / 1 {

if (`i' == 1) local c $cov

eststo fs`i': reg classize z c.z#c.w w `c', cluster(clid)

eststo rf`i': reg `y' z c.z#c.w w `c', cluster(clid)

eststo iv`i': ivregress 2sls `y' c.z#c.w w `c' (classize = z), cluster(clid)

}

noi esttab fs0 fs1 rf0 rf1 iv0 iv1, b(a2) nogaps se mti(FS FS RF RF IV IV) order(classize)

}

\*\*\*\*\*

**// FS: z tells me the change in average class size around the cutoff.**

**// The first 2 columns show the first stage and tell us that passing the threshold of 41 students reduces average class size by 13 students.**

**// w is telling the slope of the relationship between class size and enrolmment before the cutoff (left part of the graph) => As the figure shows this slope is positive**

**// w\*z captures the CHANGE in the slope after the cutoff +> The figure shows that this slope is less steep (right part of the graph)**

**// RF: Estimates the impact of being just above the cutoff (meaning: in a smaller class)**

**// The 3rd and 4th column show that test scores increases by about 3 points at the cutoff (for both math and verbal tests).**

**//IV: The last 2 columns are the IV estimates, which captures the impact of 1 extra-student in a classroom on the average test score. Note that the IV coefficient is the reduced from coefficient (+3 point) divided by the first stage (-13 students)**

\*\*\*\*\*

// CONTROLS: Adding controls in columns 2, 4 and 6 doesn't seem to impact the coefficient of interest (in col 1, 3, 5). This is also what we would expect, as the controls are added to improve precision but are not necessary to reduce bias. We see that the standard errors in columns 4 and 6 are indeed smaller than their counterparts in columns 3 and 5.

\*\*\*\*\*

// 10 Balance

global grti ytitle(Percentage boy) xtitle(Enrollment count) ylabel(0(20)100)

rdplot cohpctboy enroll, c(41) nbins(10 10) graph\_options(name(gr10a) $grti) p(1)

global grti ytitle(Percentage disadvantaged ) xtitle(Enrollment count) ylabel(0(20)100)

rdplot cohpctda enroll, c(41) nbins(10 10) graph\_options(name(gr10b) $grti) p(1)

// GOOD NEWS: We don't see any jumps in pre-determined characteristics at the cutoff!

qui eststo b1: reg cohpctboy z c.z#c.w w

// Here we regress one pre determined characteristic on passing the threshold. The coefficient of z should capture any discontinuity at the cutoff.

qui eststo b2: reg cohpctda z c.z#c.w w

// Here we regress the other pre determined characteristic on passing the threshold. The coefficient of z should capture any discontinuity at the cutoff.

noi esttab b1 b2, b(a2) nogaps se

qui suest b1 b2

// Suest b1 b2 pools the 2 regressions together

test z

// We don't see any statistically significant jumps in pre determined characteristics at the cutoff, as the p-value is higher than 0.05

// 11 Bunching

global gr ytitle(Density) xtitle(Enrollment count) legend(off) ylabel(0(0.01)0.04)

rddensity enroll, c(41) plot p(3) graph\_opt($gr name(gr11a))

rddensity enroll if classid==1, c(41) plot p(3) graph\_opt($gr name(gr11b))

//We only want each school to show up once in the density plot of enrollment numbers. Since our observations are on school-classroom levels, not using "if classid==1" would result in schools with multiple classes to show up multiple times. In short: To avoid duplications of schools with more classes for the same grade.

// The p-values are lower than 0.05. The density of enrollment numbers is not smooth, indicating potential manipulation of the running variable

/\*

//Running variable: enroll.

------------------------------------------

Method | T P>|T|

-------------------+----------------------

Robust | 2.4227 0.0154

------------------------------------------

\*/

//Given the evidence in the graph (bunching to the right) schools prefer (have incentives) to have more (and smaller) classes.

\*\*\*\*

// 12 Bandwidth & polynomial sensitivity

qui foreach y of varlist classize avgmath avgverb {

if ("`y'"=="classize") local bin

else local bin nbins(10 10)

//start lin spec

local s c.w

//local

rdbwselect `y' enroll, c(41)

global bw = e(h\_mserd)

eststo p0: reg `y' z c.z#(`s') `s' if inrange(w, -$bw, $bw ), cluster(clid)

rdplot `y' enroll, c(41) h($bw $bw) p(1) `bin' graph\_options(legend(off) name(p0, replace))

//global

forv i = 1 / 5 {

eststo p`i': reg `y' z c.z#(`s') `s', cluster(clid)

rdplot `y' enroll, c(41) p(`i') `bin' graph\_options(legend(off) name(p`i', replace))

local s `s'##c.w

}

noi esttab p1 p2 p3 p4 p5 p0, b(a2) nogaps se

graph combine p1 p2 p3 p4 p5 p0, xcommon ycommon imargin(0 0 0 0) l1title(`y') ///

b1title(Enrollment count) name(gr\_`y')

graph drop p\*

}

// In the first 5 graphs we've fitted polynomials of different orders to the data left and right of the threshold. We see that the results are relatively robust for different orders of polynomials, except for very high orders. Gelman and Imbens (2019) advise against using polynomials with high-order polynomials for Regression Discontinuity (p=3 or higher is discouraged). Given that class size just below the cutoff does not seem to fit the linear trend in the 1st graph very well, one could argue that quadratic polynomials is the best in this situation

// The last graph shows a linear polynomial fit for a optimally chosen bandwidth. This means that we are only using observations relatively close to the threshold. This specification also seems to give results that are comparable to the quadratic specification shown in the 2nd graph/column 2.