HCMG 901 Problem Set 4

Andrés Rovira

April 9, 2023

1. It will help you to first read the 1999 QJE article so you become familiar with what was done there and why.

Done.

2. Estimate basic bivariate regressions of students' test scores on class size. What patterns do you find? Describe one source of endogeneity and discuss how this might bias these estimates.

Table 1: Bivariate Regressons of Test Scores on Class Size

	4th Graders		5th Graders	
	(1)	(2)	(3)	(4)
	Grammar Score	Math Score	Grammar Score	Math Score
Class Size	0.141***	0.221***	0.221***	0.322***
	(0.0300)	(0.0327)	(0.0285)	(0.0333)
Constant	68.21***	62.19***	67.77***	57.66***
	(0.975)	(1.055)	(0.915)	(1.053)
Observations	2049	2049	2019	2018

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

See Table 1. The general pattern is that larger class sizes are associated with higher test scores, contrary to our expectations of the causal relationship, and contrary to the IV estimates of Angrist and Lavy (1999). Clearly class size is endogenous. One source of this endogeneity could be school administrators anticipating positive benefits to small classes, and assigning underpriveleged or poorer-performing children to the smaller classes to compensate for the greater difficulty of teaching these students. This form of selection would likely bias OLS estimates of class size on test scores upwards, as we see in Table 1.

3. Calculate the average class size and number of classes in a grade and school that would result if schools mechanically applied Maimonides' Rule (that is, calculate the predicted average class size and the predicted number of classes). Graph the relationship between enrollment and predicted class size. Describe potential sources of variation in class size that you might be able to use for identification.

See Figures 1 and 2 below. Potential sources of variation in class size that might be useful for identification are the discrete jumps around enrollment multiples of 40. When comparing classes close to either side of these cutoffs, we might expect the classes to be similar along all other dimensions except for class size as induced by the cutoff rule. Thus, we can potentially think of this as an exogenous source of variation in class size, which is what we would want for the purposes of identifying the causal effect of class size on other outcomes.

Figure 1: Predicted Number of Classes by Enrollment

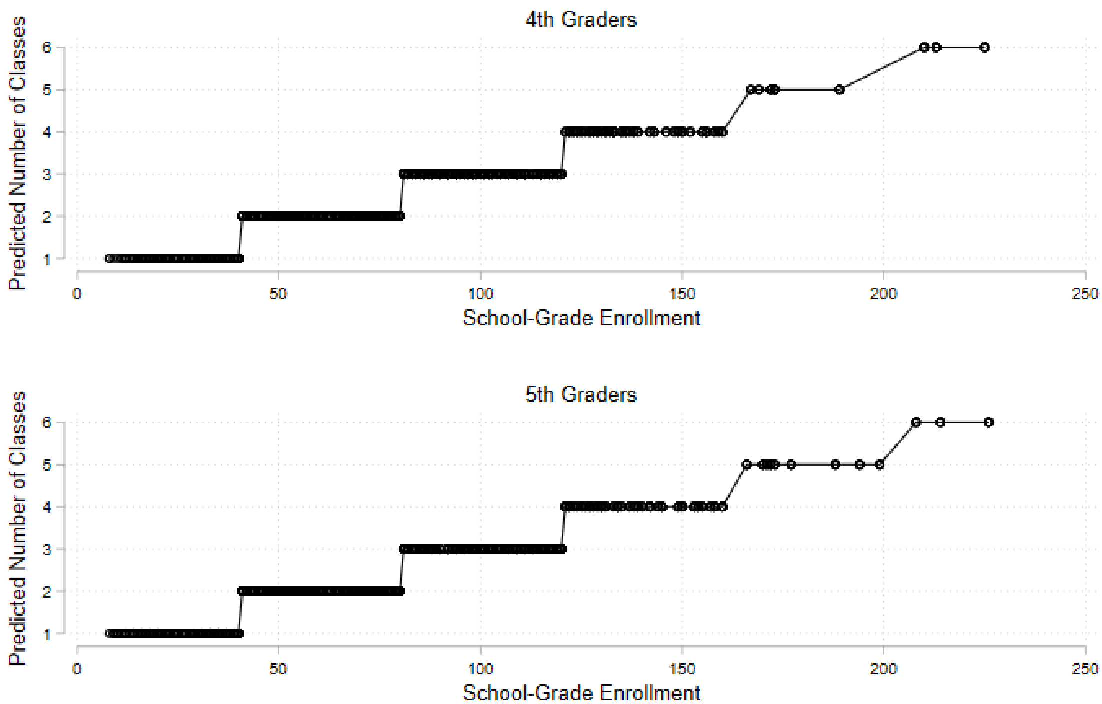
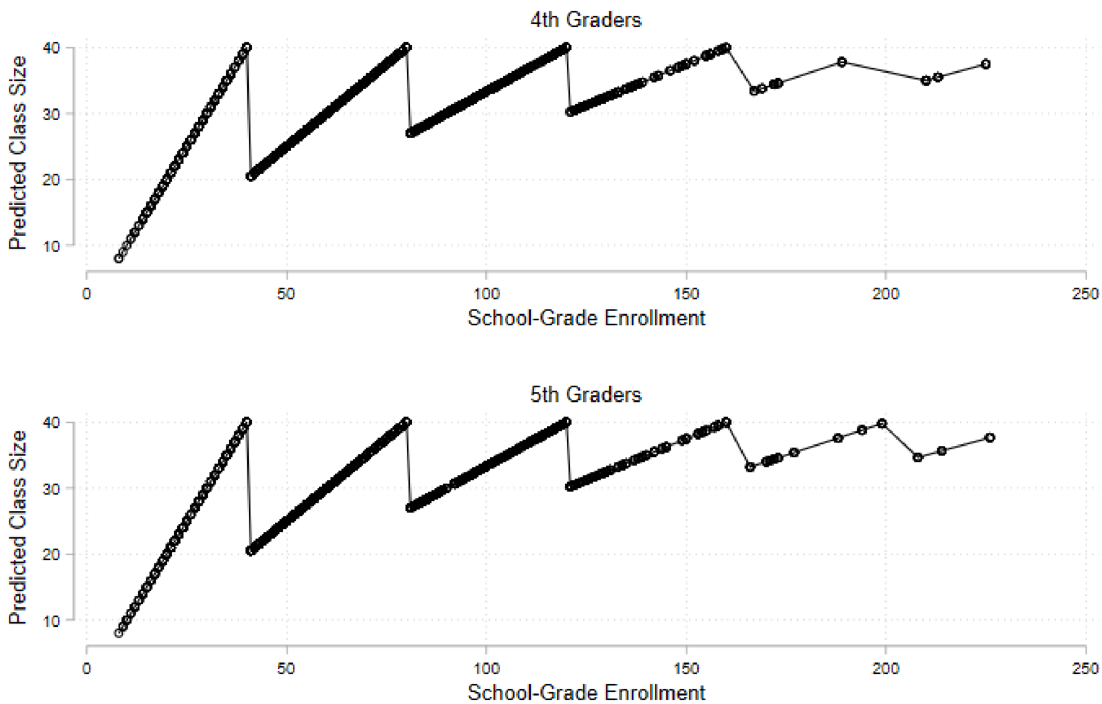


Figure 2: Predicted Average Class Size by Enrollment



4. Calculate the average class size students in a given school and grade actually experience. Present descriptive statistics on this variable and overlay actual class size and predicted class size against enrollment on a graph. What percentage of schools either overshoot the class size cap or create a new class before they are required to do so by Maimonides' Rule? How do these schools' characteristics compare with schools who 'comply' with Maimonides' Rule? Discuss whether (or in what circumstances) noncompliance with Maimonides' Rule might affect your identification strategy.

See Table 2 for descriptive statistics on the mean class size students actually experience within a given school and grade.

Table 2: Descriptive Statistics on Actual Mean Class Size

	Mean	SD	Min	Q1	Median	Q4	Max
Mean Class Size	30.1	6.1	8.0	26.0	30.8	34.7	44.0

Figure 3 overlays actual mean class size against predicted mean class size as a function of enrollment.

Figure 3: Actual and Predicted Average Class Size by Enrollment

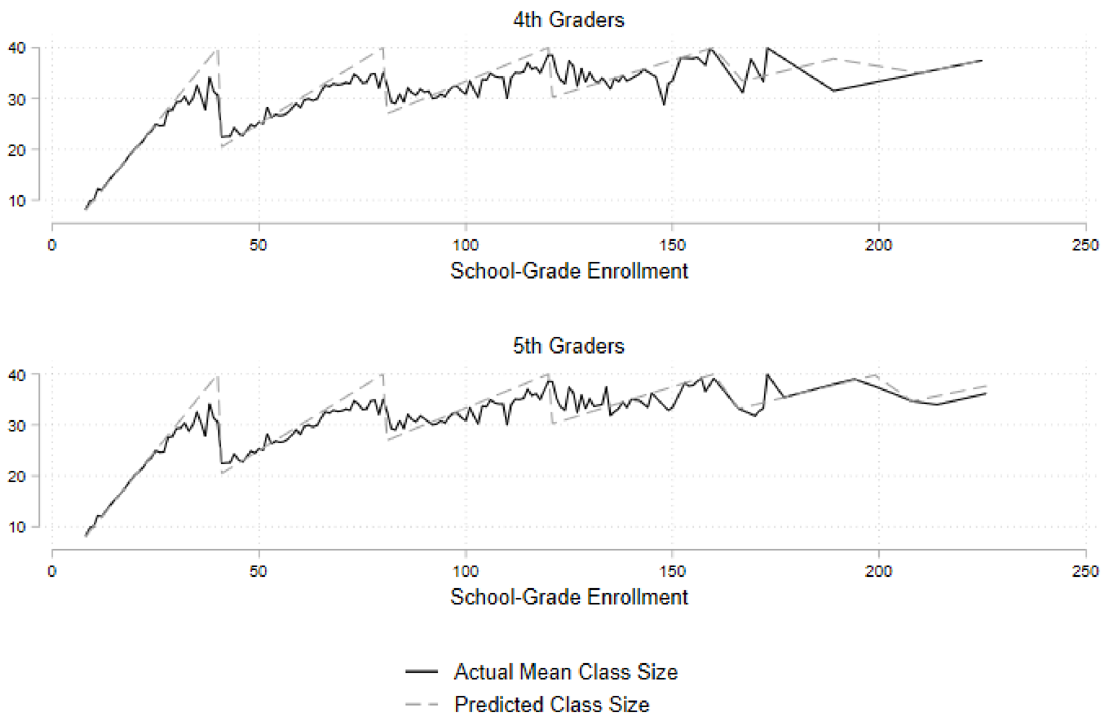


Table 3 moves to the school-grade level and gives the percentages of these that are “compliers”, “overshooters”, and “early splitters”, in terms of the class size cap, and then lists the means and standard deviations of covariates, broken down by these groups.

I should note that I use the total enrollment variable (“c_size”) used by Angrist and Lavy, but this variable sometimes translates poorly to the sum of class sizes (“classize”) across all classes in that school-grade. As a result, actual and predicted mean class sizes differ for many schools. However, my definition of “compliers” as I have interpreted it from your question, is more forgiving, as it includes any school not classified as an “overshooter” or “early splitter”. Overshooters are schools whose actual mean class size exceeds 40, and early splitters are schools whose actual number of classes exceeds the predicted number of classes.

Table 3: Characteristics of Schools by Rule Compliance

	All		Compliers (96.4%)		Early Splitters (2.5%)		Overshooters (1.1%)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Mean Class Size	28.9	6.7	28.9	6.6	23.4	4.4	41.4	1.3
Percent Disadvantaged	15.8	15.1	15.8	15.2	18.2	12.8	12.8	14.2
Mean Math Score	67.3	9.0	67.4	9.0	63.9	9.7	69.8	7.4
Mean Grammar Score	73.0	7.9	73.1	7.9	69.8	8.1	73.9	8.5
School-Grade Enrollment	63.8	35.4	63.4	35.3	71.4	32.5	84.8	36.3

Evidently, early splitters have more disadvantaged students and lower test scores than compliers, and overshooters have less disadvantaged students and higher test scores. This would be consistent with schools trying to assign class size in compensatory fashion so that less advantaged and poorer performing students get smaller classes. If these same types of schools are also willing and able to manipulate their enrollment counts to position themselves on the desired side of class size cutoffs, then the key identification assumption of the RD design will be violated. That is, school characteristics will not vary smoothly at the cutoffs, so class sizes as determined by position relative to the cutoff cannot be interpreted as exogenous.

5. Next, you will use an RD design to estimate the impact of class size on achievement at the first "discontinuity" created by Maimonides' rule. Following Lee and Lemieux's Table 2 that we saw in the RD class, estimate a series of fuzzy RD (IV) models, in which you vary the bandwidth (i.e., window around the discontinuity) and degree of polynomial in the running variable. How robust are you estimates to smaller bandwidths and higher degree polynomials?

See Tables 4 and 5 below. Clustered standard errors at the school level are in parentheses. Bandwidths ("BW") are restrict student enrollment to within the given distance in student counts from the cutoff of 40. The estimates are not very robust to different choices of bandwidth and polynomial degree, as the coefficients move around a good bit. Standard errors indicate that most estimates are not significantly different from 0, but some are, particularly with larger bandwidths, which makes sense as they are able to use more data points.

Table 4: Fuzzy RD Estimates of of Class Size on Math Score

Polynomial Degree	BW = 2	BW = 5	BW = 8	BW = 10	BW = 15	BW = 20	BW = 25	BW = 30
0	–0.21	–0.33	–0.29	–0.32	–0.58	–1.77	1.63	0.78
	(0.27)	(0.19)	(0.18)	(0.17)	(0.24)	(0.91)	(0.80)	(0.24)
1	–0.05	–0.21	–0.30	–0.19	–0.12	–0.19	–0.20	–0.06
	(0.66)	(0.37)	(0.26)	(0.23)	(0.17)	(0.14)	(0.13)	(0.11)
2	–0.13	–0.23	–0.39	–0.31	–0.16	–0.21	–0.26	–0.21
	(0.67)	(0.37)	(0.28)	(0.24)	(0.17)	(0.13)	(0.12)	(0.11)
3	–0.13	0.06	–0.05	–0.31	–0.23	–0.12	–0.15	–0.23
	(0.67)	(0.42)	(0.39)	(0.35)	(0.25)	(0.20)	(0.17)	(0.15)
4	0.17	0.06	–0.07	–0.31	–0.40	–0.17	–0.14	–0.22
	(0.57)	(0.42)	(0.39)	(0.35)	(0.28)	(0.21)	(0.18)	(0.16)

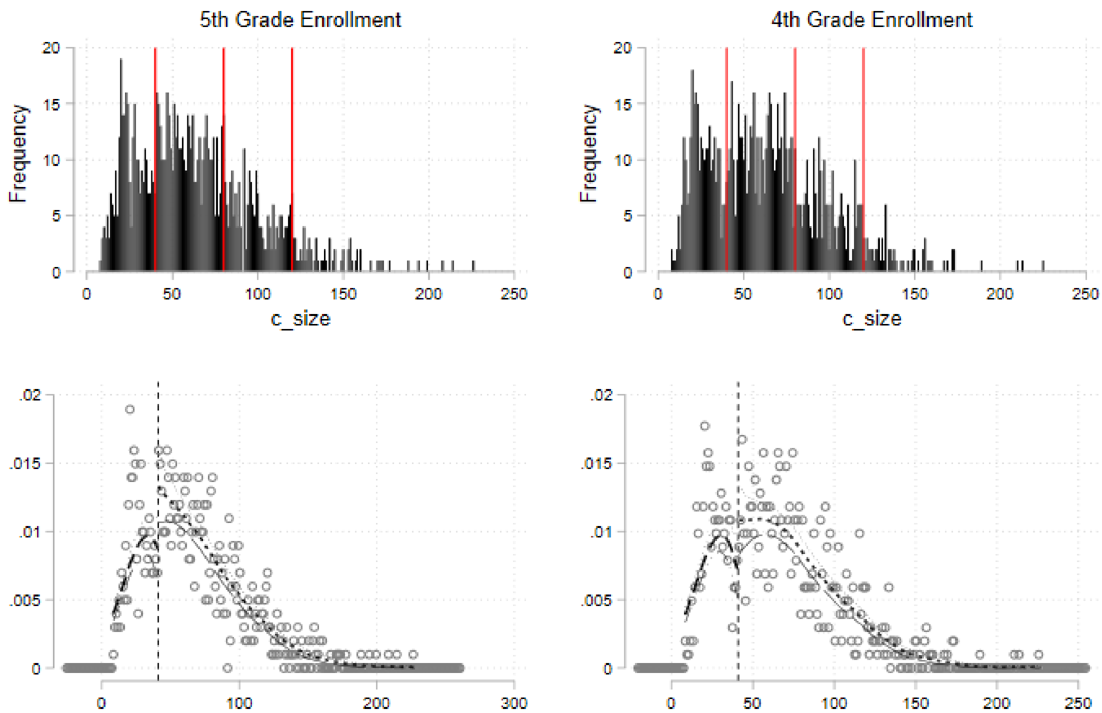
Table 5: Fuzzy RD Estimates of of Class Size on Grammar Score

Polynomial Degree	BW = 2	BW = 5	BW = 8	BW = 10	BW = 15	BW = 20	BW = 25	BW = 30
0	–0.35	–0.52	–0.36	–0.31	–0.45	–1.20	0.90	0.40
	(0.26)	(0.18)	(0.16)	(0.15)	(0.20)	(0.72)	(0.63)	(0.21)
1	–0.70	–0.39	–0.57	–0.47	–0.30	–0.29	–0.25	–0.14
	(0.67)	(0.34)	(0.25)	(0.22)	(0.15)	(0.13)	(0.11)	(0.10)
2	–0.76	–0.42	–0.59	–0.56	–0.31	–0.27	–0.29	–0.25
	(0.65)	(0.35)	(0.27)	(0.23)	(0.16)	(0.12)	(0.11)	(0.10)
3	–0.76	–0.25	–0.45	–0.63	–0.50	–0.38	–0.35	–0.35
	(0.65)	(0.37)	(0.39)	(0.34)	(0.25)	(0.19)	(0.16)	(0.14)
4	–0.52	–0.25	–0.46	–0.58	–0.65	–0.44	–0.33	–0.33
	(0.53)	(0.37)	(0.40)	(0.35)	(0.28)	(0.21)	(0.17)	(0.15)

6. The 2019 AER Insights article revisits this setting with newer data and finds that there was some manipulation in class size after all and once they account for this, there are zero effects of class size on test scores! Unfortunately the newer data is non-public so we cannot replicate the new analysis. However, the new paper also looks for manipulation on the older data. In particular, I'd like you to replicate their Figure A6 where they look for jumps in enrollment at the thresholds and perform a McCrary test. This analysis reveals some jumps in enrollment frequency at the thresholds. How should we interpret this?

See Figure 4. We should interpret the jumps in enrollment frequency at the thresholds as indicating manipulation of the running variable at the threshold, which, as described before, invalidates the identifying assumption of the RD design. Evidently, some schools are bunching just above the threshold, probably because they feel greater need for smaller classes, in anticipation of positive effects.

Figure 4: Replication of Figure A6



7. In response, the authors do a robustness check where they drop data close to the thresholds. They refer to this as a 'donut' analysis. This is presented in Table A6. Replicate this table. What does this tell us? How is it helpful?

See Table 6. This table tells us that the donut estimates are quite close to the baseline estimates which Angrist et al. (2019) replicate in their Table A4, regardless of the width of the donut hole or the type of test score. This is helpful because it suggests that manipulation of enrollment counts very near the threshold do not substantially affect the estimates of interest. Thus, perhaps we need not be so concerned about the bunching above the cutoff we saw previously.

Table 6: Replication of Table A6

Grade	Donut	Language (1)	Language (2)	Math (3)	Math (4)
5	[39,41]	–0.2341	–0.2010	–0.1947	–0.2144
		(0.0762)	(0.0954)	(0.1018)	(0.1306)
5	[38,42]	–0.2406	–0.2072	–0.2000	–0.2213
		(0.0776)	(0.0987)	(0.1044)	(0.1368)
5	[37,43]	–0.2152	–0.1696	–0.1930	–0.2024
		(0.0777)	(0.0991)	(0.1054)	(0.1388)
4	[39,41]	–0.1267	–0.0581	–0.0544	–0.0353
		(0.0612)	(0.0690)	(0.0749)	(0.0858)
4	[38,42]	–0.1187	–0.0431	–0.0438	–0.0208
		(0.0632)	(0.0719)	(0.0775)	(0.0899)
4	[37,43]	–0.1166	–0.0390	–0.0467	–0.0227
		(0.0649)	(0.0743)	(0.0794)	(0.0927)
Controls:					
Percent Disadvantaged		X	X	X	X
Enrollment		X	X	X	X
Enrollment Squared /100			X		X

Stata code

```
/*******************************************************************************
AUTHORS: Andres Rovira and Fabio Schanaider
CREATED: 2023-04-01
PURPOSE: Solve HCMG 901 Problem Set 4
*******************************************************************************/

global home "~"
if "$S_OS" == "Windows" global home "`:env USERPROFILE'"
global code "$home/Dropbox (Penn)/Classes/2_health_applied_metrics/ps4/code"
global tex  "$home/Dropbox (Penn)/Apps/Overleaf/hcmg901_ps4"
cd "$code"

// ssc install blindschemes
// ssc install texsave
// ssc install outreg2
// ssc install estout
// net install grc1leg
copy "https://eml.berkeley.edu/~jmccrary/DCdensity/DCdensity.ado" "./", replace

set scheme plotplainblind, perm

/* Data and reference code sources:
- Angrist and Lavy (1999): https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive
- Angrist et al. (2019): https://www-aeaweb-org.proxy.library.upenn.edu/articles?id=10.1257/aeri.2018
Components are copied to code/reference and input folders
*/

********************************************************************************
**# Part 2
********************************************************************************

// Combine 4th and 5th grader data
use "../input/al_1999/final4.dta", clear
rename *4* *5*
append using "../input/al_1999/final5.dta"
rename *5* *0*

// Clean some variables like AL1999 do
replace avgverb = avgverb - 100 if avgverb > 100
replace avgmath = avgmath - 100 if avgmath > 100
replace avgverb  = . if verbsize == 0
replace passverb = . if verbsize == 0
replace avgmath  = . if mathsize == 0
replace passmath = . if mathsize == 0

// Use their sample restrictions?
keep if 1 < classize & classize < 45 & c_size > 5
keep if c_leom == 1 & c_pik < 3
```

```
label var c_size   "School-Grade Enrollment"
label var classize "Class Size"
label var avgmath  "Math Score"
label var avgverb  "Grammar Score"
label var tipuach  "Percent Disadvantaged"

save "../intermediate/al_1999_data.dta", replace
use "../intermediate/al_1999_data.dta", clear

local test_score_vars avgverb avgmath

eststo clear
forvalues g = 4/5 {
        foreach var of varlist `test_score_vars' {
                eststo: regress `var' classize if grade == `g', robust
        }
}

esttab * using "$tex/tab_2.tex", replace label se booktabs ///
        mgroups("4th Graders" "5th Graders", pattern(1 0 1 0) span prefix(\multicolumn{@span}{c}{) su
        title("Bivariate Regressons of Test Scores on Class Size")


********************************************************************************
**# Part 3
********************************************************************************

// Predictions of Maimonides' Rule
gen pred_num_classes = int((c_size - 1) / 40) + 1
gen pred_class_size = c_size / pred_num_classes

label var pred_num_classes "Predicted Number of Classes"
label var pred_class_size  "Predicted Class Size"

sort c_size

// Predicted number of classes by enrollment
tw conn pred_num_classes c_size if grade == 4, name(grade4, replace) title("4th Graders")
tw conn pred_num_classes c_size if grade == 5, name(grade5, replace) title("5th Graders")
graph combine grade4 grade5, cols(1)
graph export "$tex/fig_3_num_classes.png", replace

// Predicted class size by enrollment
tw conn pred_class_size c_size if grade == 4, name(grade4, replace) title("4th Graders")
tw conn pred_class_size c_size if grade == 5, name(grade5, replace) title("5th Graders")
graph combine grade4 grade5, cols(1)
graph export "$tex/fig_3_class_size.png", replace


********************************************************************************
**# Part 4
```

```
********************************************************************************

// Actual vs Predicted Mean Class Sizes by Enrollment
egen mean_class_size_byenroll = mean(classize), by(c_size)
label var mean_class_size_byenroll "Actual Mean Class Size"
tw line mean_class_size_byenroll pred_class_size c_size if grade == 4, ///
        name(grade4, replace) title("4th Graders")
tw line mean_class_size_byenroll pred_class_size c_size if grade == 5, ///
        name(grade5, replace) title("5th Graders")
grc1leg grade4 grade5, cols(1)
graph export "$tex/fig_4.png", replace

// Mean class size at school-grade level
egen mean_class_size = mean(classize), by(schlcode grade)
label var mean_class_size "Mean Class Size"

eststo clear
estpost tabstat mean_class_size, c(stat) stat(mean sd min p25 p50 p75 max)
esttab using "$tex/tab_4_descriptives.tex", replace booktabs nonumber noobs label ///
        cells("mean(fmt(%6.1fc)) sd(fmt(%6.1fc)) min p25 p50 p75 max") ///
        collabels("Mean" "SD" "Min" "Q1" "Median" "Q4" "Max") ///
        title("Descriptive Statistics on Actual Mean Class Size")

// Collapse to school-grade level and summarize mean covariates within
collapse (mean) avgmath avgverb (count) num_classes = classid (sum) classize, ///
        by(schlcode grade mean_class_size pred_class_size pred_num_classes tipuach c_size cohsize)
isid schlcode grade
label var avgmath  "Mean Math Score"
label var avgverb  "Mean Grammar Score"

// Compliance must use number of classes due to imprecision in translation of cohort size to class si

// Overshooters: mean class size exceeds 40
gen overshooter = (mean_class_size > 40)
// Early Splitters: actual number of classes exceeds predicted number
gen earlysplitter = (num_classes > pred_num_classes)
// Compliers: all others. even if their number of classes or class size misses prediction
gen complier = (overshooter == 0) & (earlysplitter == 0)

eststo clear
local covars mean_class_size tipuach avgmath avgverb c_size
local stats "c(stat) stat(mean sd)"
eststo m1: estpost tabstat `covars', `stats'
eststo m2: estpost tabstat `covars' if complier == 1, `stats'
        summarize complier, meanonly
        local pc_comply = string(100 * r(mean), "%4.1fc")
eststo m3: estpost tabstat `covars' if earlysplitter == 1, `stats'
        summarize earlysplitter, meanonly
        local pc_earlysplit = string(100 * r(mean), "%4.1fc")
eststo m4: estpost tabstat `covars' if overshooter == 1, `stats'
        summarize overshooter, meanonly
```

```
        local pc_overshoot = string(100 * r(mean), "%4.1fc")
esttab m* using "$tex/tab_4_characteristics.tex", replace booktabs nonumber noobs label ///
        cells("mean(fmt(%6.1fc)) sd(fmt(%6.1fc))") ///
        collabels("Mean" "SD") ///
        mgroups("All" "Compliers (`pc_comply'\%)" ///
                "Early Splitters (`pc_earlysplit'\%)" ///
                        "Overshooters (`pc_overshoot'\%)", ///
                pattern(1 1 1 1) span prefix(\multicolumn{@span}{c}{) suffix(})) ///
        title("Characteristics of Schools by Rule Compliance")



********************************************************************************
**# Part 5
********************************************************************************

local test_score_vars avgmath avgverb
local bandwidths 2 5 8 10 15 20 25 30
local poly_degs  0 1 2 3 4

foreach test_score of varlist `test_score_vars' {

        use "../intermediate/al_1999_data.dta", clear
        gen above_cutoff_1 = (c_size > 40)

        local test_score_lbl : var label `test_score'

        tempname p
        tempfile f
        postfile `p' bw deg b se using `f'
        foreach bw of local bandwidths {
                gen in_bw_`bw' = (c_size >= 40 - `bw') & (c_size <= 40 + `bw')

                foreach d of local poly_degs {

                        // Create the polynomial terms if not already there
                        local c_size_poly
                        forvalues i = 1/`d' {
                                cap gen c_size_`i' = c_size^`i'
                                local c_size_poly `c_size_poly' c_size_`i'
                        }

                        // Run IV regression
                        ivregress 2sls `test_score' `c_size_poly' (classize = above_cutoff_1) ///
                                if in_bw_`bw' == 1, vce(cluster schlcode)
                        post `p' (`bw') (`d') (_b[classize]) (_se[classize])
                }
        }
        postclose `p'
        use `f', clear
```

```
        rename b est1
        rename se est2
        reshape long est, i(bw deg) j(etype)
        reshape wide est, i(deg etype) j(bw)

        format est* %13.2fc
        tostring deg est*, replace force usedisplayformat
        foreach var of varlist est* {
                replace `var' = "(" + `var' + ")" if etype == 2 & !missing(`var')
                local lbl : var label `var'
                local lbl = subinstr("`lbl'", " est", "", .)
                local lbl = "BW = `lbl'"
                label var `var' "`lbl'"
        }
        replace deg = "" if etype == 2
        drop etype
        label var deg "Polynomial Degree"
        texsave * using "$tex/tab_5_`test_score'.tex", varlabels frag replace location("H") ///
                title("Fuzzy RD Estimates of of Class Size on `test_score_lbl'")

}


********************************************************************************
**# Part 6
********************************************************************************

forvalues g = 4/5 {

        use "../intermediate/al_1999_data.dta", clear
        contract schlcode c_size if grade == `g'
        isid schlcode
        hist c_size, discrete freq ///
                addplot(pci 0 40 20 40, lcolor(red) ///
                        || pci 0 80 20 80, lcolor(red) ///
                        || pci 0 120 20 120, lcolor(red)) ///
                legend(off) title("`g'th Grade Enrollment") ///
                xtitle("") ///
                saving("../output/hist_`g'.gph", replace)

        use "../intermediate/al_1999_data.dta", clear
        contract schlcode c_size if grade == `g'
        isid schlcode
        DCdensity c_size, breakpoint(41) generate(Xj Yj r0 fhat se_fhat) b(1)
        graph save "../output/mccrary_`g'.gph", replace

}

graph combine "../output/hist_5.gph" "../output/hist_4.gph" ///
        "../output/mccrary_5.gph" "../output/mccrary_4.gph", rows(2) cols(2)
```

```
graph export "$tex/fig_6.png"


********************************************************************************
**# Part 7
********************************************************************************

use "../intermediate/al_1999_data.dta", clear
gen pred_num_classes = int((c_size - 1) / 40) + 1
gen pred_class_size = c_size / pred_num_classes
gen above_cutoff_1 = (c_size > 40)
gen c_size_2_d100 = (c_size^2) / 100

local grades 5 4
local donut_widths 1 2 3

tempname p
tempfile f
postfile `p' grade donut est_type lang1 lang2 math1 math2 using `f'
foreach g of local grades {
        foreach dw of local donut_widths {

                cap gen in_donut_`dw' = (c_size < 40 - `dw') | (c_size > 40 + `dw')

                ivregress 2sls avgverb tipuach c_size (classize = pred_class_size) ///
                                if in_donut_`dw' == 1 & grade == `g', vce(cluster schlcode)
                        local l1_b  =  _b[classize]
                        local l1_se = _se[classize]

                ivregress 2sls avgverb tipuach c_size c_size_2_d100 (classize = pred_class_size) ///
                                if in_donut_`dw' == 1 & grade == `g', vce(cluster schlcode)
                        local l2_b  =  _b[classize]
                        local l2_se = _se[classize]

                ivregress 2sls avgmath tipuach c_size (classize = pred_class_size) ///
                                if in_donut_`dw' == 1 & grade == `g', vce(cluster schlcode)
                        local m3_b  =  _b[classize]
                        local m3_se = _se[classize]

                ivregress 2sls avgmath tipuach c_size c_size_2_d100 (classize = pred_class_size) ///
                                if in_donut_`dw' == 1 & grade == `g', vce(cluster schlcode)
                        local m4_b  =  _b[classize]
                        local m4_se = _se[classize]

                post `p' (`g') (`dw') (0) (`l1_b')  (`l2_b')  (`m3_b')  (`m4_b')
                post `p' (`g') (`dw') (1) (`l1_se') (`l2_se') (`m3_se') (`m4_se')
        }
}
postclose `p'
use `f', clear
```

```
gen donut_lower = 40 - donut
gen donut_upper = 40 + donut
format lang* math* %13.4fc
tostring grade donut* lang* math*, replace force usedisplayformat
foreach var of varlist lang* math* {
        replace `var' = "(" + `var' + ")" if est_type == 1 & !missing(`var')
}
gen donut_int = "[" + donut_lower + "," + donut_upper + "]"
foreach var of varlist grade donut_int {
        replace `var' = "" if est_type == 1
}
drop donut donut_lower donut_upper est_type
order grade donut_int

// Manually enter notes to append to end of table
preserve
        clear
        input str200 grade str200 donut_int str200 lang1 str200 lang2 str200 math1 str200 math2
                "Controls:"                "" ""  ""  ""  ""
                "Percent Disadvantaged"    "" "X" "X" "X" "X"
                "Enrollment"               "" "X" "X" "X" "X"
                "Enrollment Squared /100" "" ""  "X" ""  "X"
        end
        tempfile end_of_table
        save `end_of_table', replace
restore
append using `end_of_table'

label var grade "Grade"
label var donut_int "Donut"
label var lang1 "Language (1)"
label var lang2 "Language (2)"
label var math1 "Math (3)"
label var math2 "Math (4)"
texsave * using "$tex/tab_7.tex", varlabels frag replace location("H") hlines(6 12) ///
        title("Replication of Table A6")
```