Lab #8 - Class Size and Test Scores

Econ 224

September 20th, 2018

Angrist and Lavy (1999)

This lab is adapted from one of Josh Angrist's problem set questions for 14.32 at MIT.

The Angrist data archive https://economics.mit.edu/faculty/angrist/data1/data/anglavy99 contains data from the article "Using Maimonides Rule to estimate the Effect of Class Size on Student Achievement" by Angrist & Lavy, published in the *Quarterly Journal of Economics*, May 1999. This article uses the fact that Israeli class sizes are capped at 40 to estimate the effects of class size on test scores. We have not yet studied the methods used in the paper, so in this lab we'll examine the dataset using linear regression. The dataset we'll examine is `final.dta` which contains data for 5th grade classes:

Name	Description
`c_size`	September grade enrollment at the school
`classize`	class size: number of students in class in the spring
`tipuach`	percent of students in the school from disadvantaged backgrounds
`avgverb`	average composite reading score in the class
`avgmath`	average composite math score in the class
`mathsize`	number of students who took the math test
`verbsize`	number of students who took the reading test

Note: you do *not* have to use robust standard errors in this lab, although you are welcome to do so if you wish.

1. Load and clean the dataset:

   (a) Download the file `final5.dta` from the Angrist Data Archive at the url listed above and save it on your machine. (This file contains data for 5th graders.)

   (b) Read `final5.dta` into an R dataframe called `final5.dta` using the function `read.dta` from the package `foreign`. (This file was created with an old version of STATA and for mysterious reasons does not load correctly using `read_csv` from `readr`.)

   (c) Convert `final5` to a tibble using the function `as_tibble` from `dplyr`.

   (d) Look up the `dplyr` function `rename`. Once you understand how it works, use it to re-name `c_size` to `enroll`, and `tipuach` to `pdis`.

   (e) Use `dplyr` to restrict `final5` so that it contains only observations for schools with 5th grade enrollment of at least 5 students, and classrooms with fewer than 45 students.

   (f) Select only the columns we will use later in the analysis: `classize`, `enroll`, `pdis`, `verbsize`, `mathsize`, `avgverb`, and `avgmath`.

   (g) There was a data entry error for one value of `avgmath`: `181.246` should be `81.246` since the test score is out of `100`. Correct this.

   (h) There was a data entry error for one value of `avgread`: `187.606` should be `87.606` since the test score is out of `100`. Correct this.

   (i) There is a classroom with `mathsize` equal to zero, i.e. no students in this class took the math test, which has a *non-missing* value for `avgmath`. This is an error: since no one in this class took the test, there is no average math score for this class. Replace all values of `avgmath` for classes with `mathsize` equal to zero with `NA`.

2. Create a table of descriptive statistics:

   (a) Download the Angrist & Lavy paper and consult Table I on page 539.
   (b) Use `stargazer` to replicate the top panel of Table I, i.e. the panel with information on 5th grade classes. You do not have to display the 10th and 90th percentiles of the data: the quartiles, mean, and standard deviations are sufficient.

3. Regress achievement on class size:

   (a) Carry out a regression predicting average verbal scores from class size. Create a nicely formatted table of results using the package of your choice: `stargazer` or `texreg`.
   (b) Repeat part (a) but predict average math test scores.
   (c) Discuss your results from (a) and (b). If smaller classes improve student achievement, what sign should the coefficient estimates from your regression have? What kind of relationship do you find? Is it large enough to be of practical importance? Statistically significant?

4. Control for school size:

   (a) A possible explanation for your findings in question 3 is that larger schools have larger classes *and* better students. Repeat question 3 but add `enrollment`, which measures the size of the 5th grade at the school, to your regressions. How do the results change? Combine the results for all four regressions into a single table to make it easier to compare them.

5. Construct the correlation matrix of math test scores, class size, and enrollment. Use this matrix and your regression results to explain why and how the coefficient on class size changes when you control for enrollment.

6. Control for percent disadvantaged:

   (a) Repeat question 4 but add the percent of students who came from disadvantaged backgrounds `pdis` in place of enrollment. How does this affect the results?
   (b) Calculate the correlation matrix for math test scores, class size, and `pdis`. Using this information along with the correlation matrix from question 5 above and your regression results, why does controlling for `pdis` have a larger effect on the estimated coefficient for class size than controlling for `enroll`?

7. Regress math and verbal test scores on class size controlling for *both* `pdis` and `enroll`. Discuss your results in light of questions 4-6 above. All told, do your results for this dataset suggest that smaller classes are good, bad, or neutral?

Solutions

1 - Load and Clean the Data

```
library(tidyverse)
library(foreign)
final5 <- read.dta('~/econ224/labs/final5.dta')
final5 <- as_tibble(final5)
final5 <- final5 %>%
  rename(enroll = c_size, pdis = tipuach) %>%
  filter((enroll >= 5) & (classize < 45)) %>%
  select(classize, enroll, pdis, verbsize, mathsize, avgverb, avgmath)
math_error <- which(final5$avgmath > 100)
```

```
final5$avgmath[math_error] <- 81.246
verbal_error <- which(final5$avgverb > 100)
final5$avgverb[verbal_error] <- 187.606
change_to_NA <- which(final5$mathsize == 0)
final5$avgmath[change_to_NA] <- NA
final5
```

```
## # A tibble: 2,025 x 7
##    classize enroll  pdis verbsize mathsize avgverb avgmath
##       <int>  <int> <int>    <int>    <int>   <dbl>   <dbl>
##  1       28     54    24       28       28    70.6    74.1
##  2       26     54    24       27       27    75      71.1
##  3       22     37    38       15       15    75.5    64
##  4       15     37    38       20       20    60.6    50
##  5       32     32     6       32       32    74.0    68.4
##  6       34     68     3       22       22    69.6    59.9
##  7       34     68     3       30       31    68.1    61.9
##  8       30     86     8       30       30    65.6    61.1
##  9       26     86     8       24       24    69.9    59.4
## 10       31     86     8       29       27    73.1    66.4
## # ... with 2,015 more rows
```

2 - Create Table of Descriptive Statistics

```
library(stargazer)
stargazer(as.data.frame(final5),
          type = 'latex',
          title = 'Unweighted Descriptive Statistics',
          digits = 1,
          header = FALSE,
          covariate.labels = c('Class size',
                               'Enrollment',
                               'Percent disadvantaged',
                               'Reading size',
                               'Math size',
                               'Average verbal',
                               'Average math'),
          summary.stat = c('mean',
                           'sd',
                           'p25',
                           'median',
                           'p75'))
```

3 through 7 - Various Regressions

```
reg1v <- lm(avgverb ~ classize, final5)
reg2v <- lm(avgverb ~ classize + enroll, final5)
reg3v <- lm(avgverb ~ classize + pdis, final5)
```

Table 2: Unweighted Descriptive Statistics

Statistic	Mean	St. Dev.	Pctl(25)	Median	Pctl(75)
Class size	29.9	6.6	26	31	35
Enrollment	77.9	39.1	50	72	100
Percent disadvantaged	14.1	13.5	4	10	20
Reading size	27.3	6.6	23.0	28.0	32.0
Math size	27.7	6.7	23.0	28.0	33.0
Average verbal	74.4	8.1	69.8	75.4	79.8
Average math	67.3	9.6	61.1	67.8	74.1

```r
reg4v <- lm(avgverb ~ classize + enroll + pdis, final5)

reg1m <- lm(avgmath ~ classize, final5)
reg2m <- lm(avgmath ~ classize + enroll, final5)
reg3m <- lm(avgmath ~ classize + pdis, final5)
reg4m <- lm(avgmath ~ classize + enroll + pdis, final5)

stargazer(reg1v, reg2v, reg3v, reg4v,
          reg1m, reg2m, reg3m, reg4m,
          type = 'latex',
          header = FALSE,
          digits = 2,
          covariate.labels = c('Class Size', 'Enrollment', 'Percent Disadvantaged'),
          dep.var.labels = c('Verbal', 'Math'),
          title = 'Test Scores Regression Controlling for Enrollment',
          omit.stat = c('f', 'ser', 'adj.rsq'))
```

Table 3: Test Scores Regression Controlling for Enrollment

	Dependent variable:							
	Verbal				Math			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Class Size	0.22^{***}	0.12^{***}	-0.03	-0.03	0.32^{***}	0.16^{***}	0.07^{**}	0.01
	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)
Enrollment		0.03^{***}		-0.001		0.04^{***}		0.02^{***}
		(0.01)		(0.005)		(0.01)		(0.01)
Percent Disadvantaged			-0.35^{***}	-0.35^{***}			-0.34^{***}	-0.33^{***}
			(0.01)	(0.01)			(0.01)	(0.01)
Constant	67.89^{***}	68.75^{***}	80.24^{***}	80.23^{***}	57.79^{***}	59.19^{***}	69.94^{***}	70.22^{***}
	(0.83)	(0.84)	(0.81)	(0.81)	(0.97)	(0.99)	(1.01)	(1.02)
Observations	2,020	2,020	2,020	2,020	2,019	2,019	2,019	2,019
R^2	0.03	0.04	0.32	0.32	0.05	0.07	0.25	0.25
Note:								*p<0.1; **p<0.05; ***p<0.01

```
final5 %>%
  summarize(SD_verb = sd(avgverb, na.rm = TRUE),
            SD_math = sd(avgmath, na.rm = TRUE),
            SD_classize = sd(classize, na.rm = TRUE))
```

```
## # A tibble: 1 x 3
##   SD_verb SD_math SD_classize
##     <dbl>   <dbl>       <dbl>
## 1    8.08    9.60        6.57
```

```
final5 %>%
  select(avgmath, enroll, classize) %>%
  cor(use = 'complete.obs') %>%
  round(2)
```

```
##          avgmath enroll classize
## avgmath     1.00   0.24     0.22
## enroll      0.24   1.00     0.62
## classize    0.22   0.62     1.00
```

```
final5 %>%
  select(avgmath, pdis, classize) %>%
  cor(use = 'complete.obs') %>%
  round(2)
```

```
##          avgmath  pdis classize
## avgmath     1.00 -0.49     0.22
## pdis       -0.49  1.00    -0.35
## classize    0.22 -0.35     1.00
```