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, namely instrumental variables and regression discontinuity, 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. To be clear, each observation in the dataset is a *5th grade class* rather than an individual student. The variables we will use in the lab are as follows:

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

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. First regression:

    (a) Economists and educators have long debated whether it's worth paying the extra labor costs (i.e., teachers' wages) required to reduce class size. What should the sign of the achievement/class-size relationship be if the investment is worthwhile? Regress average math and verbal scores on class size. What is the sign of this relationship? Is it significantly different from zero? How does it look so far for the class size optimists?
    (b) Weighted regression!

4. A possible concern with the bivariate regression of test scores on class size is that bigger schools have bigger classes and also better students. Check this by adding enrollment (in the grade, measured at the school level) to your regressions.

5. Construct the correlation matrix of test scores (math or verbal, your choice), class size, and enrollment. Use this matrix and your regression results to explain why and how the coefficient on class size changes when you add enrollment controls. (Hint: This requires an application of the omitted variables bias formula)

6. Add the percent of students who came from disadvantaged backgrounds (PD) instead of enrollment. How does this affect the class size coefficient? Use the correlation matrix for test scores, class size, and PD, along with the correlation matrix in Part (c) above and your regression results, to explain why the coefficient changes more here than when you added enrollment in response to Part (?).

7. Estimate the effects of class size in a model that includes both PD and enrollment controls. Is the class size coefficient here closer to the one in Part (?) or Part (?)? Why?

8. All told, does the analysis in this question suggest that smaller classes are good, bad, or neutral?

Solutions

1 - Load and Clean the Data

```r
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

```r
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'))
```

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