Case Study: Do Smaller Class Sizes Improve Test Scores? (OLS on CASchool)

Lin_Yu

2025-09-27

Project: Do Smaller Class Sizes Improve Student Test Scores?

This case study uses the CASchool dataset to examine whether smaller class sizes improve student test scores. I begin with exploratory analysis and a simple OLS regression, then extend the model by adding income, poverty, and English learner controls to address omitted variable bias. The results show that smaller student–teacher ratios are consistently associated with higher scores, though socioeconomic factors also play an important role. I conclude by discussing model fit, limitations of OLS, and potential future approaches such as nonlinear models and causal inference designs.

Research Question:

Does lowering the student–teacher ratio (STR) improve academic performance in California elementary schools?

Data

Dataset: CASchool.Rdata (California school districts). Outcome (Y): test.score (average reading+ math scores). Main regressor (X): str (student–teacher ratio). Controls: - income: average district income - meal: % eligible for subsidized lunch (proxy for poverty) - english: % English learners (control)

```
# load our dataset and create variables needed in our potential model
# install.packages("AER")  # if needed
library(AER)
```

```
## Warning: package 'AER' was built under R version 4.4.1
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: lmtest
```

```
## Warning: package 'lmtest' was built under R version 4.4.1
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.4.1
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Warning: package 'sandwich' was built under R version 4.4.1
```

```
## Loading required package: survival
```

```
library(lmtest)     # for coeftest
library(sandwich)  # for vcovHC
library(car)       # for linearHypothesis

data("CASchools", package = "AER")
df <- CASchools
names(df)
```

```
##  [1] "district"    "school"      "county"      "grades"      "students"
##  [6] "teachers"    "calworks"    "lunch"       "computer"    "expenditure"
## [11] "income"      "english"     "read"        "math"
```

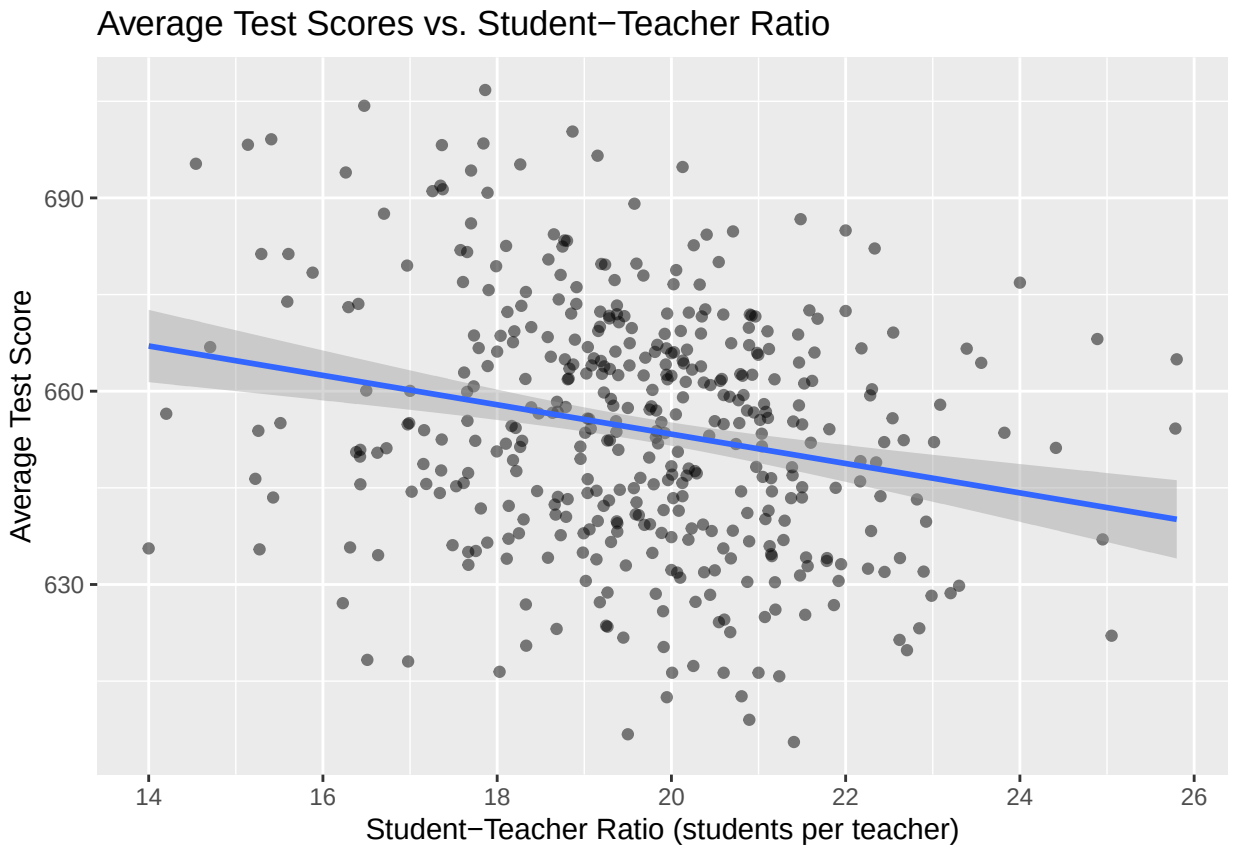
```
# Create STR (student-teacher ratio) and test score;
df$STR <- df$students / df$teachers
df$test.score<- (df$math+df$read)/2
```

##Exploratory Data Analysis (EDA)

```
library(ggplot2)

# Scatter of test scores vs STR with a linear fit
ggplot(df, aes(x = STR, y = test.score)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    title = "Average Test Scores vs. Student-Teacher Ratio",
    x = "Student-Teacher Ratio (students per teacher)",
    y = "Average Test Score"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



###Observations: The scatterplot of average test scores against the student–teacher ratio (STR) shows a downward trend: districts with smaller class sizes (lower STR) tend to achieve higher test scores, while those with larger class sizes generally have lower scores. Though the relationship is not perfectly linear and there is noticeable variation across districts, the fitted regression line confirms a negative correlation between STR and performance.

OLS Models:

Bivariate OLS

```
library(lmtest)     # for coeftest
library(sandwich)   # for vcovHC
library(car)        # for linearHypothesis

Y   <- df$test.score
X1  <- df$STR

model_1 <- lm(Y ~ X1, data = df)
coeftest(model_1, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 698.93295   10.36436 67.4362 < 2.2e-16 ***
## X1           -2.27981    0.51949 -4.3886 1.447e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion: If we assume the student–teacher ratio (STR) is the only factor affecting student performance and its relationship with student performance is linear, we can run a simple linear regression. The t-test results show that the coefficient on STR is –0.915 with a p-value < 0.01, indicating strong statistical significance. This means we reject the null hypothesis that class size has no effect on test scores. The negative coefficient implies that smaller class sizes are associated with higher student achievement, which is consistent with the downward trend observed in the ggplot visualization.

Multivariate Linear Regressions:

However, there are many important socioeconomic and demographic differences across school districts that may independently influence student achievement, such as wealth levels and language barriers. For example, higher household income is often associated with greater educational resources and stronger performance. Eligibility for subsidized meal programs reflects poverty levels, which can hinder learning outcomes. Likewise, a higher proportion of English learners may lower average test scores due to language challenges. By controlling for these factors, we can better isolate the effect of class size (STR) on student performance and reduce omitted variable bias in our regression. By assuming all these varibles' influences on test.core is linear, we use the following model:

\[\text{test.score}_i = \beta_0 + \beta_1 str_i + \beta_2 income_i + \beta_3 meal_i + \beta_4 english_i + u_i \]

```
Y   <- df$test.score
X1  <- df$STR
X2 <- df$income
X3  <- df$lunch
X4 <- df$english

model_2 <- lm(Y ~ X1 + X2 + X3 + X4, data = df)
coeftest(model_2, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 675.608208   6.201868 108.9363 < 2.2e-16 ***
## X1           -0.560389   0.255064  -2.1971   0.02857 *
## X2            0.674984   0.083716   8.0628 8.058e-15 ***
## X3           -0.396366   0.030230 -13.1116 < 2.2e-16 ***
## X4           -0.194328   0.033245  -5.8454 1.024e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion: The estimated effect of STR decreases from –0.915 in the bivariate model to –0.373 after adding controls, yet the p-value remains well below 0.01, confirming that the negative relationship holds. We also find that socioeconomic and demographic factors significantly influence student performance: districts with higher average income tend to achieve better test results, while higher poverty rates (measured by subsidized lunch eligibility) are associated with lower scores. Likewise, a larger proportion of English learners is linked to lower average performance, highlighting the challenges faced by these students.

Model Fitness

```
#install.packages("kableExtra")

library(broom)       # for glance()
library(dplyr)       # for bind_rows, mutate, select, pipes
```

Table 1: Model Fit Comparison

model	r.squared	adj.r.squared	sigma	AIC	BIC
Bivariate OLS	0.051	0.049	18.581	3650.499	3662.620
Multivariate OLS	0.805	0.803	8.448	2991.352	3015.593

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(kableExtra)  # for kbl() and kable_classic()
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
bind_rows(
  glance(model_1) %>% mutate(model = "Bivariate OLS"),
  glance(model_2) %>% mutate(model = "Multivariate OLS")
) %>%
  select(model, r.squared, adj.r.squared, sigma, AIC, BIC) %>%
  kbl(digits = 3, caption = "Model Fit Comparison") %>%
  kable_classic(full_width = FALSE)
```

Discussion: Our multivariate OLS model shows improved fit relative to the simple bivariate regression, with higher explanatory power and lower information criteria. This suggests that accounting for socioeconomic and demographic controls provides a clearer picture of student performance. However, OLS relies on strong assumptions, and the linear specification may oversimplify the relationship between class size and achievement. Future work could explore nonlinear effects, such as diminishing returns to smaller classes, or examine whether the impact of class size varies across districts with different poverty or language profiles. More advanced approaches like robust regression, hierarchical (multilevel) modeling, or causal designs such as instrumental variables, regression discontinuity, or difference-in-differences would provide stronger evidence on whether smaller classes truly cause better outcomes.