Econometrics 1: Case Study 1

Determinants of Testscores in Public Schools

Bischoy Bert

November 2025

1 Introduction

A classic topic in Economics concerns estimating the "returns to education", that is, the causal effect of education on labor market earnings. A related question is what properties of a school are predictive for their students' achievements, measured in terms of standardized test score outcomes. Understanding the determinants of being a "good school" would help to improve individual decision making, and would be informative for economic policy. We shall investigate this question in the present case study.

As for this analysis, we shall only be interested in public schools. We use the dataset `CASchools_EE141_InSample.xlsx` from the Stock and Watson textbook resources, available at `https://www.princeton.edu/~mwatson/Stock-Watson_4E/Stock-Watson-Resources-4e.html`. The dataset has been edited to include only public schools. For our purposes, we shall call it `CASchool`.

```
CASchool <- read_xlsx("CASchools_EE141_InSample.xlsx") |>
  filter(charter_s == 0)
```

2 Data Acquisition

According to the description of the dataset, the variables are defined as follows:

- `testscore`: The test scores for each school as a sum of math and English/language arts scores for 5th grade students
- `str_s`: The student-to-teacher ratio (full-time equivalent, FTE)
- `med_income_z`: The median income for the 15+ population per zip code

We expect both `str_s` and `med_income_z` to be negatively and positively correlated with `testscore`, respectively. A lower student-to-teacher ratio implies fewer students per teacher, which should facilitate a more personalized learning experience, thereby improving learning outcomes. Furthermore, higher median income should correlate positively with test scores, as higher-income families typically have greater resources to support academic achievement and, assuming a correlation between education and income, may place stronger emphasis on academic success.
Following, we will compute the empirical correlation between both `med_income_z`, `str_s` and `testscore`, respectively. To do this, we shall use the `R` base function `cor()`.

```
CASchool |>
  summarise(correlation_test_income = cor(testscore, med_income_z))
```

```
## # A tibble: 1 x 1
##   correlation_test_income
##                     <dbl>
## 1                   0.595
```

The computation indicates a strong positive correlation between testscores and median personal income and a slightly negative correlation between testscores and student-to-teacher ratios. These results supports our previously stated arguments albeit we expected a stronger (negative) correlation between the latter.

3 Descriptive Statistics

In this section, we shall create a histogram to display the distribution of testscores in schools.

```
histogram <- CASchool |>
  ggplot(aes(x = testscore)) +
  geom_histogram(binwidth = 50, fill = "lightblue", color = "white") +
  labs(
    title = "Distribution of Test Scores",
    x = "Average Testscore",
    y = "Frequency"
  )
histogram
```

Next, we want to determine which school has the highest testscores and which school the lowest.

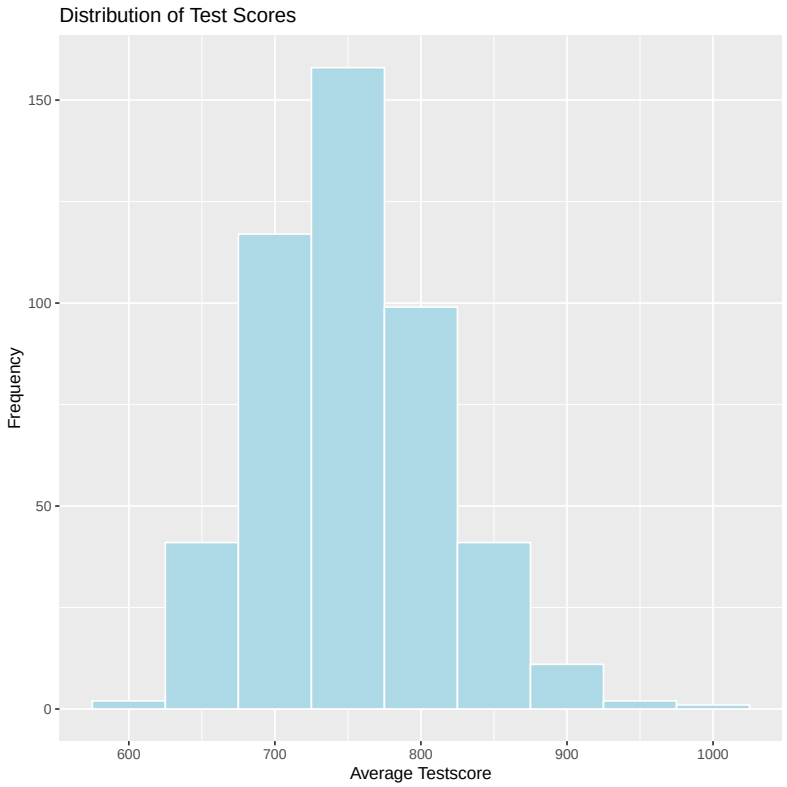

Figure 1: Histogram for Testscores

```
CASchool |>
  filter(testscore == max(CASchool$testscore) |
           testscore == min(CASchool$testscore)) |>
  select(zipcode, schoolname, testscore) |>
  arrange(testscore)
```

```
## # A tibble: 2 x 3
##   zipcode schoolname               testscore
##     <dbl> <chr>                        <dbl>
## 1   92281 Westmorland Elementary        602.
## 2   95135 Tom Matsumoto Elementary      981.
```

Next, we want to determine the values for the variables `te_salary_avg_d`, `st_ratio`, and `med_income_z` for those two schools and interpret the results.

```
result1 <- CASchool |>
  filter(schoolname %in% c("Tom Matsumoto Elementary", "Westmorland Elementary")) |>
  select(schoolname, te_salary_avg_d, str_s, med_income_z, testscore) |>
  arrange(testscore)
result1
```

```
## # A tibble: 2 x 5
##   schoolname               te_salary_avg_d str_s med_income_z testscore
##   <chr>                              <dbl> <dbl>        <dbl>     <dbl>
## 1 Westmorland Elementary             59715  15.8        15000      602.
## 2 Tom Matsumoto Elementary           80971  25.4        51556      981.
```

As we can see, Tom Matsumoto Elementary has significantly higher average teacher salaries, a significantly higher median income within its district but also a higher student-to-teacher ratio than Westmorland Elementary. This aligns with our assumption that median income (and hence teacher salaries) positively correlate with testscores. However, the higher student to teacher ratio appears counterintuitive. It may be the case that a possible income effect heavily outweighs any higher student to teacher ratio, although we might want to emphasize that the negative correlation was very small to begin with.

4 Plots

For this section, we shall create a new variable and add it to our dataset. We will call the new variable `group_strs` and define it as follows:

- equals -1 if an observation's `str_s` value is strictly smaller than the 20% quantile of the variable `str_s`,

- equals 1 if an observation's `str_s` value is strictly larger than the 80% quantile of the variable `str_s`,
- and equals 0 otherwise.

```
CASchool <- CASchool |>
  mutate(
    group_strs = case_when(
      str_s < quantile(str_s, 0.2, na.rm = TRUE) ~ -1,
      str_s > quantile(str_s, 0.8, na.rm = TRUE) ~ 1,
      TRUE ~ 0
    )
  )
```

Next, we shall create three boxplots displaying the distributions of testscores within each defined quantile of student to teacher ratio.

5 Scatterplot

For this section we shall take the natural logarithm of `med_income_z` and create a new variable `lmed_income` which we will add to our `CASchool` dataset.

```
CASchool <- CASchool |>
  mutate(
    lmed_income = log(med_income_z)
  )
```

Now, we will generate a scatterplot illustrating the relationship between `testscore` and our newly created variable `lmed_income`

```
scatterplot <- CASchool |>
  ggplot() +
  geom_point(aes(x = lmed_income, y = testscore))
```

The plot indicates a strong, linear relationship between `lmed_income` and `testscore`. Next, we will try and fit a linear model through our scatterplot by guessing a reasonable regression line. To do this, the intercept was chosen to be `-1300` and the slope `200`.

```
scatterplot_abline_guess <- scatterplot +
  geom_abline(intercept = -1300, slope = 200, color = "red")
scatterplot_abline_guess
```

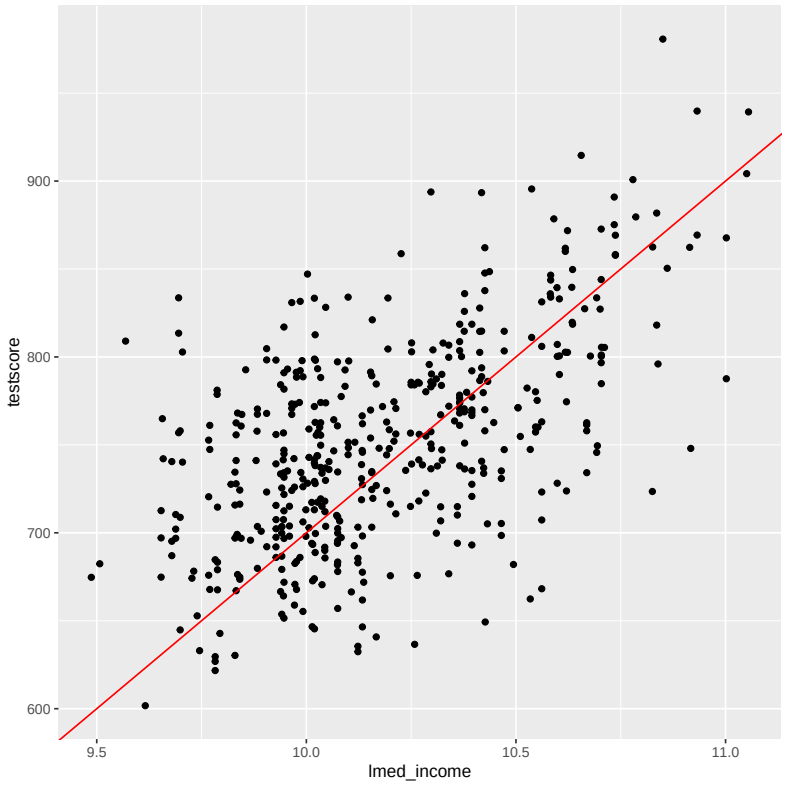

Figure 2: Scatterplot with our guessed regression line

6 OLS Estimation and Illustration

In this section we shall fit a linear regression of the form

$$Y = \beta_0 + \beta_1 \log(X) + u, \quad E[u \mid X] = 0$$

with `testscore` as Y and `lmed_income` as X.

```
model <- lm(data = CASchool, testscore ~ lmed_income)
```

The estimated intercept is -396.8125783, and the estimated coefficient for `lmed_income` is 112.8072619. The graphic below shows the computed regression line above our previously created scatterplot.

```
scatterplot_abline <- scatterplot +
  geom_abline(intercept = model$coefficients["(Intercept)"],
              slope = model$coefficients["lmed_income"], color = "blue")
scatterplot_abline
```

Next, we shall predict `testscore` for a new observation with `med_income_z` == `22026.47` by: (i) reading the graph, (ii) manually using the estimated model equation, and (iii) using the `R` function `predict`.

(i) By reading the graph for `lmed_income` $= 10$ ($\log(22026.47)$), we can roughly estimate our testscore variable to be 750.

(ii) To manually calculate the exact predicted value, we plug in $\log(22026.47)$ into variable X of our linear model.

$$Y = -396.8 + 112.8 \times \log(22026.47) + u = 731.2 + u, \quad E[u \mid X] = 0$$

(iii) Using the `predict` function.

```
predict(model, newdata = data.frame(lmed_income = log(22026.47)))

##        1
## 731.2601
```

Next, we want to see what the expected score is if `lmed_income` decreases by 0.5.

```
predict(model, newdata = data.frame(lmed_income = log(22026.47) - 0.5))

##        1
## 674.8564
```

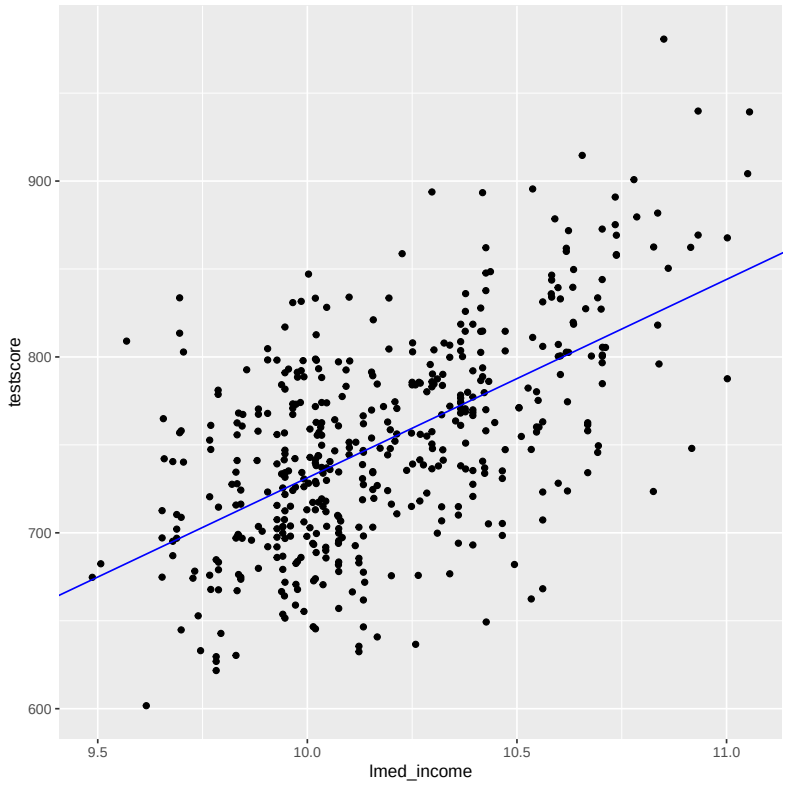

Figure 3: Scatterplot with the calculated regression line

7 Residual Plot

In this section we shall plot the residuals from the above model against the values in the regressor `lmed_income`. Using residuals as proxies of error terms, we will discuss whether the standard assumptions for the OLS appear to be satisfied or not.

```
CASchool <- CASchool |>
  mutate(residuals = residuals(model))

residuals_plot <- CASchool |>
  ggplot(aes(x = lmed_income, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(
    x = "Log Median Income",
    y = "Residuals",
    title = "Residuals for Test Scores vs Log Median Income"
  )
residuals_plot
```

Now we shall focus on the model specifications and the assumption of homoskedasticity. To do this, we shall call a diagnostic function `plot` in `R`.

```
plot(model, which = 1)
```

The residual plot shows a slightly u-shaped pattern rather than random scatter around zero. This violates the zero conditional mean assumption E[u] = 0, indicating model misspecifications thus suggesting that the relationship between `testscore` and `lmed_income` is not exactly linear. Homoskedasticity appears satisfied - the spread of residuals is roughly constant across fitted values. Given the u-shaped curve, it might be worthwhile to consider adding a quadratic term in `lmed_income` to capture potential curvature in the relationship between income and test scores. This could be modeled as:

$$Y = \beta_0 + \beta_1 \times \log(X) + \beta_2 \times \log(X)^2 + u, \quad E[u \mid X] = 0$$

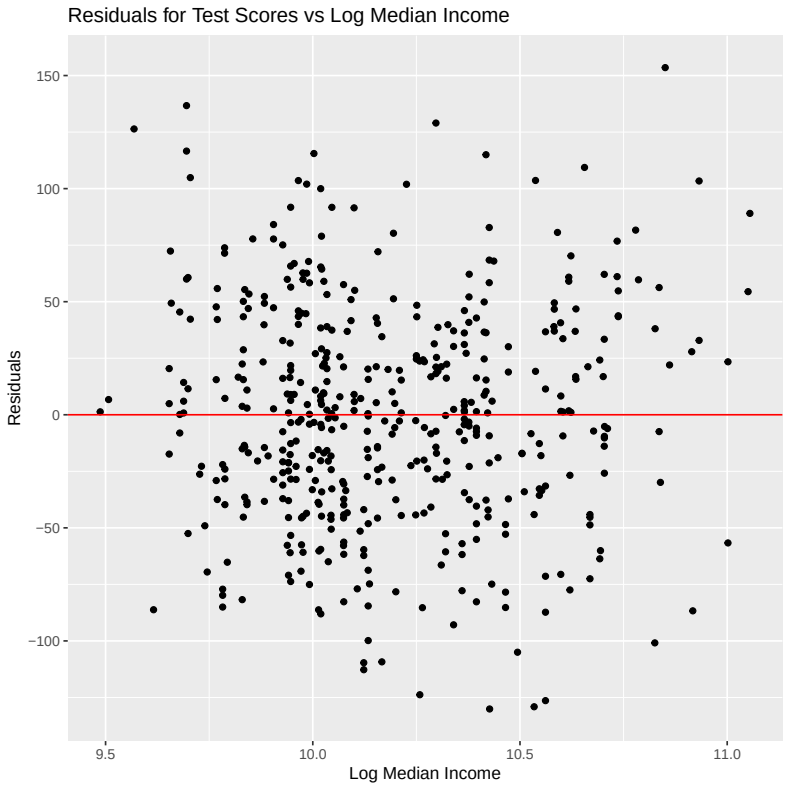

Figure 4: Residuals Plot

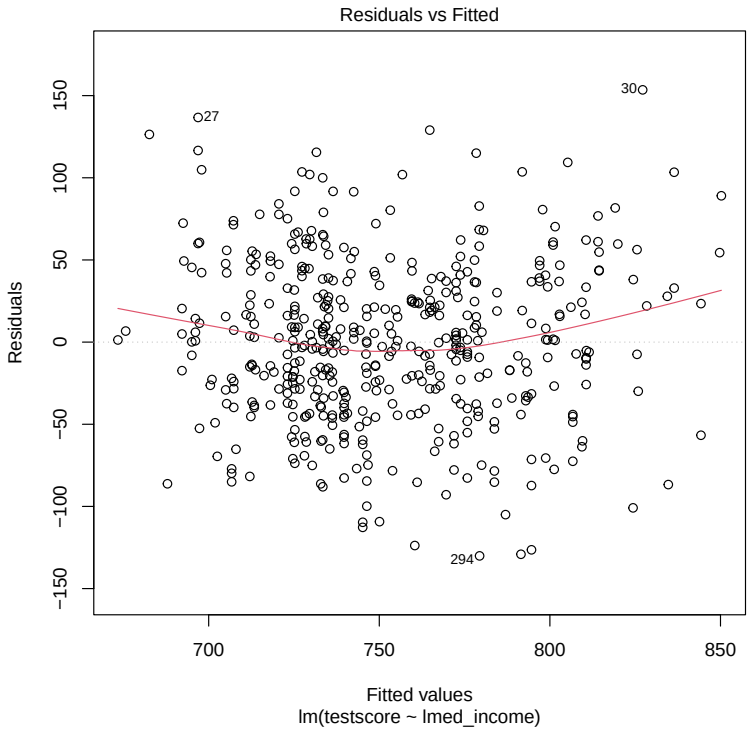

Figure 5: Residuals vs Fitted Plot