Case Study 1: Econometrics I / Ökonometrie I

WS 2025/2026

Deadline: October 20, 2025@ 23:59

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 standardised test score outcomes). Understanding the determinants of being a "good school" would help to improve individual decsion making, and would be informative for economic policy. We shall investigate this question in the present case study.

1 Data aquisition (2 points)

Download the zip-folder "CA_Schools_EE14.zip" from https://www.princeton.edu/~mwatson/Stock-Watson_4E/Stock-Watson-Resources-4e.html and extract the file "CASchools_EE141_InSample.xlsx".

Next, load that excel file into R using the function `read_xlsx` from the package **readxl** (after installing that package via `install.packages("readxl")`). We shall actually only be interested in "non-charter" (i.e., "public") schools. Select the relevant schools making use of the variable `charter_s` (0 corresponds to public school), and store the so-obtained sub-dataset in `CAschool`.

Consulting the data documentation "CASchools_EE141_Description.pdf" in the zip-file you downloaded, answer the question what the variables `testscore`, `str_s`, and `med_income_z` measure.

Briefly provide an argument supporting your believe whether (or not) high values in `med_income_z` or `str_s` should typically occur with high values in `testscore`.

Compute the empirical correlation between `testscore` and `med_income_z`. Does this support your argument?

Compute the empirical correlation between `testscore` and `str_s`. Does this support your argument?

2 Descriptive statistics (2 points)

Create a nice histogram (with suitable title, x-axis notation and in color "lightblue") for the variable `testscore`.

Which school has the smallest and which school has the largest value in the variable `testscore`? Furthermore, determine the values of the variables `te_salary_avg_d`, `str_s`, and `med_income_z` for those two schools and interpret the results.

3 Plots (1 point)

Create a variable (with values -1, 0, 1) called `group_strs` that

- 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 else.

Next create three boxplots of the variable `testscore`, one for each possible outcome of `group_strs` and interpret the results.

Case Study 1: Econometrics I / Ökonometrie I

WS 2025/2026

Deadline: October 20, 2025@ 23:59

4 Scatterplot (2 points)

Take the natural logarithm of `med_income_z` and add the so-obtained data vector as an additional variable called `lmed_income` to the dataset `CAschool`.

Generate a scatterplot illustrating the relationship between `lmed_income` and `testscore` and briefly interpret the result.

Furthermore, use the function `abline` and add by **guessing** a reasonable regression line. To this end, choose the parameters for `a` and `b` the function `abline` requires (e.g., by determining the slope and intercept from guessing the values of the regression line at two points, e.g., for x = 10 and x = 11).

5 OLS estimation and illustration (2 points)

Consider fitting a simple linear regression of the form

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

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

Report the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for that model, and display the regression line (in blue) on top of a scatter plot created as in the previous exercise.

Predict the testscore for a new observation with `med_income_z = 22026.47` by: (1) reading from the graph, (2) manually using the estimated model equation, and (3) using the **R** function `predict`. What is the expected score if the value (associated to that new observation) in `lmed_income` then decreases by 0.5.

6 Residual plot (1 point)

Plot the residuals from the above model against the values in the regressor `lmed_income`. Using residuals as proxies of error terms, discuss whether the standard assumptions for the OLS appear to be satisfied or not.

Specifically, focus on the model specification and the homoskedasticity assumption. Which transformation applied to the dependent variable could be useful in this case (also inspect the second plot you get after calling `plot(reg1, which = 1)`)?