Introduction to Linear Regression
(SW Chapter 4)

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w

Brooklyn College

Empirical problem: Class size and educational output

Policy question: What is the effect of reducing class size by one student per class? by 8 students/class?

What is the right output (performance) measure?

- parent satisfaction
- student personal development
- future adult welfare
- future adult earnings
- performance on standardized tests

What do data say about class sizes and test scores?

The California Test Score Data Set

All K-6 and K-8 California school districts (n = 420)

Variables:

- 5th grade test scores (Stanford-9 achievement test, combined math and reading), district average
- Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teachers

An initial look at the California test score data

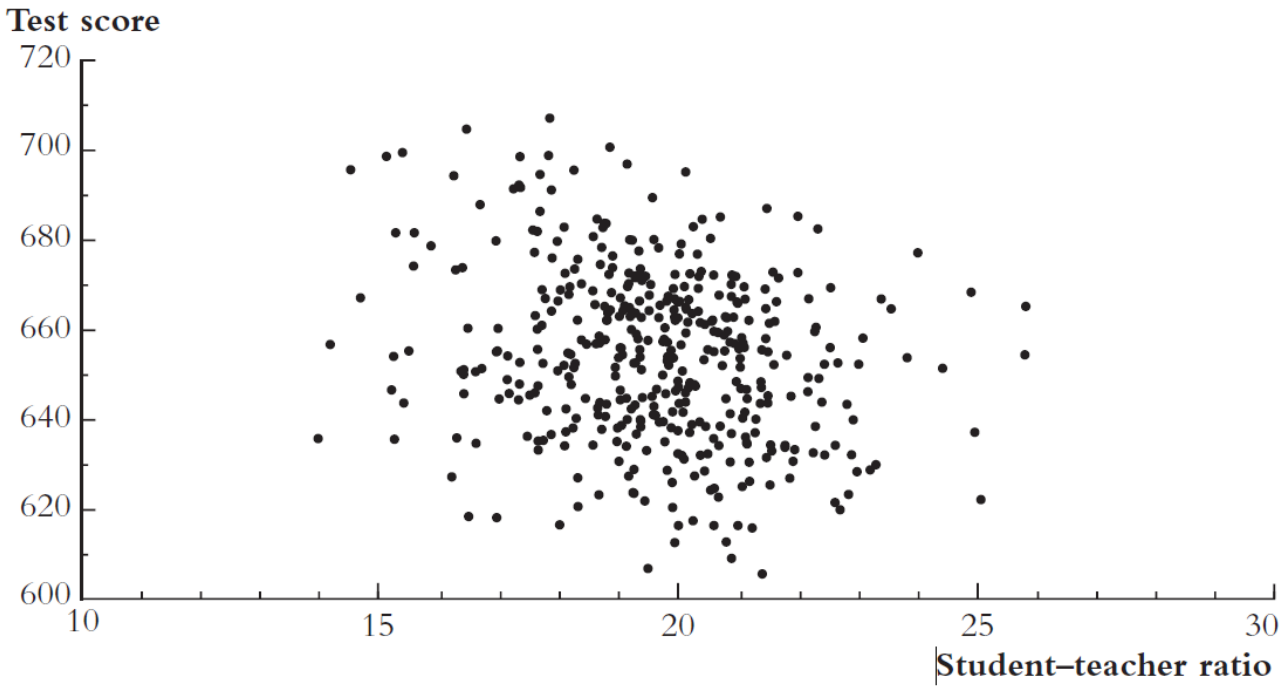
TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1999

			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

Do districts with smaller classes (lower STR) have higher test scores?

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.



Some Notation and Terminology

- The ***population regression line*:**

$$TestScore = \beta_0 + \beta_1 STR$$

β_1 = slope of population regression line

$$= \frac{\Delta Test\ Score}{\Delta STR}$$

= change in test score for a unit change in *STR*

Why are β_0 *and* β_1 *"population" parameters*?

- We would like to know the population value of β_1.
- We don't know β_1, so must estimate it using data

The Ordinary Least Squares Estimator

We will focus on the least squares ("*ordinary least squares*" or "*OLS*") estimator of the unknown parameters β_0 and β_1.

The OLS estimator solves,

$$\min_{b_0,b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem can be solved using calculus (App. 4.2).
- **The result is the OLS estimators of β_0 and β_1.**

Why use OLS, rather than some other estimator?

- OLS is a generalization of the sample average: if the "line" is just an intercept (no X), then the OLS estimator is just the sample average of $Y_1, \ldots Y_n$ ($\bar{Y}$).
- Like $\bar{Y}$, the OLS estimator has some desirable properties: under certain assumptions, it is unbiased (that is, $E(\widehat{\beta}_1) = \beta_1$), and it has a tighter sampling distribution than some other candidate estimators of β_1 (more on this later)
- Importantly, this is what everyone uses – the common "language" of linear regression.

The OLS Estimator, Predicted Values, and Residuals

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.8}$$

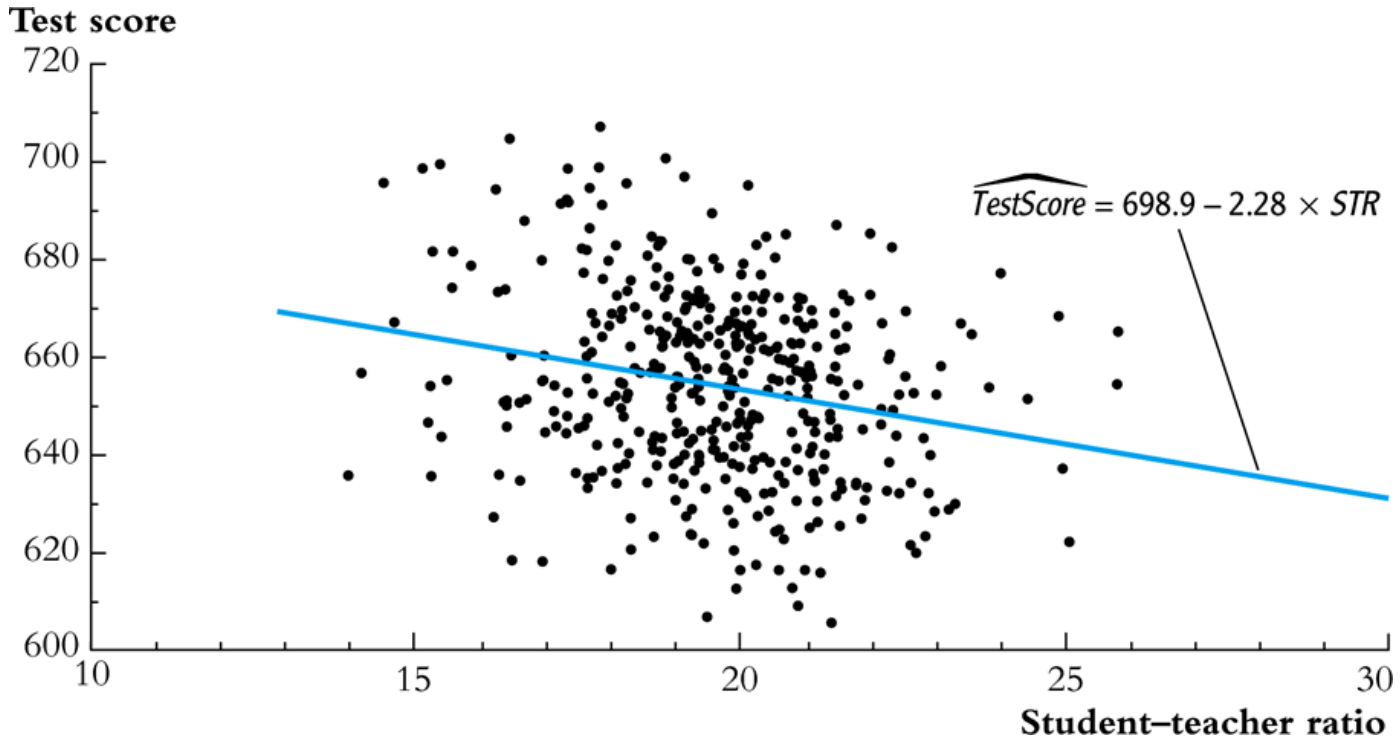
The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n. \tag{4.10}$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{u}_i$) are computed from a sample of n observations of X_i and $Y_i, i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

Application to the California *Test Score vs Class Size* data



- Estimated slope $= \hat{\beta}_1 = -2.28$
- Estimated intercept $= \hat{\beta}_0 = 698.9$
- Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

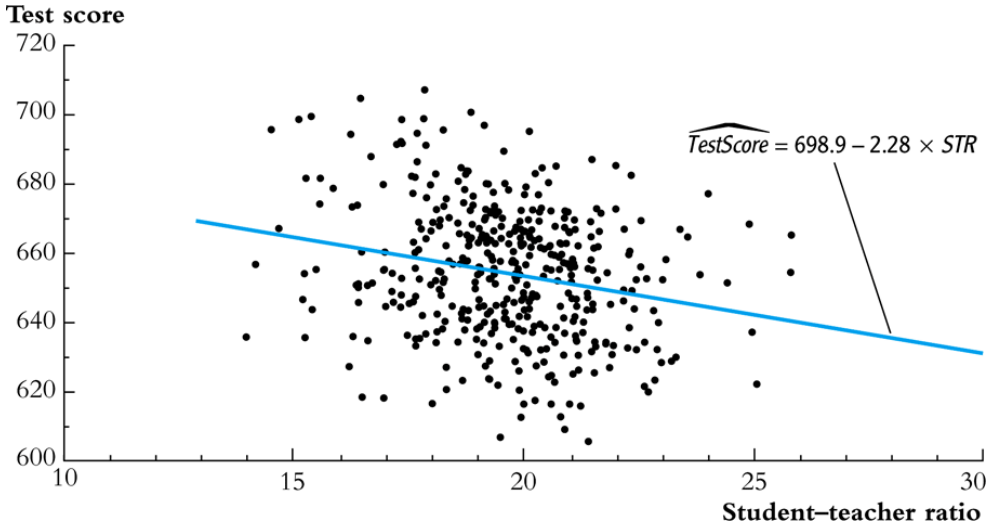
Interpretation of the estimated slope and intercept

- $\widehat{TestScore} = 698.9 - 2.28 \times STR$

- Districts with one more student per teacher on average have test scores that are 2.28 points lower.
- That is, $\frac{\Delta E(Test\ score|STR)}{\Delta STR} = -2.28$

- The intercept (taken literally) means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9. But this interpretation of the intercept makes no sense – it extrapolates the line outside the range of the data – here, the intercept is not economically meaningful.

Predicted values & residuals



One of the districts in the data set is Antelope, CA, for which *STR* = 19.33 and *Test Score* = 657.8

predicted value: $\hat{Y}_{Antelope} = 698.9 - 2.28 \times 19.33 = 654.8$

residual: $\hat{u}_{Antelope} = 657.8 - 654.8 = 3.0$

Looking ahead

The OLS regression line is an estimate, computed using our sample of data; a different sample would have given a different value of $\hat{\beta}_1$.

How can we:

- quantify the sampling uncertainty associated with $\hat{\beta}_1$?
- use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$?
- construct a confidence interval for β_1?

Like estimation of the mean, we proceed in four steps:

1. The probability framework for linear regression
2. Estimation
3. Hypothesis Testing
4. Confidence intervals

Excel live session: Analysis Toolpak Addin

- https://wwwedu.github.io/BC4400S23/Lecture2/Excel_Formulas_and_Functions.pdf#page=25

Excel live session: CASchools dataset

- https://wwwedu.github.io/BC4400S23/Lecture8/CASchools.csv