Quantitative Social Science in the Age of Big Data and AI

Lecture 4: Regression Analysis

Zhaopeng Qu
Hopkins-Nanjing Center
April 14 2025

Review the previous lecture

Causal Inference and RCT

- **Causality** is our main goal in the studies of empirical social science.
- To build a **reasonable counterfactual world** or to find a **proper control group** is the core of causal inference.
- The existence of **selection bias** makes social science more difficult than science.
- Although RCTs is a powerful tool for economists, every project or topic can **NOT** be carried on by it.
    - The main reason is the **ethical concerns** of RCTs.
- This is why modern econometric methods exists and develops. The main job of econometrics(causal inference) is using **non-experimental** data to **making convincing causal inference**.

Furious Seven Weapons

- To build a **reasonable counterfactual world** or to find a **proper control group** is the core of econometric methods.

  1. **Randomized controlled trial(RCTs)**
  2. **Regression**
  3. **Matching and Propensity Score**
  4. **Instrumental Variable**
  5. **Regression Discontinuity**
  6. **Panel Data and Difference in Differences**
  7. **Synthetic Control Method**

- The most fundamental of these tools is **regression**. It compares treatment and control subjects with the same observable characteristics **in a generalized manner**.
- It paves the way for the more elaborate tools used in the class that follow.
- **Let's start our exciting journey from it**.

OLS Regression: Simple Regression

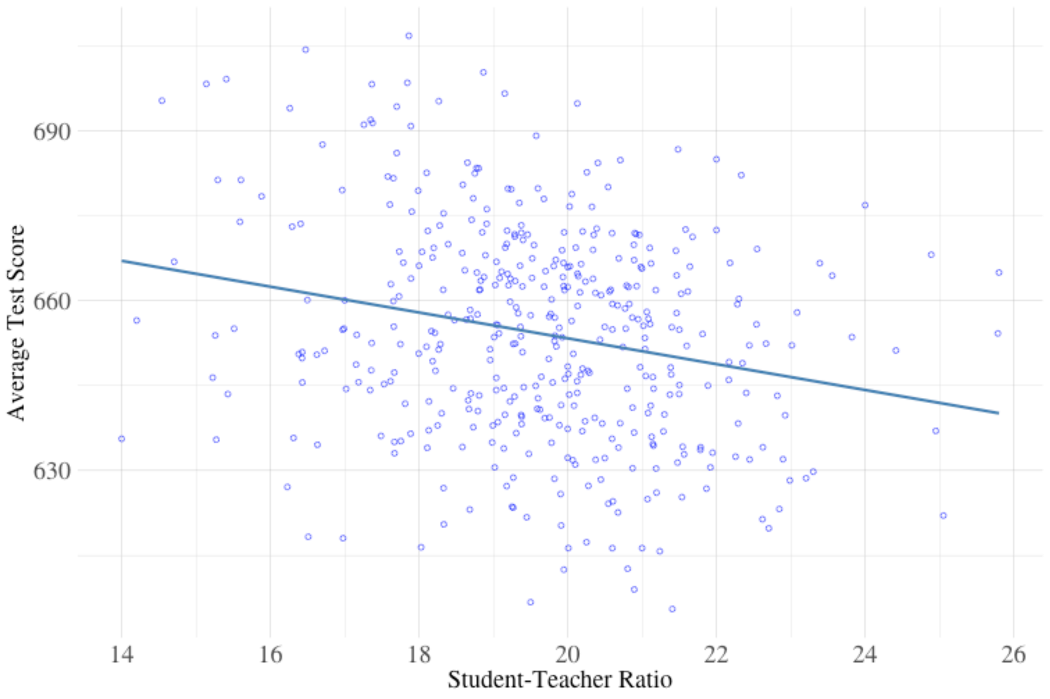
Class Size and Student's Performance

- **Topic**: Class Size and Student's Performance in California
- **Data**: California Test Score Data(`CASchools`)
  - The data used here are from all 420 K-6 and K-8 districts in California in 1998 and 1999.
  - **Test scores** are on the Stanford 9 standardized test administered to 5th grade students.
  - **School characteristics**(average across the district): class size, number of computers per classroom, expenditures per student, etc.
  - **Student characteristics**(average across the district): percentage of students in the public assistance program CalWorks (formerly AFDC), average family income, the percentage of students that qualify for a reduced price lunch, and the percentage of students that are English learners (that is, students for whom English is a second language).

Class Size and Student's Performance

- **Please answer the following questions**:
  - *What is the population in this study?*
  - *What is the sample in this study?*
  - *What is the unit of observation in this study?*
  - *What is the variable of interest(outcome)?*
  - *What is the treatment variable(cause)?*
  - *Could you think of any other variables that might affect the outcome?*

Class Size and Student's Performance



- What can we tell from the plot?
- Specifically, is there a relationship between class size and student's performance?
  - Positive or Negative?
  - How strong is the relationship?
  - The relationship is linear or other forms?

Quantitative Question and Modeling

- **Specific Quantitative Question**:
    - *What is the effect on district* ***test scores*** *if we would increase district average* ***class size*** *by 1 student (or one unit of Student-Teacher's Ratio)*

- If we could know the full relationship between two variables which can be summarized by a **model**, which can be written as a real value function $f(\cdot)$, thus

$$Testscore = f(ClassSize)$$

- Unfortunately, the function form is always unknown.

- Now our job is to find a **reasonable approximation** of the function $f(\cdot)$.

Modeling and Estimation Methods

- **Two** basic methods to obtain the function $f(\cdot)$:
  - **non-parametric**: we don't care the specific form of the function, unless we know all the values of two variables or some key characteristics of the function. We can specify the *whole distributions* of class size and test scores.
  - **parametric**: we have to suppose the basic form of the function, then to find values of some *unknown parameters* to determine the specific function form.

Modeling and Estimation Methods

- Suppose we choose **parametric** method, we need to know the real value of a **parameter** β_1 to describe the relationship between Class Size and Test Scores as

$$\beta_1 = \frac{\Delta Testscore}{\Delta ClassSize}$$

- Next step, we have to suppose specific forms of the function $f(\cdot)$, still two categories: **linear** and **non-linear**

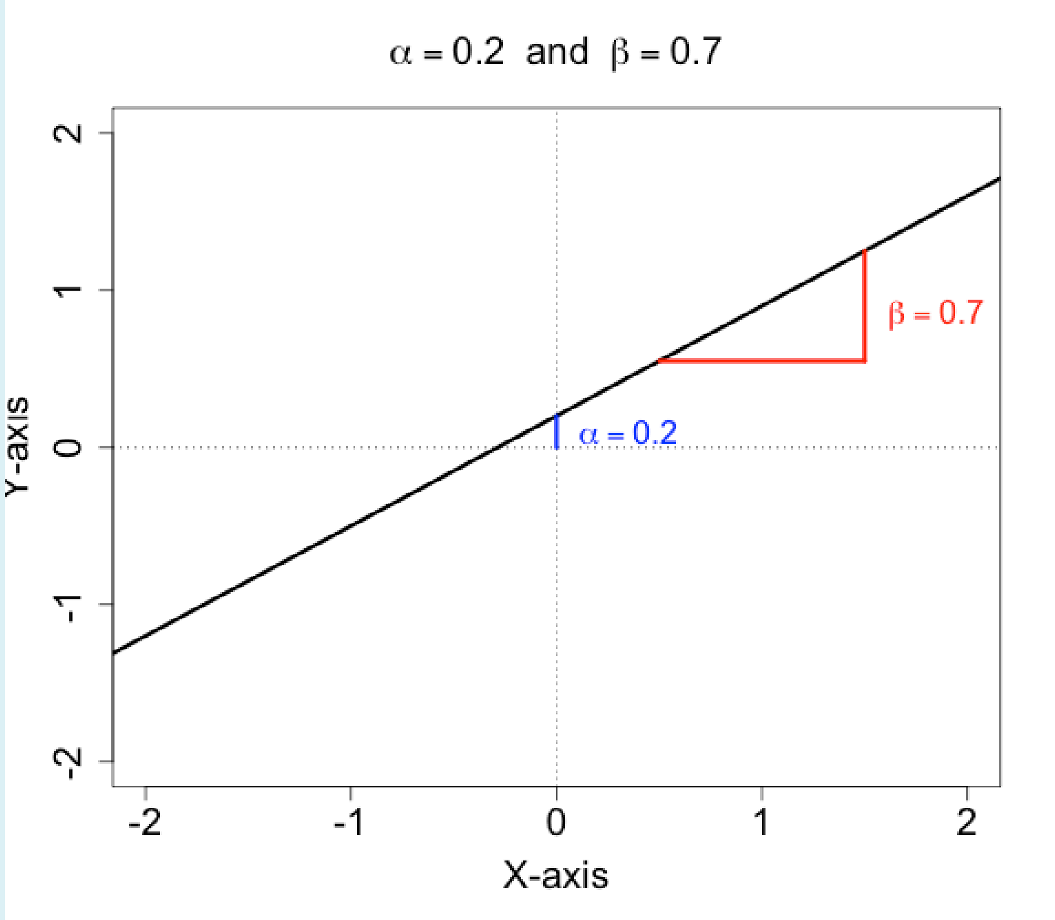
- And we start to use the *simplest* function form: a **linear** equation, which is graphically **a straight line**, to summarize the relationship between two variables.

$$Test\ score = \beta_0 + \beta_1 \times Class\ size$$

  where β_1 is actually the **the slope** and β_0 is the **intercept** of the straight line.

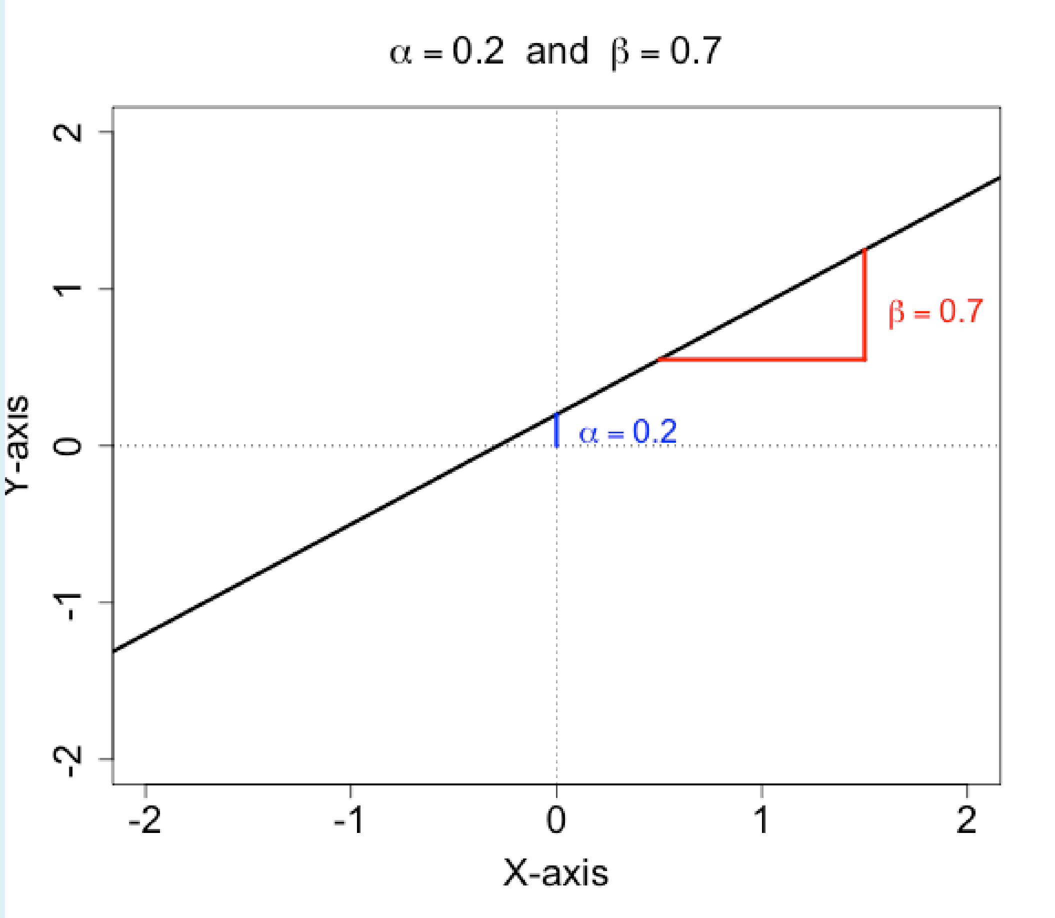
Review: Linear functions or relationships

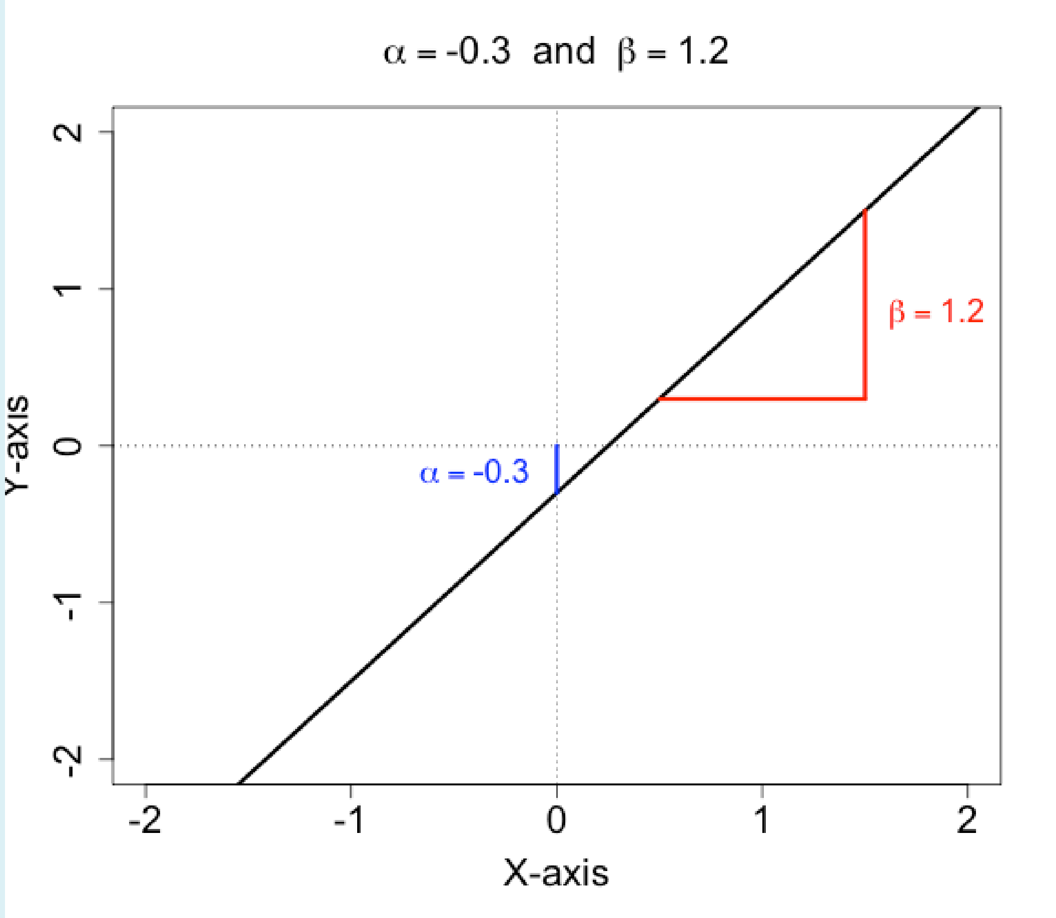
- A **straight line** can be represented by a **linear** equation: $Y = \alpha + \beta X$



- α is the **intercept**: the value of Y where $X = 0$.
- β is the **slope**: the amount that Y increases when X increases by one unit.
- Here, a one-unit increase in X is associated with a 0.7-unit increase in Y.

Review: Linear functions or relationships





From Functional Model to Estimation Model

- BUT the average test score in district i does not **only** depend on the average class size
- It also depends on **other factors** such as
    - Student background
    - Quality of the teachers
    - School's facilitates
    - Quality of text books
    - Random deviation......
- Therefore, the equation describing the linear relation between Test score and Class size is **better** written as

$$Test\ score_i = \beta_0 + \beta_1 \times Class\ size_i + u_i$$

where u_i lumps together all **other factors** that affect average test scores.

Terminology for Simple Regression Model

- The linear regression model with one regressor is denoted by

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Where

  - Y_i is the **dependent variable**(Test Score)
  - X_i is the **independent variable** or regressor(Class Size or Student-Teacher Ratio)

- The intercept β_0 and the slope β_1 are the **coefficients** of the **population regression line**, also known as the **parameters** of the population regression line.

- u_i is the **error term** which contains all the other factors **besides** X that determine the value of the dependent variable, Y, for a specific observation, i.

Population Regression: relationship in average

- The linear regression model with one regressor is denoted by$$Y{i}=\beta{0}+\beta{1}X{i}+u_{i}$$
- Then $\beta_0 + \beta_1 X_i$ is the **population regression line** or the **population regression function**
- Both side to conditional on X, then

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i + E[u_i|X_i]$$

- Suppose $E[u_i|X_i] = 0$ then

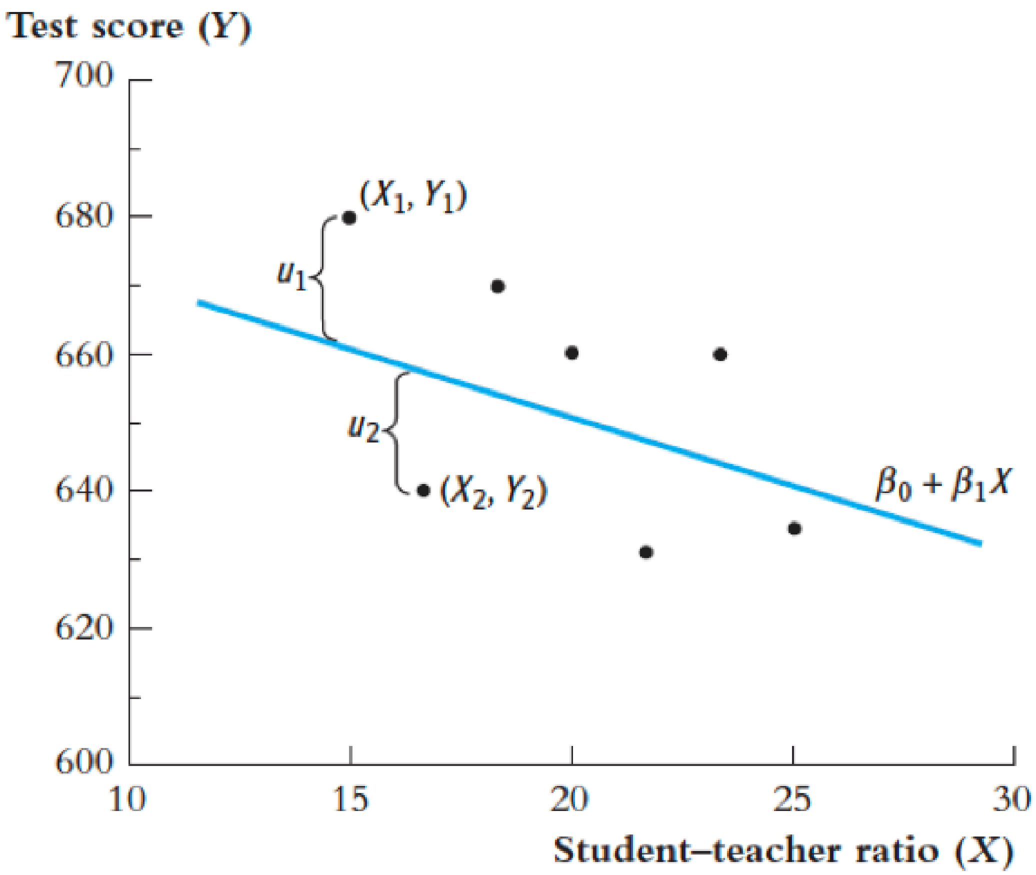
$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

- Population regression function is the relationship that holds between Y and X **on average over the population**.

Graphics for Simple Regression Model

FIGURE 4.1 Scatterplot of Test Score vs. Student–Teacher Ratio (Hypothetical Data)

The scatterplot shows hypothetical observations for seven school districts. The population regression line is $\beta_0 + \beta_1 X$. The vertical distance from the i^{th} point to the population regression line is $Y_i - (\beta_0 + \beta_1 X_i)$, which is the population error term u_i for the i^{th} observation.

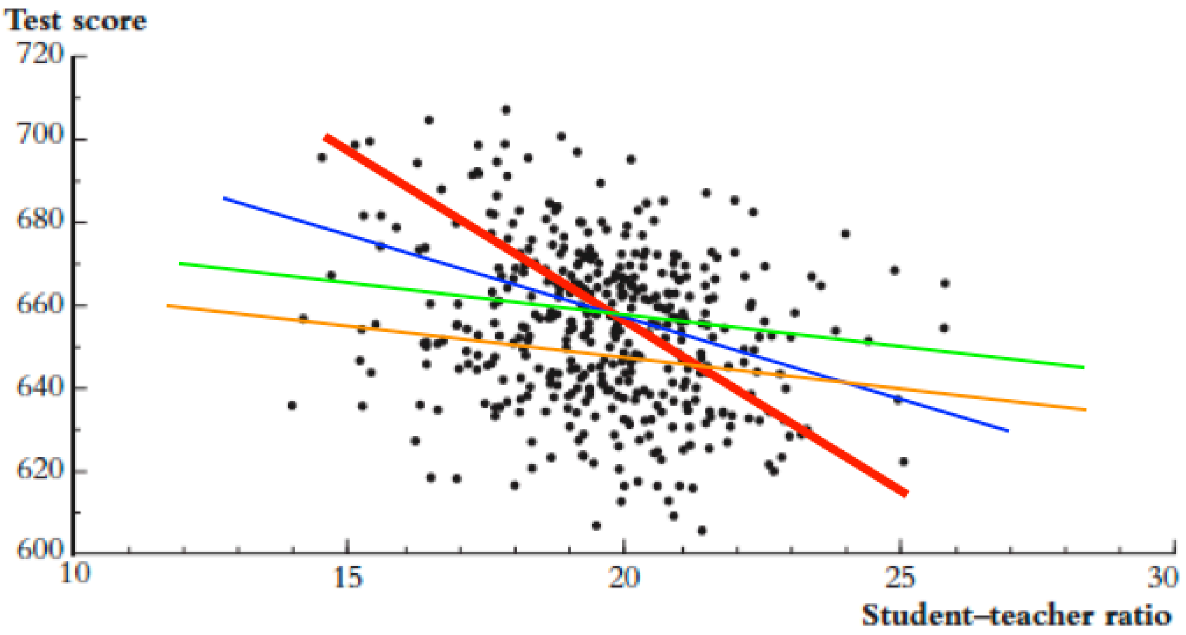


How to find the "best" fitting line?

- In general we don't know β_0 and β_1 which are parameters of **population regression function** but have to calculate them using a bunch of data: **the sample**.

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.



- How to find the line that fits the data best?

The Ordinary Least Squares(OLS)

The OLS estimator

- Chooses the **best** regression coefficients so that the estimated regression line is **as close as possible** to the observed data, where **closeness** is measured by **the sum of the squared mistakes** made in predicting Y given X.

- Let b_0 and b_1 be **estimators** of β_0 and β_1,thus

$$b_0 \equiv \hat{\beta}_0 \text{ and } b_1 \equiv \hat{\beta}_1$$

- The predicted value of Y_i given X_i using these estimators is $b_0 + b_1 X_i$, or $\hat{\beta}_0 + \hat{\beta}_1 X_i$ formally denotes as $\hat{Y}_i$, thus

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- The **prediction mistake** is the difference between Y_i and $\hat{Y}_i$,which denotes as $\hat{u}_i$, **the residual**

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$$

The Ordinary Least Squares(OLS)

The OLS Estimator

- Math Review:

  - **Summation**: $\sum_{i=1}^{n} u_i = u_1 + u_2 + u_3 + \ldots + u_n$
  - **Minimization**: $\underset{b_0, b_1}{min}$ which means we need to find the values of b_0 and b_1 that **minimize** the expression
  - **Derivative**: $\frac{\partial}{\partial b_0}$ and $\frac{\partial}{\partial b_1}$

- The OLS estimator minimizes the *sum of squared prediction mistakes*:

$$\underset{b_0, b_1}{min} \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

- Solve the problem by optimization: **F.O.C**(the first order condition)

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = 0$$

The Ordinary Least Squares(OLS)

The OLS Estimator

- Step 1: OLS estimator of β_0

$$b_0 \equiv \hat{\beta}_0 = \overline{Y} - b_1\overline{X}$$

- Step 2: OLS estimator of β_1

$$b_1 \equiv \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})}$$

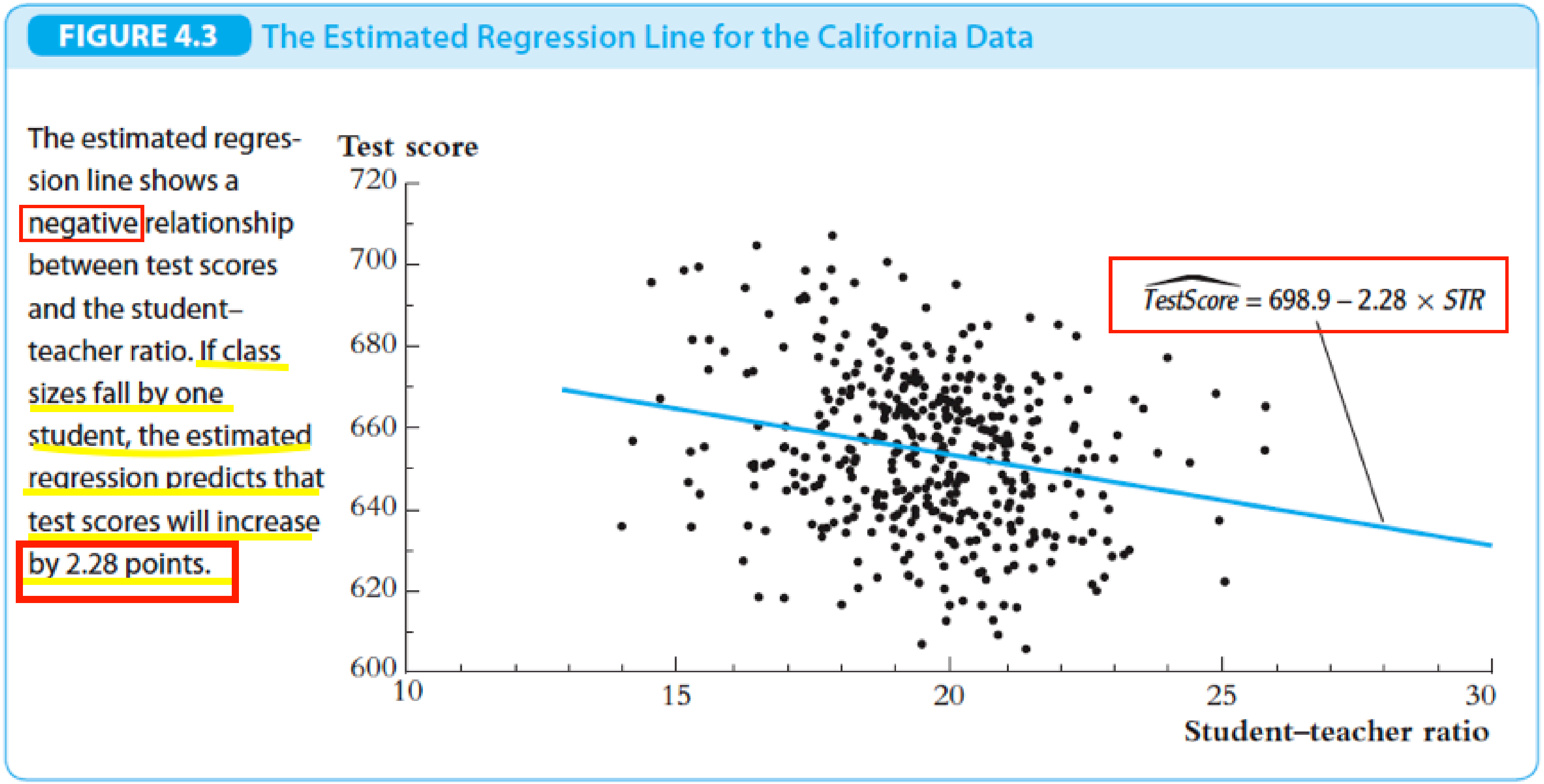
- Now as long as we have data, we can calculate $\hat{\beta}_0$ and $\hat{\beta}_1$, and then we can obtain the estimated regression line.

The Estimated Regression Line

- Obtain the values of OLS estimator for a certain data,

$$\hat{\beta}_1 = -2.28 \; and \; \hat{\beta}_0 = 698.9$$

- Then the regression line is



Measures of Fit: The R^2

- Math Review:
  - **Variance**: $Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$
  - **Standard Deviation**: $SD(X) = \sqrt{Var(X)}$
- Because the **variation** of Y can be summarized by **Variance**,Then the total variation of Y_i, which are also called as the **total sum of squares** (TSS), is:

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Measures of Fit: The R^2

- Because Y_i can be decomposed into the fitted value plus the residual: $Y_i = \hat{Y}_i + \hat{u}_i$,then likewise Y_i, we can obtain

$$TSS = ESS + SSR$$

- The **explained sum of squares** (ESS):

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

- The **sum of squared residuals** (SSR):

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2 = \sum_{i=1}^{n} \hat{u}_i^2$$

Measures of Fit: The R^2

- R^2 or **the coefficient of determination**, is the fraction of the sample variance of Y_i explained/predicted by X_i

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- So $0 \leq R^2 \leq 1$, it measures that *how much can the variations of Y be explained by the variations of X_i in share*.
- **Question**: *If R-squares is bigger, is the regression better?*
- **Answer**: **Not necessarily**,especially when we make **causal inference** in cross-sectional data.

Least Squares Assumptions

1. Assumption 1: Conditional Mean is Zero
2. Assumption 2: Random Sample
3. Assumption 3: Large outliers are unlikely

- If the 3 least squares assumptions hold the OLS estimators will be
  - **unbiased**
  - **consistent**
  - **normal sampling distribution**

Simple OLS and RCT

Simple OLS and RCT

- We have known that RCT is the **"golden standard"** for causal inference.Because it can naturally eliminate **selection bias**.
- So far, we did not discuss the relationship between RCT and OLS regression, which means that we can not be sure that the result from an OLS regression can be explained as "causal".
- Instead of using a continuous regressor X, the regression where D_i is a binary variable, a so-called **dummy variable**, will help us to unveil the relationship between RCT and OLS regression.
- For example, we may define D_i as follows:
- The regression can be written as

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

Regression when X is a Binary Variable

- More precisely, the regression model now is

$$TestScore_i = \beta_0 + \beta_1 D_i + u_i$$

  - With D as the regressor, it is not useful to think of β_1 as a slope parameter.
  - Since $D_i \in \{0, 1\}$, i.e., we only observe two discrete values instead of a continuum of regressor values.

- There is no continuous line depicting the conditional expectation function $E(TestScore_i|D_i)$ since this function is solely defined for x-positions 0 and 1.

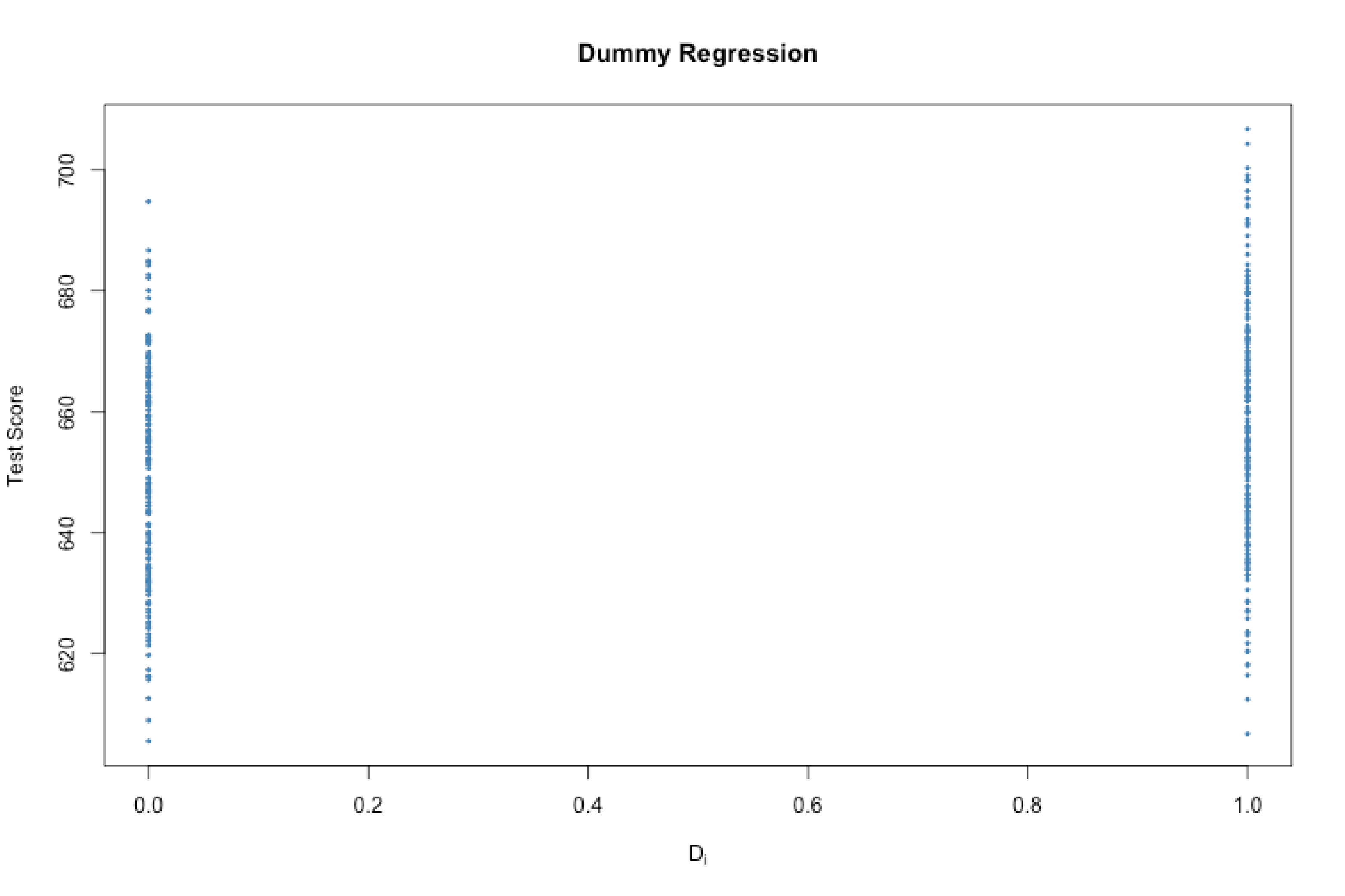
- The interpretation of the coefficients in this regression model is as follows:

  - $E(Y_i|D_i = 0) = \beta_0$, so β_0 is the expected test score in districts where $D_i = 0$ where STR is below 20.
  - $E(Y_i|D_i = 1) = \beta_0 + \beta_1$ where STR is above 20

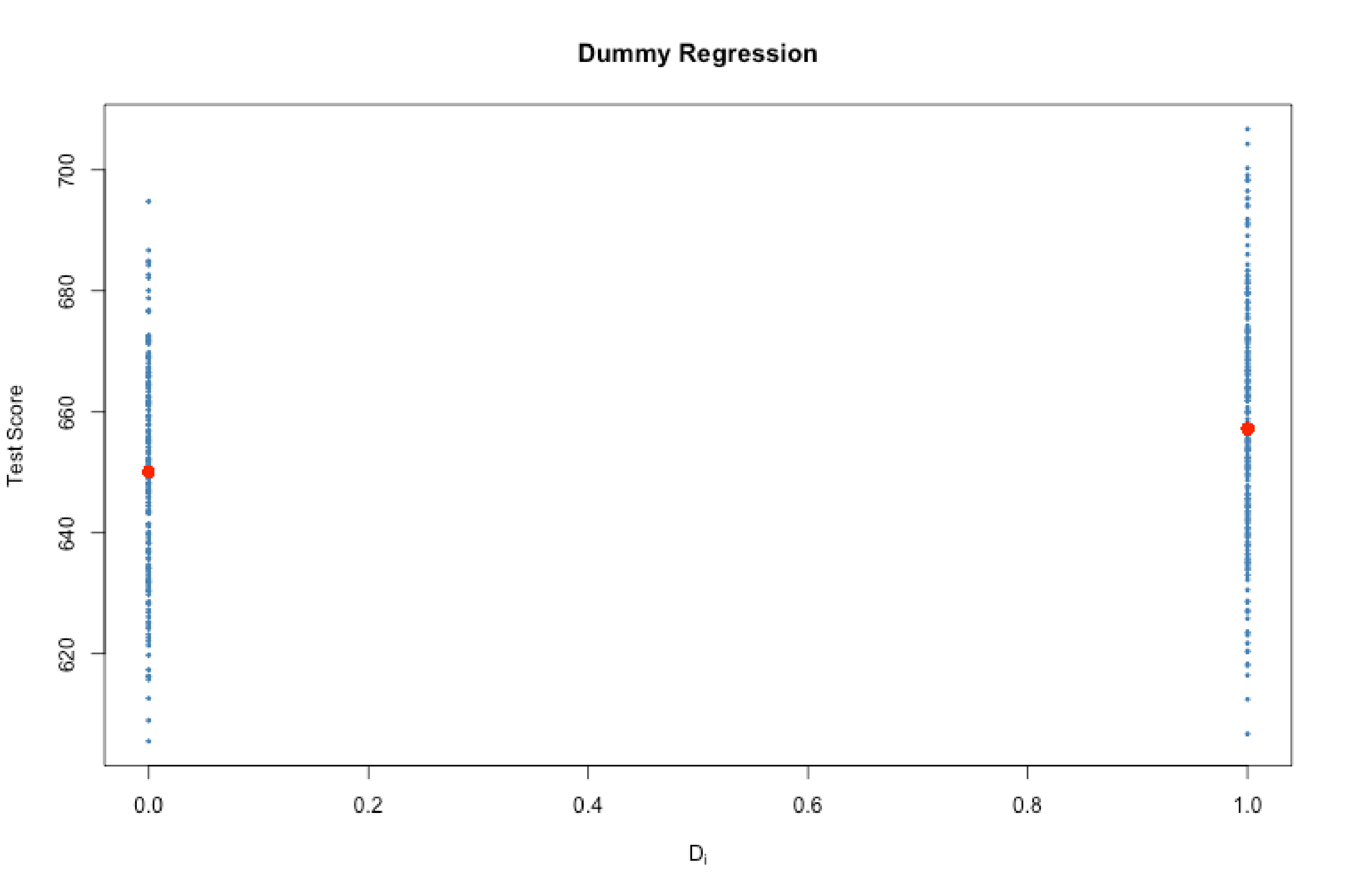
- Thus, β_1 is **the difference in group specific expectations**, i.e., the difference in expected test score between districts with $STR < 20$ and those with $STR \geq 20$,

$$\beta_1 = E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$$

Class Size and Test Score: D is a Binary Variable



Class Size and Test Score: D is a Binary Variable



Causality and OLS

- **Recall**, the *individual treatment effect*(ICE) is defined as

$$ICE = Y_{1i} - Y_{0i} = \delta_i$$

- The ATE is the average of the ICE and ATT is the average of the ICE for the treated group.

$$\rho = E(\delta_i) \ or \ \rho = E(\delta_i | D = 1)$$

- Either way, the treatment effect is a constant, i.e., it does not depend on the individual.
- **OLS regression** is to estimate a constant treatment effect ρ, thus

$$\mathrm{Y}_i = \underbrace{\alpha}_{E[\mathrm{Y}_{0i}]} + \mathrm{D}_i \underbrace{\delta_i}_{\mathrm{Y}_{1i} - \mathrm{Y}_{0i}} + \underbrace{\eta_i}_{\mathrm{Y}_{0i} - E[\mathrm{Y}_{0i}]}$$

Causality and OLS

- Now write out the conditional expectation of Y_i for both levels of D_i

$$E\left[Y_i \mid D_i = 1\right] = E\left[\alpha + \delta_i + \eta_i \mid D_i = 1\right] = \alpha + \rho + E\left[\eta_i | D_i = 1\right]$$
$$E\left[Y_i \mid D_i = 0\right] = E\left[\alpha + \eta_i \mid D_i = 0\right] = \alpha + E\left[\eta_i \mid D_i = 0\right]$$

- Take the difference

$$E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right] = \rho + \underbrace{E\left[\eta_i | D_i = 1\right] - E\left[\eta_i \mid D_i = 0\right]}_{\text{Selection bias}}$$

- Again, our estimate of the **treatment effect** (ρ) is only going to be as good as our ability to shut down the **selection bias**.

- *Selection bias in regression model:*

$$E\left[\eta_i | D_i = 1\right] - E\left[\eta_i \mid D_i = 0\right]$$

- Thus there is something in our disturbance η_i that is affecting Y_i and is also correlated with D_i.

Simple OLS Regression v.s. RCT

- In a simple regression model, OLS estimators are just a **generalizing continuous version of RCT** when least squares assumptions are hold.
- Ideally,regression is a way to control observable confounding factors, which assume the source of selection bias is only from the difference in **observed characteristics**.
- But in contrast to RCT, in observational studies, researchers cannot control the assignment of treatment into a **treatment group versus control group**,which means that the two groups are **incomparable**.
- To make two groups comparable, we need to keep treatment and control group **other thing equal** in observed characteristics and unobserved characteristics.
- OLS regression is **valid** only when least squares assumptions are hold.
- In most cases,it is not easy to obtain. We have to know how to make a convincing causal inference when these assumptions are not hold.