Introduction to Econometrics

Lecture 3 : Regression: CEF and Simple OLS

Zhaopeng Qu

Business School,Nanjing University

Oct. 9th, 2017

Outlines

1. Review the Previous Lecture

2. Make Regression Make Sense
   - Economic Relationships and the CEF
   - The Properties of CEF
   - Linear Regression and the CEF: Why Regress?

3. Simple Regression: Ordinary Least Squares

4. The Least Squares Assumptions

Review the Previous Lecture

Causal Inference in Social Science

- Causality is our main goal in the studies of empirical social science.
- The existence of Selection Bias makes social science more difficult than science.
- Although experimental method is a powerful tool for economists, every project can not be carried on by it.
- It is the main reason *why modern econometrics exists and develops.*
- *"The modern menu of econometric methods can seem confusing, even to an experienced number cruncher. Luckily, not everything on the menu is equally valuable or important. Some of the more exotic items are needlessly complex and may even be harmful. On the plus side, the core methods of applied econometrics remain largely unchanged, while the interpretation of basic tools has become more nuanced and sophisticated."* *Angrist and Pischke(2009)*

Furious Seven Weapons (七种武器)

- Build a reasonable counterfactual world or find a proper control group is the core of econometrical methods.
    1. **Random Trials(随机试验)**
    2. **Regression(Ordinary Least Squares)(OLS 回归)**
    3. **Matching and Propensity Score (匹配与倾向得分)**
    4. **Decomposition (分解)**
    5. **Instrumental Variable (工具变量)**
    6. **Regression Discontinuity (断点回归)**
    7. **Difference in Differences （双差分或倍差法)**
- The most basic of these tools is regression, which compares treatment and control subjects who have the *same observed* characteristics.
- Regression concepts are foundational, paving the way for the more elaborate tools used in the class that follow.
    - *So let's start our exciting journey from it.*

Make Regression Make Sense

Regression: What You Need to Know

- We spend our lives running regressions (I should say: ”regressions run me”). And yet this basic empirical tool is often misunderstood. So I begin with a recap of key regression properties. (Angrist, 2014)
- Our Regression Agenda
  1. The CEF is all you need
  2. What is Regression and Why We Regress
  3. Regression and Causality

Conditional Expectation Function(CEF): Education and Earnings

- Most of what we want to do in the social science is learn about how **two variables** are related, such as *Education and Earnings*.
- On average, people with more schooling earn more than people with less schooling.
    - The connection between schooling and average earnings has considerable predictive power, in spite of the enormous variation in individual circumstances.
    - The fact that more educated people earn more than less educated people does not mean that schooling causes earnings to increase.
    - However, it's clear that education predicts earnings in a narrow statistical sense.
- This predictive power is compellingly summarized by the **Conditional Expectation Function.**

Probability Review: Conditional Expectation Function(CEF)

- Both X and Y are r.v., then conditional on X, Y's probability density function is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f(x)}$$

- Conditional on X, Y's expectation is

$$E(Y|X) = \int_Y y f_{Y|X}(y|x)\,dy = \int_Y y\frac{f(x,y)}{f(x)}\,dy$$

- So Conditional Expectation Function(CEF) is a function of x, since x is a random variable, so CEF is also a random variable
- **直观理解：期望就是求平均值，而条件期望就是"分组取平均"或"在... 条件下的均值"。**

Conditional Expectation Function(CEF): Education and Earnings

- Most of what we want to do in the social science is learn about how two variables are related, such as Education and Earnings.
- On average, people with more schooling earn more than people with less schooling.
  - The connection between schooling and average earnings has considerable predictive power, in spite of the enormous variation in individual circumstances.
  - The fact that more educated people earn more than less educated people does not mean that schooling causes earnings to increase.
  - However, it's clear that education predicts earnings in a narrow statistical sense.
- This predictive power is compellingly summarized by the **Conditional Expectation Function.**
- The Law of Iterated Expectations(LIE)

$$E[Y] = E[E[Y \mid X]]$$

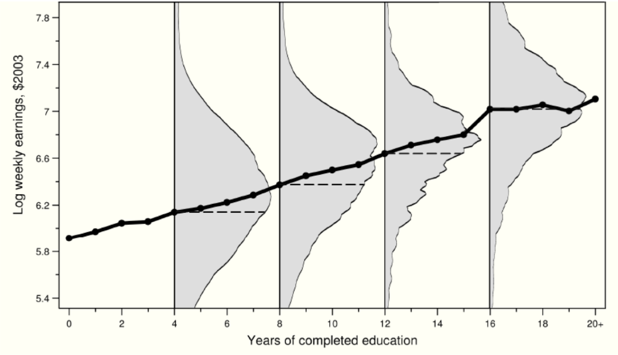
The Law of Iterated Expectations(LIE)

Proof.

$$\begin{aligned}
E[E(Y|X)] &= \int E[Y \mid X = u] g_X(u)\,du \\
&= \int \left[\int t f_Y(t \mid X = u)\,dt \right] g_X(u)\,du \\
&= \iint t f_Y(t \mid X = u) g_X(u)\,dt\,du \\
&= \int t \left[\int f_Y(t \mid X = u) g_X(u)\,du \right] dt \\
&= \int t \left[\int f_{XY}(t, u)\,du \right] dt \\
&= \int t f_Y(t)\,dt
\end{aligned}$$

□

Education and Wage: CEF View



- The figure plots the CEF of **log weekly wages** given **schooling** for a sample of middle-aged white men from the 1980 census.
- The CEF in the figure captures the fact that the enormous variation individual circumstances notwithstanding - people with more schooling generally earn more, on average.

The CEF Decomposition Property

Theorem

Every random variable such as Y_i can be written as

$$Y_i = E[Y_i \mid X_i] + \varepsilon_i$$

where ε_i is mean-independent of X_i, i.e., $E[\varepsilon_i \mid X_i] = 0$. and therefore ε_i is uncorrelated with any function of X_i.

Proof.

Skipped □

- This theorem says that any random variable, Y_i, can be decomposed into two parts
  - a piece that's "explained by X_i", i.e. the CEF,
  - a piece left over which is orthogonal to (i.e. uncorrelated with) any function of X_i.

The CEF-Prediction Property

Theorem

*Let $m(X_i)$ be any function of X_i. The CEF is **the Minimum Mean Squared Error**(MMSE) predictor of Y_i given X_i. Thus*

$$E[Y_i \mid X_i] = \underset{m(X_i)}{argmin E}\left[[Y_i - M(X_i)]^2\right]$$

Proof.

$$\begin{aligned} (Y - m(X_i))^2 &= [(Y_i - E[Y_i \mid X_i]) + (E[Y_i \mid X_i] - m(X_i))]^2 \\ &= (Y_i - E[Y_i \mid X_i])^2 \\ &+ 2(Y_i - E[Y_i \mid X_i])(E[Y_i \mid X_i] - m(X_i)) \\ &+ (E[Y_i \mid X_i] - m(X_i))^2 \end{aligned}$$

The last term is minimized at zero when $m(X_i)$ is the CEF. □

The CEF-Prediction Property

- Suppose we are interested in predicting Y using some function $m(X_i)$, the optimal predictor under the **MMSE** (Minimized Mean Squared Error) criterion is CEF.
- Therefore ,CEF is a natural summary of the relationship between Y and X under MMSE.
- *It means that if we can know CEF, then we can describe the relationship of Y and X.*

CEF and Regression

- We have already learned CEF is a natural summary of the relationships which we would like to know it.
- But CEF is an unknown functional form, so the next question is How to model CEF, $E(Y \mid X)$?
- Answer: Two basic approaches
  - Nonparametric(Matching, Kernel Density etc.)
  - Parametric(Regression)
- Regression estimates provides a valuable baseline for almost all empirical research because Regression is tightly linked to CEF.

Estimating the CEF: two discrete values

- Suppose a binary X case : X only take on two values, 0 and 1(like our formal example: here X is a treatment).
- We've been writing and for the means in different groups。
  - Then the mean in each group is just a conditional expectation:
  - The fact that more educated people earn more than less educated people does not mean that schooling causes earnings to increase.
  - However, it's clear that education predicts earnings in a narrow statistical sense.
- How to estimate $\hat{E}[Y \mid X_i = x]$? it means that we have to use sample data to inference the population.
- we could use **sample means** within each group.

$$\begin{aligned}
\hat{E}[Y \mid X_i = 1] &= \frac{1}{n_1} \sum_{i:X_i=1} Y_i \\
\hat{E}[Y \mid X_i = 0] &= \frac{1}{n_0} \sum_{i:X_i=0} Y_i
\end{aligned}$$

here $n_0$and n_1 are numbers of men and women in the sample.

Estimating the CEF: multiple discrete values

- What if X takes on > 2 discrete values?
- we can still estimate $\hat{E}[Y \mid X_i = x]$ with the sample mean among those who have $X_i = x$, thus

$$\hat{E}[Y \mid X_i = x] = \frac{1}{n_x} \sum_{i:X_i=x} Y_i$$

where $n_x$and is the number of group x in the sample.

Estimating the CEF: continuous

- What if X is continuous? Can we calculate a mean for every value of X_i.
- Because the probability could take values only in an interval for a continuous variable. So we could turn it into a discrete variable. This is called as **stratification**.
  - Once it's discrete, we can just calculate the means within each strata.
- The stratification approach was fairly crude: it assumed that means were constant within strata, but that seems wrong.
- Now we will think about $E[Y \mid X_i = x]$ as a function. What does this function look like?
  - unknown functions in the population! make producing an estimator very difficult!

Population Regression: What is a Regression?

Definition

population regression ("regression" for short) as the solution to the population least squares problem. Specifically, the K×1 regression coefficient vector β is defined by solving

$$\beta = \underset{b}{arg\,min} E\left[(Y_i - X_i'b)^2\right]$$

- Using the first order condition

$$E[X_i(Y_i - X_i'b)] = 0$$

- The solution for b can be written

$$\beta = E[X_iX_i']^{-1}E[X_iY_i']$$

Three Reasons to Regress

- There are three reasons (three justifications) why the vector of population regression coefficient might be of interest.
  1. The Best Linear Predictor Theorem
  2. The Linear CEF Theorem
  3. The Regression-CEF Theorem

Regression Justification I

Theorem

The Best Linear Predictor Theorem

Regression solves the population least squares problem and is therefore the Best Linear Predictor(BLP) of Y_i given X_i.

Proof.

By definition of regression. □

- In other words, just as CEF, which is the best predictor of Y_i given X_i in the class of all functions of X_i, the population regression function is the best we can do in the class of linear functions.

Regression Justification II

Theorem

The Linear CEF Theorem

Suppose the CEF is linear. Then the Regression function is it.

Proof.

Suppose $E(Y_i|X_i) = X_i'\beta^*$ for a K×1 vector of coefficients. By the CEF decomposition property, we have

$$E[X_i(Y_i - E[Y_i \mid X_i])] = 0$$

Then substitute using $E(Y_i|X_i) = X_i'\beta^*$

$$E[X_i(Y_i - X_i'\beta^*)] = 0$$

At last find that

$$\beta^* = \beta = E[X_iX_i']^{-1} E[X_iY_i]$$

Regression Justification III

Theorem

The Regression-CEF Theorem

The population regression function $X_i'\beta$ provides the ***MMSE linear approximation*** *to $E(Y_i|X_i)$, thus*

$$\beta = \underset{b}{arg\,min} E\left[\left(E[Y_i|X_i] - X_i'b\right)^2\right]$$

Proof.

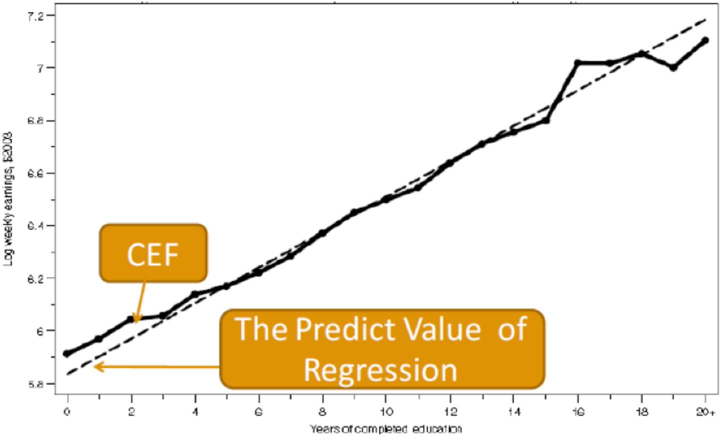
$$\begin{aligned}\left(Y_i - X_i'b\right)^2 &= \left[\left(Y_i - E[Y_i \mid X_i]\right) + \left(E[Y_i \mid X_i] - X_i b\right)\right]^2 \\ &= \left(Y_i - E[Y_i \mid X_i]\right)^2 + \left(E[Y_i \mid X_i] - X_i b\right)^2 \\ &+ 2\left(Y_i - E[Y_i \mid X_i]\right)\left(E[Y_i \mid X_i] - X_i b\right)\end{aligned}$$

The first term has no b and the last term by the CEF-decomposition property. Therefore the minimized problem has the same solution as

Regression Justification

- The Best Linear Predictor Theorem and The Regression-CEF Theorem show us two more ways to view regression:
  - Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable.
  - If we prefer to think about approximating $E(Y_i|X_i)$, as opposed to predicting Y_i, the Regression-CEF theorem tells us that even if the CEF is nonlinear, regression provides the best linear approximation to it.
- Actually, The regression-CEF theorem is our favorite way to motivate regression. The statement that regression approximates the CEF lines up with our view of empirical work as an effort to describe the essential features of statistical relationships, without necessarily trying to pin them down exactly.
- We are not really interested in predicting individual Y_i; it's the distribution of Y_i that we care about.

The CEF and Regression



CEF

The Predict Value of Regression

Sample is limited to white men, age 40-49. Data is from Census IPUMS 1980, 5% sample.

Figure 3.1.2: Regression threads the CEF of average weekly wages given schooling

Simple Regression: Ordinary Least Squares

Question: Class Size and Student's Performance

- Specific 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)

- Technically,we would like to know the real value of a parameter β_1,

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

- And β_1 is actually the definition of **the slope** of a straight line relating test scores and class size. Thus

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

where β_0 is the is the **intercept** of the straight line.

Question: Class Size and Student's Performance

- 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 .....
- So 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 district characteristics** that affect average test scores.

Terminology for Simple Regression Model

- The linear regression model with one regressor is denoted by

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

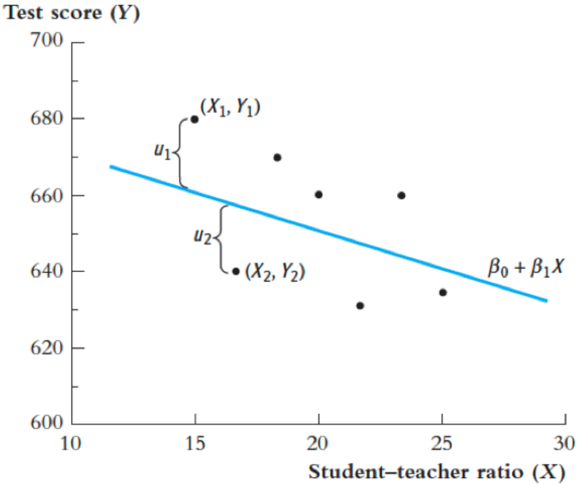
Where

- Y_i is the **dependent variable**(*Test Score*)
- X_i is the **independent variable** or regressor(Class Size or Student-Teacher Ratio)
- $\beta_0 + \beta_1 X_i$ is the **population regression line** or the **population regression function**
  - This is the relationship that holds between Y and X on average over the population. (be familiar with? Recall the concept of **conditional expectation**)
- 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.*

Terminology 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.

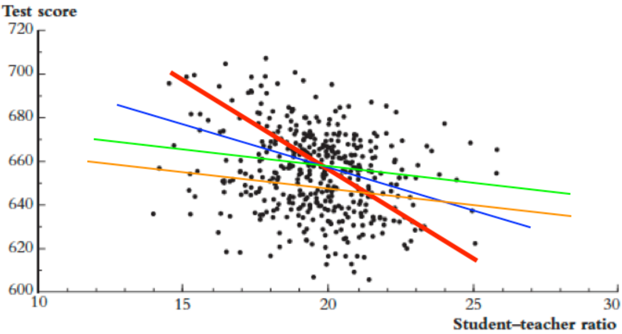


How to find the "best" fitting line?

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

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.



The Ordinary Least Squares Estimator (OLS)

- **The OLS estimator**
  - chooses the 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 $b_0 \equiv \hat{\beta}_0, b_1 \equiv \hat{\beta}_1$
  - The predicted value of Y_i given X_i using these estimators is $b_0 + b_1 X_i$, formally denotes as $\hat{Y}_i$
  - The prediction mistake is *the difference* between Y_i and $\hat{Y}_i$

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

  - The estimators of the slope and intercept that *minimize the sum of the squares of $\hat{u}_i$, thus*

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

are called the **ordinary least squares (OLS) estimators** of β_0 and β_1.

The Ordinary Least Squares Estimator (OLS)

- OLS minimizes sum of squared prediction mistakes: $\underset{b_0,b_1}{min}\sum_{i=1}^{n}\hat{u}_i^2=\sum_{i=1}^{n}(Y_i-b_0-b_1X_i)^2$
- Solve the problem by F.O.C(the first order condition)
  - Step 1:

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i-b_0-b_1X_i)^2=0$$

  - Step 2:

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i-b_0-b_1X_i)^2=0$$

Step 1: OLS estimator of β_0

- Optimization
-

$$\begin{aligned}\frac{\partial}{\partial b_0}\sum_{i=1}^{n}\hat{u}_i^2 = & \quad -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0 \\ \Rightarrow & \quad \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1 X_i = 0 \\ \Rightarrow & \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} b_0 - b_1\frac{1}{n}\sum_{i=1}^{n} X_i = 0 \\ \Rightarrow & \quad \overline{Y} - b_0 - b_1\overline{X} = 0\end{aligned}$$

OLS estimator of β_0:

$$\mathbf{b_0} = \overline{\mathbf{Y}} - \mathbf{b_1}\overline{\mathbf{X}} \; or \; \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$$

Step 2: OLS estimator of β_1

$$\begin{aligned}
\frac{\partial}{\partial b_1}\sum_{i=1}^{n}\hat{u}_i^2 = & \quad -2\textstyle\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) && = 0 \\
& \textstyle\sum_{i=1}^{n} X_i[Y_i - (\overline{Y} - b_1\overline{X}) - b_1 X_i)] && = 0 \\
\Rightarrow & \quad \textstyle\sum_{i=1}^{n} X_i[(Y_i - \overline{Y}) - b_1(X_i - \overline{X})] && = 0 \\
\Rightarrow & \quad \textstyle\sum_{i=1}^{n} X_i(Y_i - \overline{Y}) - b_1\sum_{i=1}^{n} X_i(X_i - \overline{X}) && = 0 \\
\Rightarrow & \quad \textstyle\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) - b_1\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X}) && = 0
\end{aligned}$$

Step 2: OLS estimator of β_1

- Algebra Trick

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y}) &= \sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}X_i\overline{Y}-\sum_{i=1}^{n}\overline{X}Y_i+\sum_{i=1}^{n}\overline{XY}\\ &= \sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}X_i\overline{Y}-n\overline{X}(\frac{1}{n}\sum_{i=1}^{n}Y_i)+n\overline{XY}\\ &= \sum_{i=1}^{n}X_i(Y_i-\overline{Y})\end{aligned}$$

- By a similar reasoning, we could obtain

$$\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})=\sum_{i=1}^{n}X_i(X_i-\overline{X})$$

Step 2: OLS estimator of β_1

- Thus

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}\hat{u}_i^2=\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})-b_1\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})=0$$

OLS estimator of β_1

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

- Recall the OLS **predicted values** $\hat{Y}_i$ and **residuals** $\hat{u}_i$ are:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0+\hat{\beta}_1X_i\\ \hat{u}_i &= Y_i-\hat{Y}_i\end{aligned}$$

The Estimated Regression Line

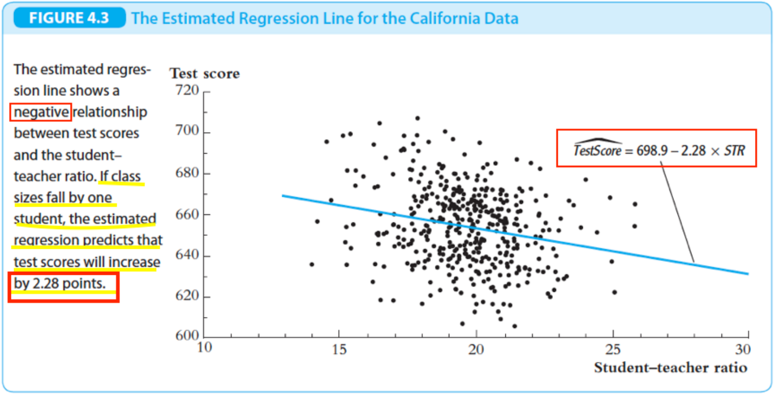


FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student–teacher ratio. If class sizes fall by one student, the estimated regression predicts that test scores will increase by 2.28 points.

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

Measures of Fit: The R^2

- Decompose Y_i into the fitted value plus the residual

$$Y_i = \hat{Y}_i + \hat{u}_i$$

- The total sum of squares (SST) = the explained sum of squares (SSE) + the sum of squared residuals (SSR):

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

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

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

So $0 \leq R^2 \leq 1$.

- It seems that T-squares is bigger, the regression is better.
- But actually we don't care much about R^2 in modern applied econometrics.

The Least Squares Assumptions

Assumption of the Linear regression model

- In order to investigate the statistical properties of OLS, we need to make some statistical assumptions

Linear Regression Model

The observations, (Y_i, X_i) come from a random sample(i.i.d) and satisfy the linear regression equation,

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

and $E[u_i \mid X_i] = 0$

Conditional Mean is Zero

Assumption 1: Zero conditional mean of the errors given X

The error,u_i has expected value of 0 given any value of the independent variable

$$E[u_i \mid X_i = x] = 0$$

- an weaker condition that u_i and X_i are *uncorrelated*:

$$Cov[u_i, X_i] = E[u_i X_i] = 0$$

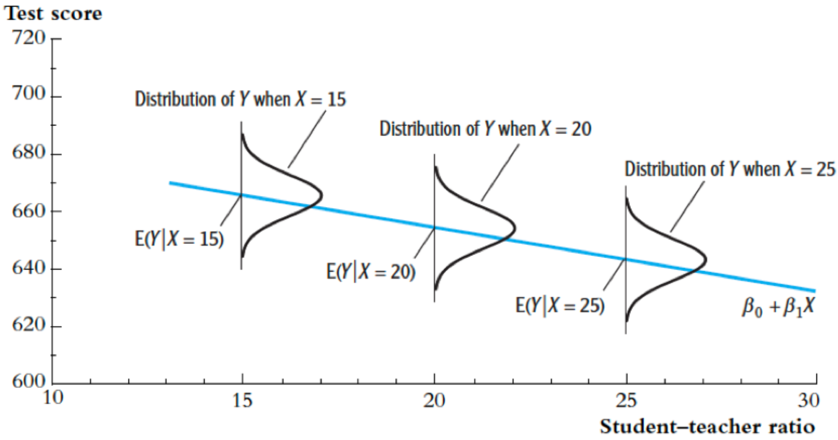
  - if both are correlated, then Assumption 1 is violated.
- Equivalently, the population regression line is the conditional mean of Y_i given X_i , thus

$$E(\,Y \mid X = x) = \beta_0 + \beta_1 X_i$$

(Exercise 4.6)

Conditional Mean is Zero

FIGURE 4.4 The Conditional Probability Distributions and the Population Regression Line



The figure shows the conditional probability of test scores for districts with class sizes of 15, 20, and 25 students. The mean of the conditional distribution of test scores, given the student–

Random Sample

Assumption 2: Random Sample

We have a i.i.d random sample of size , $\{(X_i, Y_i), i = 1, ..., n\}$ from the population regression model above.

- This is an implication of random sampling.
- generally won't hold in other data structures.
  - Violations: time-series, selected samples.

Large outliers are unlikely

Assumption 3: Large outliers are unlikely

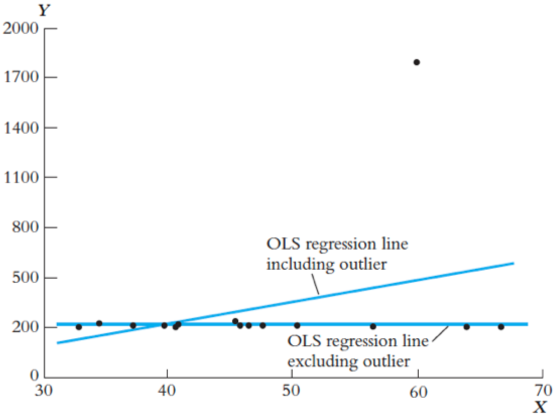
observations with values of X_i, Y_i or both that are far outside the usual range of the data(Outlier)-are unlikely. mathematically, it assume that X and Y have nonzero finite fourth moments.

- Large outliers can make OLS regression results misleading.
- One source of large outliers is data entry errors, such as a typographical error or incorrectly using different units for different observations.
- Data entry errors aside, the assumption of finite kurtosis is a plausible one in many applications with economic data.

Large outliers are unlikely

FIGURE 4.5 The Sensitivity of OLS to Large Outliers

This hypothetical data set has one outlier. The OLS regression line estimated with the outlier shows a strong positive relationship between *X* and *Y*, but the OLS regression line estimated without the outlier shows no relationship.



Underlying assumptions of OLS

- The OLS estimator is **unbiased**, **consistent** and has **asymptotically normal sampling distribution** if
  1. Random sampling.
  2. Large outliers are unlikely.
  3. The conditional mean of ui given Xi is zero
- OLS is an estimator—it's a machine that we plug data into and we get out estimates.
- It has a sampling distribution, with a sampling variance/standard error, etc.
  - Just like the sample mean, sample difference in means, or the sample variance
- We will discuss these characteristics of OLS in the next lecture.