

Lecture 5: Linear Regression with One Regressor

Zheng Tian

March 28, 2016

1 Introduction

This chapter introduces the linear regression model relating one variable, X , to another, Y , which is called the linear regression model with one regressor or the simple linear regression model. This chapter describes the ordinary least squares (OLS) methods for estimating a simple linear regression model, discusses the algebraic and statistical properties of the OLS estimator, introduces two measures of fit, and lays out three least squares assumptions for the linear regression model. We apply the OLS estimation method to a linear model of test scores and class sizes in California school districts.

This chapter lays out foundations for all chapters that follow. Although in practice we seldom use a linear regression model with only one regressor, the essential principles of the OLS estimation method are the same for a linear regression model with multiple regressors.

2 The Linear Regression Model

2.1 What is regression?

Definition of *regress* in Merriam-Webster's dictionary

- 1** an act or the privilege of going or coming back
- 2** movement backward to a previous and especially worse or more primitive state or condition
- 3** the act of reasoning backward

The meaning of regression in statistics?

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables.¹ Specifically, most regression analysis focus on the conditional mean of the dependent variable given the independent variables, which is a function of x . A very simple but common form of the function is a linear function, i.e., $\beta_0 + \beta_1 x$, such as

$$E(Y|X = x) = f(x) = \beta_0 + \beta_1 x$$

2.2 An example: Test scores versus class size

The background of the story

The superintendent of an elementary school district is facing a trade-off:

- Hiring more teachers to improve students' test performance
- Keeping the budget tight

Research question:

Can reducing class size increase students' test scores?

How can we answer the question?

- We randomly choose 42 students and divide them into two classes, with one having 20 students and another having 22. And they are taught with the same subject and by the same teachers.
- Randomization ensures that it is only the difference in class sizes of the two classes that will affect test scores.
- After a test for both classes, we then compute the average test scores which can be expressed as,

$$E(\text{TestScore}|\text{ClassSize} = 20)$$

$$E(\text{TestScore}|\text{ClassSize} = 22)$$

- Then the effect of class size on test scores can be reflected as

$$E(\text{TestScore}|\text{ClassSize} = 20) - E(\text{TestScore}|\text{ClassSize} = 22)$$

¹Read the article in Wikipedia https://en.wikipedia.org/wiki/Regression_analysis

- If the difference is large enough, we can say that reducing class can improve students' test performance.
- Question: how can we relate test scores to class sizes? The key is in $E(\text{TestScore}|\text{ClassSize})$, which dictates a regression function.

A general model of test scores v.s. class size

Generally, we can write the **population regression line** as

$$E(\text{TestScore}|\text{ClassSize}) = \beta_0 + \beta_1 \text{ClassSize} \quad (1)$$

By calculating the conditional expectation, some other factors apart from class sizes may be canceled out. These factors may also influence test scores, but they are either unimportant in your reasoning or unable to be measured. We can lump all these factors into a single term, and set up a general regression model as

$$\text{TestScore} = \beta_0 + \beta_1 \text{ClassSize} + \text{OtherFactors} \quad (2)$$

If we assume $E(\text{OtherFactors}|\text{ClassSize}) = 0$, then the linear regression model (Eq. 2) becomes the population regression line (Eq. 1).

Explore the data with a scatterplot

The relationship between the data points, the population regression line, and the errors (other factors) are illustrated in Figure 1.

2.3 The general linear regression model

Consider two variables Y and X . For both, there are n observations so that each observation $i (= 1, 2, 3, \dots)$ is associated with a pair of values of Y_i and X_i . What we are interested in is how a unit change in X will affect Y . For that, we can set up a linear regression model as

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \text{ for } i = 1, \dots, n \quad (3)$$

- Y_i : the dependent variable, the regressand, or the LHS (left-hand side) variable.
- X_i : the independent variable, the regressor, or the RHS (right-hand side) variable.

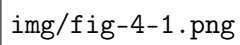
The figure is a placeholder for a graph illustrating the Population Regression Line. It is represented by the text 'img/fig-4-1.png' within a rectangular frame.

Figure 1: The Population Regression Line

- β_0 : the intercept, or the constant term. It can either have economic meaning or have merely mathematical sense, which determines the level of the regression line, that is, the point of intersection with the Y axis.
- β_1 : is the slope of the population regression line. $\beta_1 = dY_i/dX_i$ is the marginal effect of X on Y , that is, one unit change in X changes Y by β_1 units, holding other things constant.
- u_i : is the error term. $u_i = Y_i - (\beta_0 + \beta_1 X_i)$ incorporates all the other factors besides X that determine the value of Y .
- $\beta_0 + \beta_1 X_i$: the population regression function/line. We can predict the value of Y given a value of X according to the population regression line.

3 The OLS Estimation Method for a Linear Regression Model

3.1 The intuition for the OLS and minimization

The most common method to estimate a linear regression model, like Equation 3, is to use the ordinary least squares (OLS) estimator.²

Let's explain the basic idea of the OLS by dissecting its name.

Ordinary It means that the OLS estimator is a very basic method, from which we may derive some variations of the OLS estimator, such as the weighted least squares (WLS), and the generalized least squares (GLS).

Least It means that the OLS estimator tries to minimize something. The "something" is the mistakes we make when we try to guess (estimate) the values of the parameters in the model. From Equation 3, if our guess for β_0 and β_1 is b_0 and b_1 , then the mistake of our guess is $\hat{u}_i = Y_i - b_0 - b_1 X_i$.

Squares It represent the actual thing (a quantity) that we minimize. The OLS does not attempt to minimize each \hat{u}_i but to minimize the sum of the squared mistakes, $\sum_{i=1}^n \hat{u}_i^2$. Taking square is to avoid possible offsetting between positive and negative values of \hat{u}_i in $\sum_i \hat{u}_i$.

3.2 The OLS estimators for β_0 and β_1

Let b_0 and b_1 be some estimators of β_0 and β_1 , respectively. Then, the OLS estimator is the solution to the following minimization problem.

²Recall that an **estimator** is a function of a sample of data. An **estimate** is the numerical value of the estimator when it is computed using data from a sample.

$$\min_{b_0, b_1} S(b_0, b_1) = \sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 \quad (4)$$

where $S(b_0, b_1)$ is a function of b_0 and b_1 , measuring the sum of the squared prediction mistakes over all n observation.

Proof. We solve the problem by taking the derivative of $S(b_0, b_1)$ with respect to b_0 and b_1 , respectively. Then the first order conditions evaluated at the value of a solution $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\frac{\partial S}{\partial b_0}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (-2)(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \quad (5)$$

$$\frac{\partial S}{\partial b_1}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (-2)(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = 0 \quad (6)$$

Rearranging Equation 5, we get

$$\begin{aligned} \sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n X_i &= 0 \\ \text{then } \hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n Y_i - \frac{\hat{\beta}_1}{n} \sum_{i=1}^n X_i = \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned} \quad (7)$$

Rearranging Equation 6 and plugging Equation 7, we get

$$\begin{aligned} \sum_{i=1}^n X_i Y_i - \hat{\beta}_0 \sum_{i=1}^n X_i - \hat{\beta}_1 \sum_{i=1}^n X_i^2 &= 0 \\ \sum_{i=1}^n X_i Y_i - \frac{1}{n} \sum_{i=1}^n X_i \sum_{i=1}^n Y_i + \hat{\beta}_1 \frac{1}{n} \left(\sum_{i=1}^n X_i \right)^2 - \hat{\beta}_1 \sum_{i=1}^n X_i^2 &= 0 \\ \hat{\beta}_1 &= \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2} \end{aligned} \quad (8)$$

Collecting the terms in the numerator and denominator in Equation 8, we have³

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

□

³We can show the derivation reversely. For the numerator in Equation 8, we can show the following

In summary, solving the minimization problem (Equation 4), we obtain the OLS estimator of β_0 and β_1 as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (9)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (10)$$

3.3 The predicted values, residuals, and the sample regression line

After the estimator, we can compute the **predicted values** \hat{Y}_i and **residuals** \hat{u}_i for $i = 1, \dots, n$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (11)$$

$$\hat{u}_i = Y_i - \hat{Y}_i \quad (12)$$

And the line represented by Equation 11 is called **the sample regression line**. From Equation 10, we see that the sample average point (\bar{X}, \bar{Y}) is on the sample regression line.

3.4 The OLS estimates of the relationship between test scores and the student-teacher ratio

Let's go back to the application of test scores versus the student-teacher ratios in California school districts. The goal is to estimate the effect of class sizes, measured by the student-teacher ratios, on test scores. Before setting up a formal regression model, it is always a good practice to glance over the data using some exploratory data analysis techniques.

Exploratory analysis

Basic summary statistics We first need to compute basic summary statistics to see the sample distribution of the data. Some commonly used summary statistics include

$$\begin{aligned} \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_i X_i Y_i - \bar{X} \sum_i Y_i - \bar{Y} \sum_i X_i + \sum_i \bar{X} \bar{Y} \\ &= \sum_i X_i Y_i - 2n\bar{X}\bar{Y} + n\bar{X}\bar{Y} \\ &= \sum_i X_i Y_i - n\bar{X}\bar{Y} \\ &= \frac{1}{n} (n \sum_i X_i Y_i - \sum_i X_i \sum_i Y_i) \end{aligned}$$

Similarly, we can show that $\sum_i (X_i - \bar{X})^2 = \frac{1}{n} (n \sum_i X_i^2 - (\sum_i X_i)^2)$

mean, standard deviation, median, minimum, maximum, and quantile (percentile), etc. Table 1 summarizes the distribution of test scores and class sizes for the sample.

L^AT_EX % latex table generated in R 3.2.2 by xtable 1.8-0 package % Wed Mar 23 15:51:53 2016

Table 1: Summary of the Distribution of Student-Teacher Ratios and Fifth-Grade Test Scores for 420 K-8 Districts in California in 1999

	Ave	Std	Percentiles						
			10%	25%	40%	50%	60%	75%	90%
Test score	654.16	19.05	630.39	640.05	649.07	654.45	659.40	666.66	678.86
Student-teacher ratio	19.64	1.89	17.35	18.58	19.27	19.72	20.08	20.87	21.87

Scatterplot A scatterplot visualizes the relationship between two variables straightforwardly, which is helpful for us to decide what a functional form a regression model should properly take. Figure 2 shows that test scores and student-teacher ratios may be negatively related. The correlation coefficient between the two variables is -0.23, verifying the existence of a weak negative relationship.

Regression analysis

After exploratory analysis, we can estimate the linear model. Although the formula of computing β_1 and β_0 (Equations 9 and 10) seems complicated, the practical estimation procedure is simplified by using computer software, like R, Stata, and Eviews, which mostly involving just one-line command or just a few clicking of the mouse. We'll see how R estimates a linear regression model shortly. For now, let's simply present the estimation results in the following equation,

$$\widehat{TestScore} = 698.93 - 2.28 \times STR \quad (13)$$

We can draw the sample regression line represented by Equation 13 in the scatterplot to eyeball how well the regression model fits the data.

Interpretation of the estimated coefficients

Obtaining the estimates of the coefficients in the regression model is not the end of a regression analysis. What need to do next includes hypothesis tests, model specification tests, robustness (or sensitivity) test, and interpretation. Let's first see how to correctly interpret the estimation results.

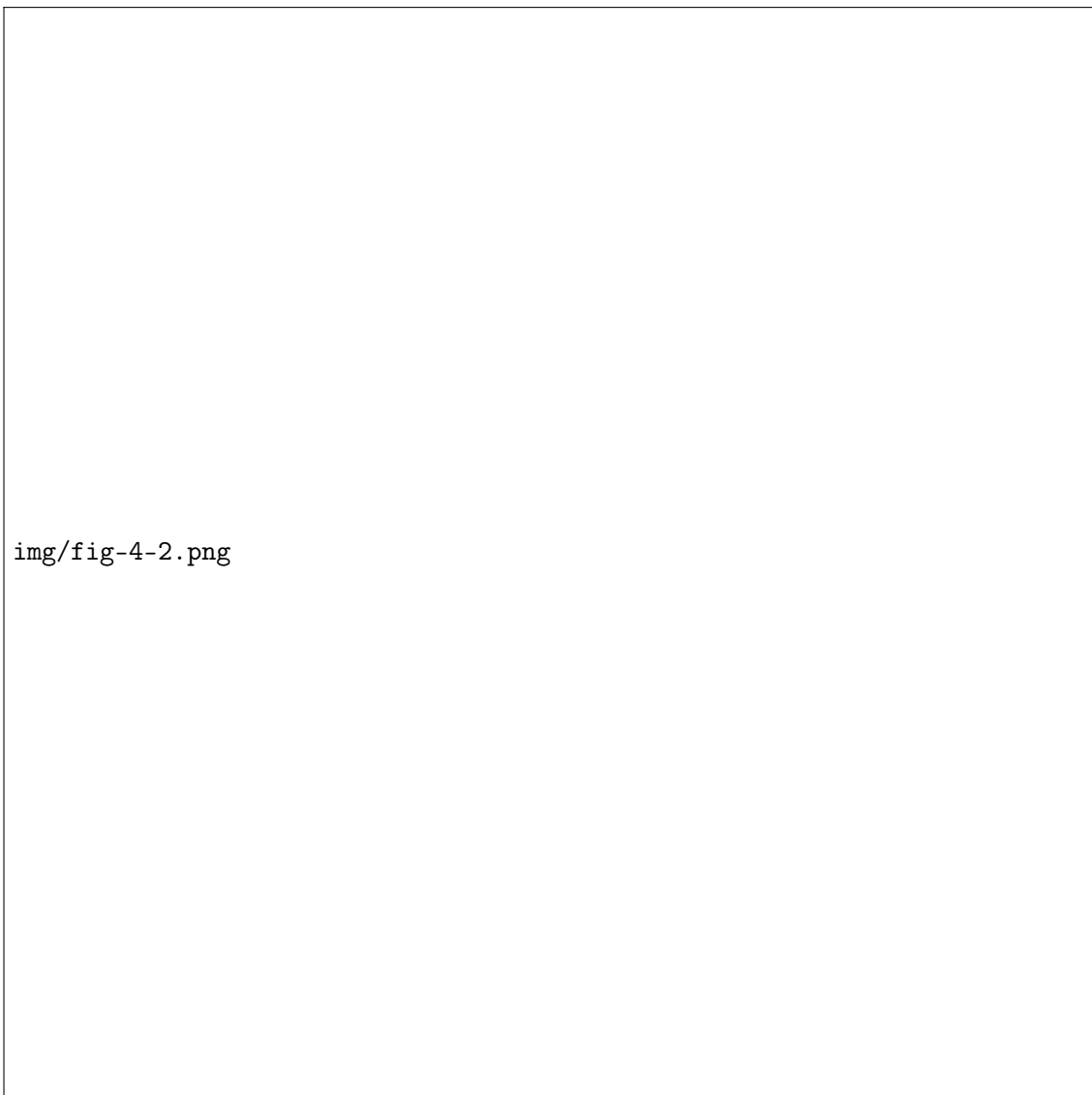


Figure 2: The scatterplot between student-teacher ratios and test scores

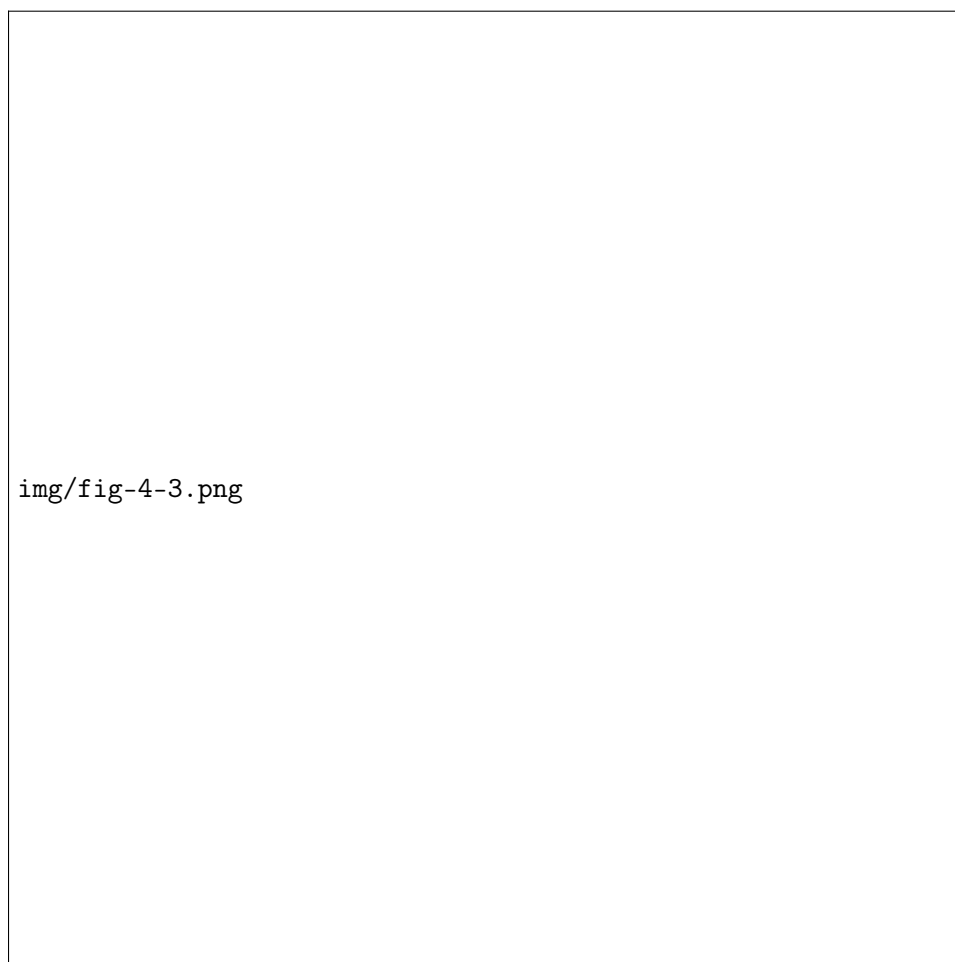


Figure 3: The estimated regression line for the California data

- What interests us the most is the slope that tell us how much a unit change in student-teacher ratios will cause test scores to change. The slope of -2.28 means that an increase in the student-teacher ratio by one student per class is, on average, associated with a decline in district-wide test scores by 2.28 points on the test.
- The intercept literally means that if the student-teacher ratio is zero, the average district-wide test scores will be 698.9. However, it is nonsense for having some positive test scores when the student-teacher ratio is zero. Therefore, the intercept term in this case merely serves as determining the level of the sample regression line.
- The mere number of -2.28 really does not make much sense for the readers of your research. We have to put it into the context of California school district to avoid ridiculous results even though the estimation itself is correct. (Read the discussion in the paragraphs in Page 117.)

4 Algebraic Properties of the OLS Estimator

4.1 TSS, ESS, and SSR

- From $Y_i = \hat{Y}_i + \hat{u}_i$, we can define
 - **the total sum of squares:** $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$
 - **the explained sum of squares:** $ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
 - **the sum of squared residuals:** $SSR = \sum_{i=1}^n \hat{u}_i^2$

Note that TSS, ESS, and SSR all take the form of "deviation from the mean". This form is only valid when an intercept is included in the regression model.⁴

⁴We are not going to prove this because it involves higher level knowledge of linear algebra. You can estimate a linear regression model of $Y_i = \beta_1 X_i + u_i$, for which TSS is simply $\sum_i Y_i^2$ and ESS is $\sum_i \hat{Y}_i^2$. Also, for this model, $\sum_i \hat{u}_i \neq 0$.

4.2 Some algebraic properties among \hat{u}_i , \hat{Y}_i , Y_i , and X_i

The OLS residuals and the predicted values satisfy the following equations:⁵

$$\sum_{i=1}^n \hat{u}_i = 0 \quad (14)$$

$$\frac{1}{n} \sum_{i=1}^n \hat{Y}_i = \bar{Y} \quad (15)$$

$$\sum_{i=1}^n \hat{u}_i X_i = 0 \quad (16)$$

$$TSS = ESS + SSR \quad (17)$$

4.3 The proof of these properties

Here, I just prove Equation 17. The proofs for the other equations above are in Appendix 4.3 in the textbook.

Proof. (a) Prove Equation 14. From Equation 10 we can write the OLS residuals as $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})$. Thus

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

By definition of the sample average, we have $\sum_{i=1}^n (Y_i - \bar{Y}) = 0$ and $\sum_{i=1}^n (X_i - \bar{X}) = 0$. It follows that $\sum_{i=1}^n \hat{u}_i = 0$.

(b) To verify Equation 15, note that $Y_i = \hat{Y}_i + \hat{u}_i$, so $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i + \sum_{i=1}^n \hat{u}_i = \sum_{i=1}^n \hat{Y}_i$. It follows that $(1/n) \sum_{i=1}^n \hat{Y}_i = \bar{Y}$.

(c) To verify Equation 16, note that $\sum_{i=1}^n \hat{u}_i = 0$ implies that $\sum_{i=1}^n \hat{u}_i X_i = \sum_{i=1}^n \hat{u}_i (X_i - \bar{X}) = \sum_{i=1}^n [(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})] (X_i - \bar{X}) = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X})^2 = 0$

(d) Prove $TSS = ESS + SSR$.

⁵Equation 14 holds only for a linear regression model with an intercept, but Equation 16 holds regardless of the presence of an intercept.

$$\begin{aligned}
TSS &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\
&= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\
&= SSR + ESS + 2 \sum_{i=1}^n \hat{u}_i \hat{Y}_i \\
&= SSR + ESS + 2 \sum_{i=1}^n \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 X_i) \\
&= SSR + ESS + 2(\hat{\beta}_0 \sum_{i=1}^n \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^n \hat{u}_i X_i) \\
&= SSR + ESS
\end{aligned}$$

where the final equality follows from Equations 14 and 16. \square

5 Measures of Fit

5.1 Goodness of Fit: R^2

R^2 is one of the commonly used measures of how well the OLS regression line fits the data. R^2 is the fraction of the sample variance of Y_i explained by X_i . The sample variance can be represented with TSS and the part of sample variance explained by X can be represented by ESS . Therefore, mathematically, we can define R^2 as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \quad (18)$$

Properties of R^2

(1) $R^2 \in [0, 1]$

$R^2 = 0$ when $\hat{\beta}_1 = 0$, that is, X cannot explain the variance in Y .

$$\hat{\beta}_1 = 0 \Rightarrow Y_i = \hat{\beta}_0 + \hat{u}_i \Rightarrow \hat{Y}_i = \bar{Y} = \hat{\beta}_0 \Rightarrow ESS = \sum_i^n (\hat{Y}_i - \bar{Y})^2 = 0 \Rightarrow R^2 = 0$$

$R^2 = 1$ when $\hat{u}_i = 0$ for all $i = 1, \dots, n$, that is, the regression line fits all the sample data perfectly.

$$\hat{u}_i = 0 \Rightarrow ESS = \sum_i^n \hat{u}_i^2 = 0 \Rightarrow R^2 = 1$$

$$(2) R^2 = r_{XY}^2$$

r_{XY} is the sample correlation coefficient, that is,

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{(\sum_i^n (X_i - \bar{X})^2 \sum_i^n (Y_i - \bar{Y})^2)^{1/2}}$$

Note: This property holds only for the linear regression model with a regressor and an **intercept**.

The use of R^2

- R^2 is usually the first statistics that we look at for judging how well the regression model fits the data.
- Most computer programs for econometrics and statistics report R^2 in their estimation results.
- However, we cannot merely rely on R^2 for judge whether the regression model is "good" or "bad". For that, we have to use some statistics that will be taught soon.

5.2 The standard error of regression (SER) as a measure of fit

Like R^2 , the standard error of regression (SER) is another measure of fit for the OLS regression.

$$\text{SER} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2} = s \quad (19)$$

- SER has the units of u , which are the units of Y .
- SER measures the average "size" of the OLS residual (the average "mistake" made by the OLS regression line).
- The root mean squared error (RMSE) is closely related to the SER:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=2}^n \hat{u}_i^2}$$

As $n \rightarrow \infty$, $\text{SER} = \text{RMSE}$.

5.3 R^2 and SER for the application of test scores v.s. class sizes

- In the application of test scores v.s. class sizes, R^2 is 0.051 or 5.1%, which implies that the regressor *STR* explains only 5.1% of the variance of the dependent variable *TestScore*.
- SER is 18.6, which means that standard deviation of the regression residuals is 18.6 points on the test. The magnitude of SER is so large that, in another way, shows that the simple linear regression model does not fit the data well.

6 The Least Squares Assumptions

6.1 Assumption 1: The conditional mean of u_i given X_i is zero

$$E(u_i|X_i) = 0 \tag{20}$$

If Equation 20 is satisfied, then X_i is called **exogenous**. This assumption can be a little stronger as $E(u|X = x) = 0$ for any value x , that is $E(u_i|X_1, \dots, X_n) = 0$.

Since $E(u|X = x) = 0$, it follows that $E(u) = E(E(u|X)) = E(0) = 0$. The unconditional mean of u is also zero.

- A benchmark for thinking about this assumption is to consider an ideal randomized controlled experiment.

Because X is assigned randomly, all other individual characteristics – the things that make up u – are distributed independently of X , so u and X are independent. Thus, in an ideal randomized controlled experiment, $E(u|X = x) = 0$

- In actual experiments, or with observational data, we will need to think hard about whether $E(u|X = x) = 0$ holds

Assumption 1 can be illustrated by Figure 4.

Correlation and conditional mean

$$E(u_i|X_i) = 0 \Rightarrow \text{Cov}(u_i, X_i) = 0$$

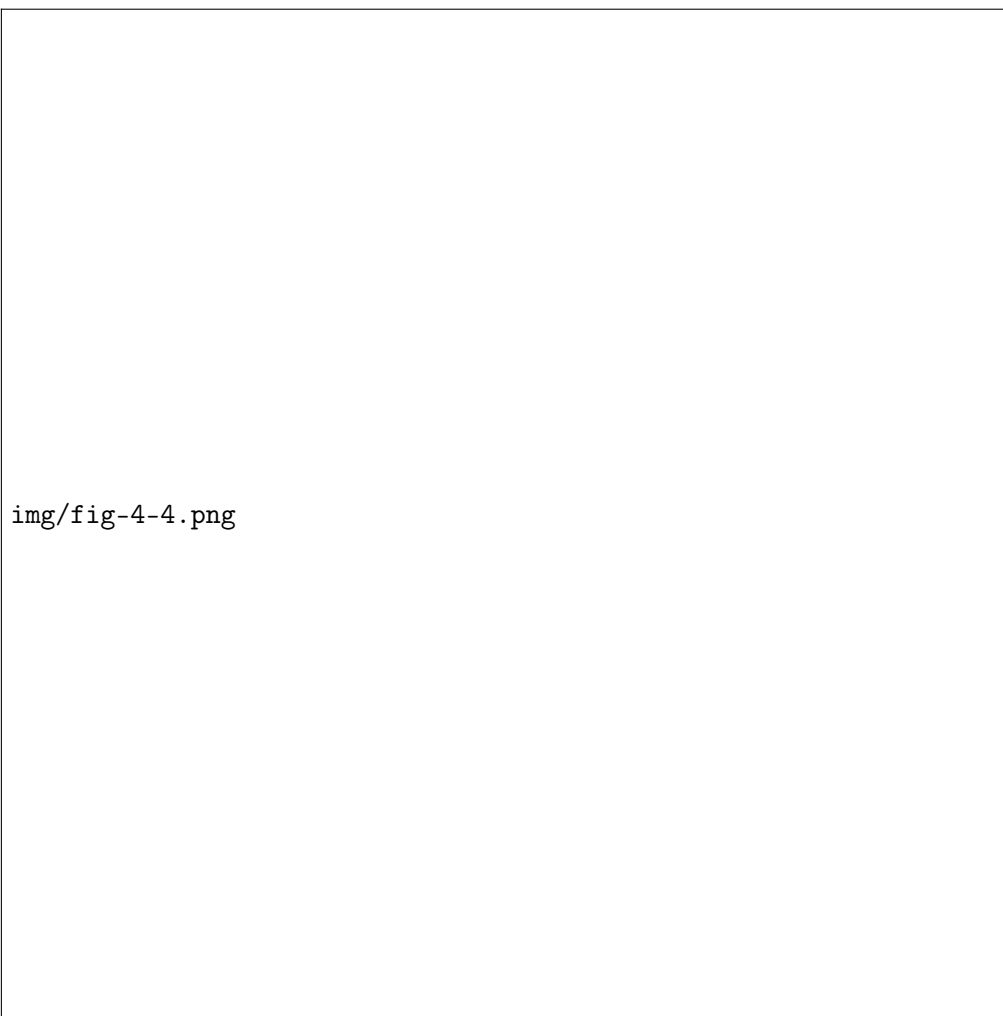


Figure 4: An illustration of $E(u|X = x) = 0$

Proof.

$$\begin{aligned}\text{Cov}(u_i, X_i) &= E(u_i X_i) - E(u_i)E(X_i) \\ &= E(E(u_i X_i | X_i)) - E(E(u_i | X_i))E(X_i) \\ &= E(X_i E(u_i | X_i)) - 0 \cdot E(X_i) \\ &= 0\end{aligned}$$

where the law of iterated expectation is used twice at the second equality. \square

It follows that $\text{Cov}(u_i, X_i) \neq 0 \Rightarrow E(u_i | X_i) \neq 0$.

6.2 Assumption 2: (X_i, Y_i) for $i = 1, \dots, n$ are i.i.d.

- The individual sample i , with which (X_i, Y_i) is associated, is selected randomly from the same joint distribution
- Since $u_i = Y_i - \beta_0 - \beta_1 X_i$, u_i is i.i.d.
- The violation of the i.i.d. assumption: time series data, $\text{Cov}(Y_t, Y_{t-1}) \neq 0$

6.3 Assumption 3: large outliers are unlikely

$$0 < E(X_i^4) < \infty \text{ and } 0 < E(Y_i^4) < \infty$$

- A large outlier is an extreme value of X or Y
- On a technical level, if X and Y are bounded, then they have finite fourth moments.
- The substance of this assumption is that a large outlier can strongly influence the results – so we need to rule out large outliers.

The influential observations and the leverage effects

- An outlier can be detected by a scatterplot. See Figure 5.
- There are also formal tests for the existence of the influential observations, some of which are coded in econometric software, like R and Stata.

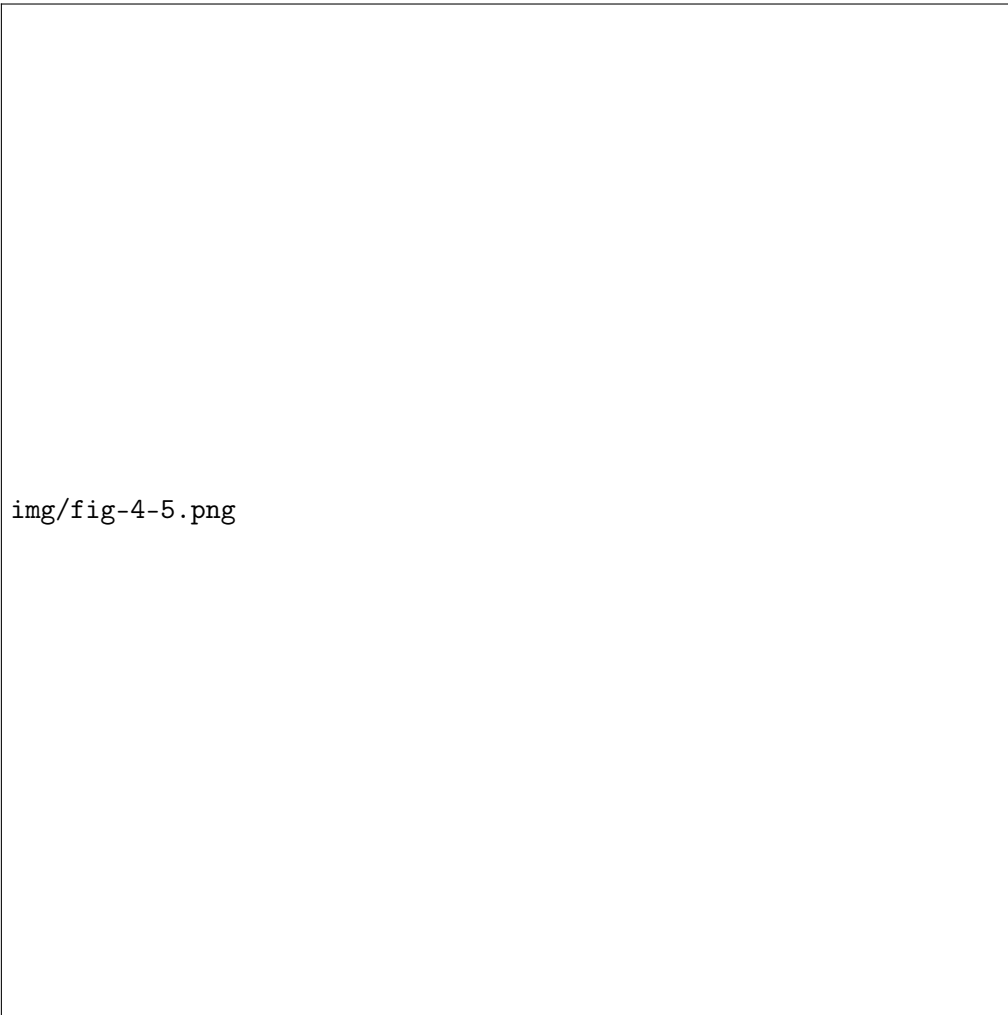


Figure 5: How an outlier can influence the OLS estimates

7 Sampling Distribution of the OLS Estimators

7.1 Unbiasedness and consistency

The unbiasedness of $\hat{\beta}_0$ and $\hat{\beta}_1$

- The randomness of $\hat{\beta}_0$ and $\hat{\beta}_1$

Since (X_i, Y_i) for $i = 1, \dots, n$ are randomly drawn from a population, different draws can render different estimates, giving rise to the randomness of $\hat{\beta}_0$ and $\hat{\beta}_1$.

- The unbiasedness of $\hat{\beta}_0$ and $\hat{\beta}_1$

Let the true values of the intercept and the slope be β_0 and β_1 . Based on the least squares assumption #1: $E(u_i|X_i) = 0$

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1$$

- Show that $\hat{\beta}_1$ is unbiased

Let's rewrite the formula of $\hat{\beta}_1$ here

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (21)$$

Proof. We first represent $\hat{\beta}_1$ with β_1, X_i , and u_i

Because $Y_i = \beta_0 + \beta_1 X_i + u_i$, $Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + u_i - \bar{u}$, which is then plugged in the numerator in Equation (21). Then,

$$\begin{aligned} \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_i (X_i - \bar{X}) [\beta_1(X_i - \bar{X}) + (u_i - \bar{u})] \\ &= \beta_1 \sum_i (X_i - \bar{X})^2 + \sum_i (X_i - \bar{X})u_i - \bar{u} \sum_i (X_i - \bar{X}) \\ &= \beta_1 \sum_i (X_i - \bar{X})^2 + \sum_i (X_i - \bar{X})u_i \end{aligned}$$

Substituting this expression in Equation (21) yields

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n} \sum_i (X_i - \bar{X})u_i}{\frac{1}{n} \sum_i (X_i - \bar{X})^2} \quad (22)$$

We prove that $\hat{\beta}_1$ is conditionally unbiased, from which the unconditional unbiased-

ness follows naturally.

$$\begin{aligned} E(\hat{\beta}_1|X_1, \dots, X_n) &= \beta_1 + E \left\{ \left[\frac{\frac{1}{n} \sum_i (X_i - \bar{X}) u_i}{\frac{1}{n} \sum_i (X_i - \bar{X})^2} \right] | X_1, \dots, X_n \right\} \\ &= \beta_1 + \frac{\frac{1}{n} \sum_i (X_i - \bar{X}) E(u_i|X_1, \dots, X_n)}{\frac{1}{n} \sum_i (X_i - \bar{X})^2} \\ &= \beta_1 \quad (\text{by assumption 1}) \end{aligned}$$

It follows that

$$E(\hat{\beta}_1) = E(E(\hat{\beta}_1|X_1, \dots, X_n)) = \beta_1$$

Therefore, $\hat{\beta}_1$ is an unbiased estimator of β_1 . □

The proof of unbiasedness of $\hat{\beta}_0$ is left for exercise.

The consistency of $\hat{\beta}_0$ and $\hat{\beta}_1$

$\hat{\beta}$ is said to be a consistent estimator of β if as n goes to infinity, $\hat{\beta}$ is in probability close to β , which can be denoted as $n \rightarrow \infty, \hat{\beta} \xrightarrow{p} \beta$, or simply as $\text{plim}_{n \rightarrow \infty} \hat{\beta} = \beta$.

And the law of large number states that for random i.i.d. samples x_1, \dots, x_n , if $E(x_i) = \mu$ and $\text{Var}(x_i) < \infty$, then $\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_i x_i = \mu$.

Then we can show that $\text{plim}_{n \rightarrow \infty} \hat{\beta}_1 = \beta_1$.

The proof is not required to understand for this course.

Proof. From Equation (22) we can have

$$\text{plim}_{n \rightarrow \infty} (\hat{\beta}_1 - \beta_1) = \text{plim}_{n \rightarrow \infty} \frac{\frac{1}{n} \sum_i (X_i - \bar{X}) u_i}{\frac{1}{n} \sum_i (X_i - \bar{X})^2} = \frac{\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_i (X_i - \bar{X}) u_i}{\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_i (X_i - \bar{X})^2}$$

The denominator of the last equality is just a consistent estimator of the sample variance of X_i , that is, $\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_i (X_i - \bar{X})^2 = \sigma_X^2$

Now we need to focus on $\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_i (X_i - \bar{X}) u_i$. To apply the law of large numbers, we need to find the expectation of $(X_i - \bar{X}) u_i$. Given that $E(X_i u_i) = E(E(X_i u_i | X_i)) = E(X_i E(u_i | X_i)) = 0$, we have

$$E((X_i - \bar{X}) u_i) = E(X_i u_i) + \frac{1}{n} \sum_i E(X_i u_i) = 0 + 0 = 0$$

So the variance of $(X_i - \bar{X})u_i$ can be expressed as

$$\begin{aligned}
\text{Var}((X_i - \bar{X})u_i) &= E((X - \bar{X})^2 u_i^2) \\
&= E(E((X - \bar{X})^2 u_i^2 | X)) \\
&= E((X - \bar{X})^2 E(u_i^2 | X)) \\
&= E((X - \bar{X})^2 \sigma_u^2) \quad (\text{by the extended assumption 4. See Chapter 17}) \\
&< \infty \quad (\text{by assumption 3})
\end{aligned}$$

Since $E((X_i - \bar{X})u_i) = 0$, $\text{Var}((X_i - \bar{X})u_i) < \infty$, and X_i, u_i for $i = 1, \dots, n$ are i.i.d, by the law of large numbers, we have

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_i (X_i - \bar{X})u_i = 0$$

Therefore, $\text{plim}_{n \rightarrow \infty} \hat{\beta}_1 = \beta_1$. □

Similarly, we can also prove that $\hat{\beta}_0$ is consistent, that is $\text{plim}_{n \rightarrow \infty} \hat{\beta}_0 = \beta_0$.

7.2 The asymptotic normal distribution

The central limit theory states that if X_1, \dots, X_n with the mean μ and the variance $0 < \sigma^2 < \infty$. Then, $\frac{1}{n} \sum_i X_i \xrightarrow{d} N(\mu, \frac{\sigma^2}{n})$.

From the proof of consistency, we have already seen that $E((X_i - \bar{X})u_i) = 0$, $\text{Var}((X_i - \bar{X})u_i) < \infty$, and X_i, u_i for $i = 1, \dots, n$ are i.i.d. By the central limit theory, we know that

$$\frac{1}{n} \sum_i (X_i - \bar{X})u_i \xrightarrow{d} N\left(0, \frac{1}{n} \text{Var}((X_i - \bar{X})u_i)\right)$$

It follows that from Equation (22) and the fact that $\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_i (X_i - \bar{X})^2 = \text{Var}(X_i)$, $\hat{\beta}_1$ is asymptotically normally distributed as

$$\hat{\beta}_1 \xrightarrow{d} N\left(\beta_1, \sigma_{\hat{\beta}_1}^2\right)$$

where

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n} \frac{\text{Var}((X_i - \bar{X})u_i)}{\text{Var}(X_i)^2} \quad (23)$$

Similarly, we can show that $\hat{\beta}_0 \xrightarrow{d} N(\beta_0, \sigma_{\hat{\beta}_0}^2)$, where

$$\sigma_{\hat{\beta}_0}^2 = \frac{1}{n} \frac{\text{Var}(H_i u_i)}{(E(H_i^2))^2}, \quad \text{where } H_i = 1 - \left(\frac{\mu_X}{E(X_i^2)}\right) X_i \quad (24)$$

- As $\text{Var}(X_i)$ increases, $\text{Var}(\hat{\beta}_1)$ decreases.

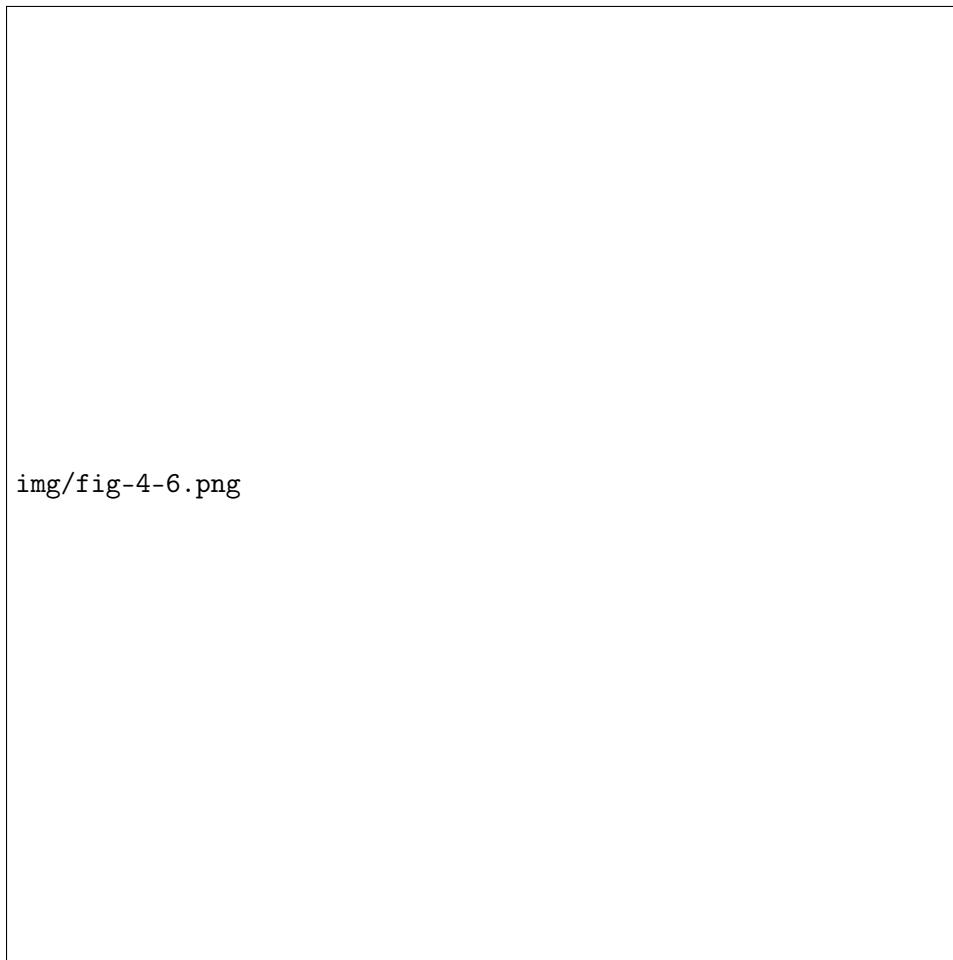


Figure 6: The Variance of $\hat{\beta}_1$ and the variance of X_i

- As $\text{Var}(u_i)$ increases, $\text{Var}(\hat{\beta}_1)$ increases.