Lecture 6: Linear Regression with One Regressor

Zheng Tian

Contents

1 Introduction

This lecture introduces a linear regression model with one regressor called a simple linear regression model. We will learn the ordinary least squares (OLS) method to estimate a simple linear regression model, discuss the algebraic and statistical properties of the OLS estimator, introduce two measures of goodness of fit, and bring up three least squares assumptions for a linear regression model. As an example, we apply the OLS estimation method to a linear model of test scores and class sizes in California school districts.

This lecture lays out foundations for all lectures to come. Although in practice we seldom use a linear regression model with only one regressor, the essential principles of the OLS estimation method and hypothesis testing 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

Merriam-Webster gives the following definition of the word "regress":

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 the values of independent variables.

A very simple functional form of a conditional expectation is a linear function. That is, we can model the conditional mean as follows,

$$\mathrm{E}(Y \mid X = x) = f(x) = \beta_0 + \beta_1 x \tag{1}$$

Equation 1 is called a **simple linear regression function**.

2.2 An example: Test scores versus class size

Let's go back to the example of California school districts, introduced in Lecture 1.

Research question:

The research question of this application is: 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 the difference in class sizes of the two classes that is the only factor affecting test scores.
- After a test for both classes, we then compute the average test scores that can be expressed as,

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

- Then the effect of class size on test scores is the difference in the conditional means, i.e.,

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

[1] Wikipedia, the free encyclopedia. Regression analysis. Retrieved from `https://en.wikipedia.org/wiki/Regression_analysis`

- If the difference is large enough, we can say that reducing class can improve students' test performance.

A simple linear regression model of test scores v.s. class size

As mentioned above, a simple linear regression function can be used to describe the relationship between test scores and class sizes. Since it regards the association between these two variable for the whole population, we call this regression function as the **population regression function** or the **population regression line**, taking the following form,

$$\mathrm{E}(TestScore|ClassSzie) = \beta_0 + \beta_1 ClassSize \tag{2}$$

By calculating the conditional expectation, some other factors, apart from class sizes, are left out of the population regression function. Although these factors may also influence test scores, 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 **simple linear regression model** as follows,

$$TestScore = \beta_0 + \beta_1 ClassSize + OtherFactors \tag{3}$$

If we assume $\mathrm{E}(OtherFactors|ClassSize) = 0$, then the simple linear regression model (Eq. 3) becomes the population regression line (Eq. 2).

A distinction between the population regression function and the population regression model

Note that here we have two concepts: the population regression function and the population regression model. What's their difference? Simply put,

- A population regression function gives us a deterministic relation between class size and the expectation of test scores. That is, when we have a value of class size and know the values of β_0 and β_1, there is one and only one expected value of test scores associated with this class size. However, we cannot compute the exact value of the test score of a particular observation.
- A population regression model, by including other factors, gives us a complete description of a data generating process (DGP). That is, when we have all the values of class sizes and other factors and know β_0 and β_1, we can generate all the values of test scores. Also, when we consider other factors as a random variable, the association between test scores and class size is not deterministic, depending on the value of other factors.

An interpretation of the population regression model

Now we have set up the simple linear regression model,

$$TestScore = \beta_0 + \beta_1 ClassSize + OtherFactors$$

What is β_1 and β_0 represent in the model?

- Interpret β_1

  Let's first look at β_1. When we hold other factors constant, the only reason for a change in test scores is a change in class size. Denote $\Delta TestScore$ and $\Delta ClassSize$ to be their respective change. According to the above regression model, holding other factors constant, we have

  $$\Delta TestScore = \beta_1 \Delta ClassSize$$

  where β_0 is removed because it is also a constant. Then, we get

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

  That is, β_1 measures the change in the test score resulting from a **one-unit change** in the class size. When $TestScore$ and $ClassSize$ are two continuous variable, we can write β_1 as

  $$\beta_1 = \frac{\mathrm{d}TestScore}{\mathrm{d}ClassSize}$$

  Hence, we often call β_1 as the **marginal effect** of the class size on the test score.

  The phrase of "holding other factors constant" is important. Without it, we cannot disentangle the effect of class sizes on test scores from other factors. "Holding other things constant" is often expressed as the notion of **ceteris paribus**.

- Interpret β_0

  β_0 is the intercept in the model. Sometimes it bears real meanings, but sometimes it merely presents as an intercept. In this regression model, β_0 is the test score when the class size and other factors are all zero, which is obviously nonsensical. Thus, β_0 does not have a real meaning in this model, and it just determines where the population regression line intersects the Y axis.

2.3 The general linear regression model

Let's generalize test scores and class sizes to be two random variables Y and X. For both, there are n observations so that each observation $i = 1, 2, 3, \ldots$ is associated with a pair of values of (X_i, Y_i).

Then a **simple linear regression model** that associates Y with X is

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \text{ for } i = 1, \ldots, n \tag{4}$$

- Y_i is called the dependent variable, the regressand, or the LHS (left-hand side) variable.
- X_i is called the independent variable, the regressor, or the RHS (right-hand side) variable.
- β_0 is 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, i.e., the point of intersection with the Y axis.
- β_1 is the slope of the population regression line. Since $\beta_1 = \mathrm{d}Y_i/\mathrm{d}X_i$, it is the marginal effect of X on Y. That is, holding other things constant, one unit change in X will make Y change by β_1 units.
- 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.
- $\beta_0 + \beta_1 X_i$ represents the population regression function(or the population regression line).

2.4 An graphical illustration of a linear regression model

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

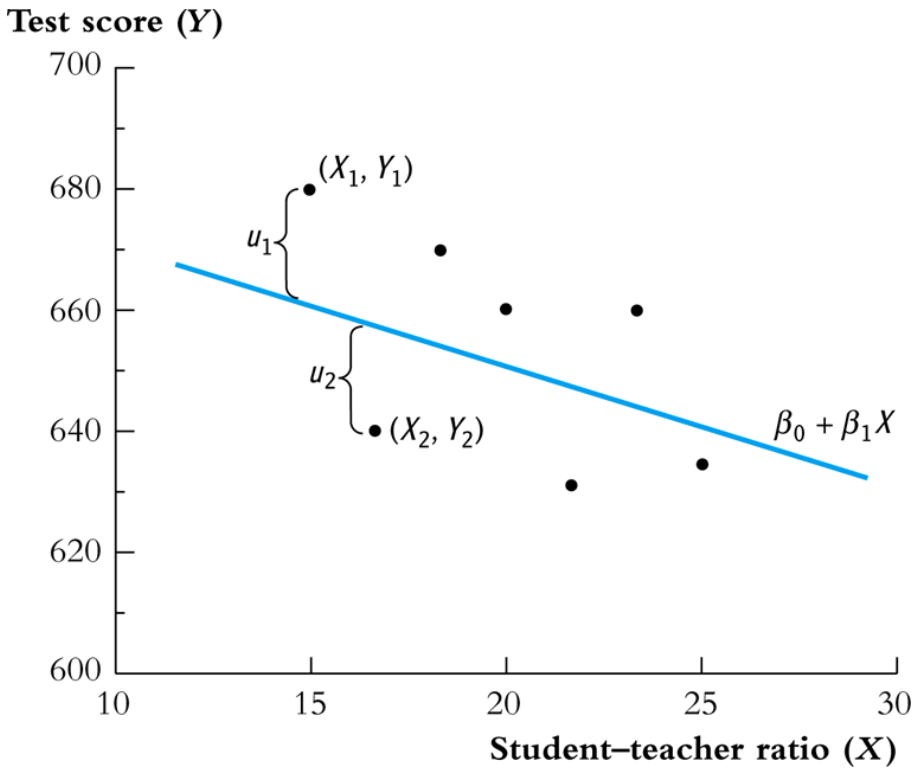

Figure 1: The Population Regression Line