Lecture 6: Linear Regression with One Regressor

Zheng Tian

Outline

1. The Linear Regression Model

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 statistics, 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,

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

The above equation is a simple linear regression function.

Research question:

The research question of this application is: Can reducing class size increase students' test scores?

- How can we answer this question?

Randomized controlled experiment

- Randomly choose 42 students and divide them into two classes, with one having 20 students and another having 22.
- 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.

Compute conditional means

- After a test for both classes, we then compute the expected values of test scores, given the different class sizes.

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

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

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

The population regression function for test scores on class sizes

- We use a linear regression function to describe the relationship between test scores and class sizes.
- The population regression function or the population regression line

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

The simple linear regression model for test scores on class sizes

- 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 becomes the population regression line.

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

- A population regression function
    - It's a deterministic relation between class size and the expectation of test scores.
    - However, we cannot compute the exact value of the test score of a particular observation.
- A population regression model
    - It's a complete description of a data generating process (DGP).
    - 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

- Denote $\Delta TestScore$ and $\Delta ClassSize$ to be their respective change.
- 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.

Marginal effect

- When *TestScore* and *ClassSize* are two continuous variable, we can write β_1 as

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

- We often call β_1 as the marginal effect of the class size on the test score.

Holding other things constant

- 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 regression model of test scores on class sizes, β_0 is the test score when the class size and other factors are all zero, which is obviously nonsensical.

The general linear regression model

- Consider 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

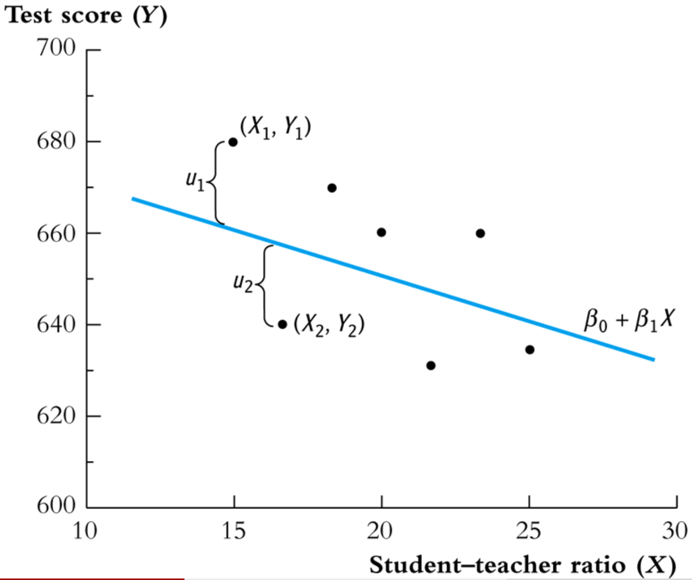
$$Y_i = \beta_0 + \beta_1 X_i + u_i, \text{ for } i = 1, \ldots, n \tag{4}$$

- 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.

The general linear regression model (cont'd)

- 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).

An graphical illustration of a linear regression model


Test score (Y)
Student–teacher ratio (X)
(X_1, Y_1)
u_1
u_2
(X_2, Y_2)
$\beta_0 + \beta_1 X$