Jared Yu

Module 1 Discussion

1. A typical simple linear regression model is , where is a response variable (also often called dependent variable), is an independent variable (also often called regressor), and is a random error with mean (also called expectation) zero. Thus and are random variables. The regressor is either a random variable or a non-random (also often called fix) variable.
   1. The regressor is non-random. What is the meaning of the expectation of , denoted ? What is the meaning of the expectation of given (or conditional on) , denoted by ? What are the differences between the two expectations?

Ans:

In equation (1), the formula for is used to replace it within the expected value equation.

In equation (2), the assumption is that and are the true parameters (rather than estimates of the true parameters) and is a fixed variable (i.e., a non-stochastic value). Therefore, these values can be seen as constants. So, it is straightforward to pull them out of the expected value equation since when is a constant and is a random variable using basic expectation rules.

In equation (3), the term is being evaluated to be 0. The error term, , is a random variable with mean 0 and so the expected value of it is also 0. The equation itself is the expected value of the simple linear regression model. It says that for some random variable , its expected value is equal to , where is the true intercept, is the true slope, and is the corresponding fixed value that pairs with .

In equation (4), it can be seen that evaluates to the same line as . This equation however is saying more specifically, that the expected value of at some point is . So, for each value of , the value of can be seen as lying on a probability distribution. The mean of this distribution is (p. 12). This probability distribution is depicted well in the Figure 1.2 on p. 3 of the textbook. It is apparent that for some value , there is a “sideways” probability distribution. This distribution is showing how that the corresponding for has a distribution at the point , spread along the vertical -axis.

* 1. The regressor is random. Discuss the questions in a) above.

Ans:

In this situation, and are jointly distributed random variables. It is unknown whether they are jointly normal, so it will not be assumed. Their distribution then is unknown. If the following 2 conditions are true, then what was said previously still applies:

1. “The conditional distribution of given is normal with conditional mean and conditional variance . (p. 53)”

2. “The ’s are independent random variables whose probability distribution does not involve , , and . (p. 53)”

In equation (5), the same step is repeated as in equation (1), where the formula for replaces it inside the expectation.

In equation (6), is being treated as a constant and the other terms are being treated as some sort of random variable, for example . Then, using the same basic expectation rules, the constant term is pulled out.

In equation (7), the summation inside is a summation of two random variables, where one of the random variables has a constant attached (i.e., is the constant being multiplied to the random variable ). They are able to be split apart using the *linearity of expectation* rule, where the expected value of a sum of random variables is equal to the sum of the expected values of the random variables.

In equation (8), the same step is happening as in equation (3), where the error term is being evaluated to 0 after having its expected value taken. Another step also is that the constant term comes out of the expected value. For example, , where is a constant and is a random variable. This formula is saying that the expected value of the random variable is equal to , which is a linear equation of the true parameters and along with the expected value of the random variable .

In equation 9, the value of is used to indicate that it’s being conditioned on in the situation where the realized value of is . Since the distributions are unknown, it can’t be fully evaluated. If, however the assumption is that and are jointly normal (i.e., they follow a bivariate normal distribution), then the following holds:

This is shown on p.54. The more obvious difference between the two equations ( versus ) is that in this scenario, we can’t treat like a constant, fixed term (unlike in part a). Solving for , we see that the expected value for the random variable depends on the expected value of the random variable , which makes sense since the former is dependent on the latter and both are unknown. Solving for , it is seen that it’s not directly solvable unless more information is known, such as the distribution of and . It is saying that, “Given that we have sampled and know its value, what is the expected value of the random variable then?” This is in contrast to just asking about the expected value of the random variable .

A notable difference also is that in part b, we are talking about sampling pairs of random variables, rather than in part a where we are sampling at a fixed level of (p. 53).

1. Under a typical simple linear regression model as given in Problem 1 above, if the value of increases by units, how much does the value of change? Is the change an increase or decrease?

The formula for simple linear regression (SLR) will be restated as follows:

If increases by units, then it would appear as follows:

In equation (12), it can be seen that an increase of leads to the additional term in the equation. In the case where is positive, then the change is an increase, however if it is negative, the change is a decrease.

1. The simple linear regression model given in Problem 1 above represents a straight-line relationship between and . If the values of and are given, a straight line can be drawn. Do all of the values of given values fall exactly on the straight line? If yes, why? If not, why not?

The SLR model will be stated again as follows:

The equation represents an SLR model for a set of data. It is showing that for a set of data, ideally a regression line can be drawn through it to show a linear relationship between and . It is possible that the data could exactly fall on this line, but in the real-world it is not too likely. This is emphasized by the term, where it is a random variable that has mean 0. This term helps to account for the variance of the data points above and below the regression line at any point .