Jared Yu

Module 1 Discussion

1. A typical simple linear regression model is $y = \beta_0 + \beta_1 x + \varepsilon$, where y is a response variable (also often called dependent variable), x is an independent variable (also often called regressor), and ε is a random error with mean (also called expectation) zero. Thus y and ε are random variables. The regressor x is either a random variable or a non-random (also often called fix) variable.
    a. The regressor x is non-random. What is the meaning of the expectation of y, denoted $E(y)$? What is the meaning of the expectation of y given (or conditional on) x, denoted by $E(y|x)$? What are the differences between the two expectations?

Ans:

$$E(y) = E(\beta_0 + \beta_1 x + \varepsilon) \tag{1}$$

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

$$= \beta_0 + \beta_1 x + E(\varepsilon) \tag{2}$$

In equation (2), the assumption is that β_0 and β_1 are the true parameters (rather than estimates of the true parameters) and x 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 $E[a + X] = a + E[X]$ when a is a constant and X is a random variable using basic expectation rules.

$$= \beta_0 + \beta_1 x + 0 = \beta_0 + \beta_1 x \tag{3}$$

In equation (3), the term $E(\varepsilon)$ is being evaluated to be 0. The error term, ε, is a random variable with mean 0 and so the expected value of it is also 0. The equation $E(y)$ itself is the expected value of the simple linear regression model. It says that for some random variable y, its expected value is equal to $\beta_0 + \beta_1 x$, where β_0 is the true intercept, β_1 is the true slope, and x is the corresponding fixed value that pairs with y.

$$E(y|x) = \mu_{y|x} = E(\beta_0 + \beta_1 x + \varepsilon) = \cdots = \beta_0 + \beta_1 x \tag{4}$$

In equation (4), it can be seen that $E(y|x)$ evaluates to the same line as $E(y)$. This equation however is saying more specifically, that the expected value of y at some point x is $\beta_0 + \beta_1 x$. So, for each value of x, the value of y can be seen as lying on a probability distribution. The mean of this distribution is $E(y|x)$ (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 x, there is a "sideways" probability distribution. This distribution is showing how that the corresponding y for x has a distribution at the point x, spread along the vertical y-axis.

    b. The regressor x is random. Discuss the questions in a) above.

Ans:

In this situation, x and y 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 y given x is normal with conditional mean $\beta_0 + \beta_1 x$ and conditional variance σ^2. (p. 53)"
2. "The x's are independent random variables whose probability distribution does not involve β_0, β_1, and σ^2. (p. 53)"

$$E(y) = E(\beta_0 + \beta_1 x + \varepsilon) \tag{5}$$

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

$$= \beta_0 + E(\beta_1 x + \varepsilon) \tag{6}$$

In equation (6), β_0 is being treated as a constant and the other terms are being treated as some sort of random variable, for example $Z = \beta_1 x + \varepsilon$. Then, using the same basic expectation rules, the constant term is pulled out.

$$= \beta_0 + E(\beta_1 x) + E(\varepsilon) \tag{7}$$

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., β_1 is the constant being multiplied to the random variable x). 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.

$$= \beta_0 + \beta_1 E(x) + 0 = \beta_0 + \beta_1 E(x) \tag{8}$$

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, $E(aX) = aE(X)$, where a is a constant and X is a random variable. This formula is saying that the expected value of the random variable y is equal to $\beta_0 + \beta_1 E(x)$, which is a linear equation of the true parameters β_0 and β_1 along with the expected value of the random variable x.

$$E(y|x = x_0) = \int_{-\infty}^{\infty} x_0 f_{y|x=x_0}(y)dy \tag{9}$$

In equation 9, the value of x_0 is used to indicate that it's being conditioned on x in the situation where the realized value of x is x_0. Since the distributions are unknown, it can't be fully evaluated. If, however the assumption is that x and y are jointly normal (i.e., they follow a bivariate normal distribution), then the following holds:

$$E(y|x) = \beta_0 + \beta_1 x. \tag{10}$$

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

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

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

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

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{11}$$

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

$$y = \beta_0 + \beta_1(x + \Delta) + \varepsilon = \beta_0 + \beta_1 x + \beta_1 \Delta + \varepsilon. \tag{12}$$

In equation (12), it can be seen that an increase of Δ leads to the additional $\beta_1\Delta$ term in the equation. In the case where β_1 is positive, then the change is an increase, however if it is negative, the change is a decrease.

3. The simple linear regression model given in Problem 1 above represents a straight-line relationship between y and x. If the values of β_0 and β_1 are given, a straight line can be drawn. Do all of the values of y given x values fall exactly on the straight line? If yes, why? If not, why not?

The SLR model will be stated again as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{13}$$

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 y and x. 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 ε 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 x.