625.661 Statistical Models and Regression

Module 1 Discussion Questions

H.M. James Hung

Please discuss all the following questions.

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 fixed) variable.

   a) The regressor x is non-random. What is the meaning of the expectation of y, denoted by $E(y)$? What is the meaning of the expectation of y given (or conditional on) x, denoted by $E(y \mid x)$? What are the differences between the two expectations?

      E(y) is the expectation of y for the entire population without considering their x values. This is a population parameter, often to be estimated by the sample mean $n^{-1}\sum_{i=1}^{n} y_i$ (or a weighted average) from a random sample, but E(y) is not the sample mean (or that weighted average).
      E(y | x) is the expectation of y of the population that has the specific x value. This is also a population parameter but given a value of x, often could be estimated by the sample mean from a random sample of subjects meeting the condition of x, but E(y | x) is not the sample mean of those y values when x condition is met.
      As an example, let us assume that x is a gender variable and y is the annual income to be earned in 2019, considering the US population. E(y | x = male) is the expected (or average) annual income in 2019 in US males, while E(y) is the expected annual income in 2019 in US males and US females.

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

Same answer as given for Problem 1. In addition, if x is a random variable, then x has a statistical distribution. Hence E(y | x) is also a random variable. Then, E(y) = E(E(y | x)); i.e., it is average of average E(y|x); average means expectation. The first E is with respect to the x distribution and the second E is with respect to the y distribution conditional on x. Using the example in Problem 1 above, after properly grouping, if US population consists of 40% males and 60% females, then E(Y) = 0.4*E(Y | X = males) + 0.6*E(Y | X = females). Note that there are no sample estimates involved.

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?

   A typical simple regression model is E(y | x) = $\beta_0 + \beta_1 x$ or y = $\beta_0 + \beta_1 x + \varepsilon$, where ε is a random error with mean zero. Under this model, if x value changes by Δ , then E(y | x+Δ) - E(y | x) = $\beta_1\Delta$; that is, the value of E(y | x) changes by $\beta_1\Delta$. The sign of the increase or decrease depends on the sign of β_1 . However, how much the value of y changes is unknown, depending on the amount of random error; that is, we only can access the value E(y | x).

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 the values of y given x values fall exactly on the straight line? If yes, why? If not, why not?

   No. We only know the expected value of y vs x will fall on the straight line because y values have the random variations from ε .