ST3006 – Regression Analysis

4. Simple Linear Regression Model

This is a process where we study the relationship between a dependent variable Y and a single independent variable X. The simplest form of relationship between two variables is a straight line and we can initially think about fitting the following functional relationship between Y and X as discussed above.

$$Y = \beta_0 + \beta_1 X \text{ ....} (1)$$

Here, β_0 is the intercept and β_1 is the slope of the fitted line. However, as we stated before, oobservations do not fall directly on the curve of a statistical relationship. Scattering of points will be there around the curve of interest due to random variation of observations because of the effects of uncontrolled and unforeseen factors on the response variable of interest Y. To account for the scattering of points around the curve, the equation (1) should be modified accordingly. Let the difference between the observed value Y and the straight line $(\beta_0 + \beta_1 X)$ be an error ε. It is a statistical error and a random variable that accounts for the failure of the model to fit the observed data exactly. The error is made up by the effects of other influential factors on the response variable Y.

Thus, the relationship between Y and X should be defined in the form of a simple linear regression model as

$$Y = \beta_0 + \beta_1 X + \varepsilon \text{..............} (2)$$

Here, β_0 and β_1 are called the regression coefficients or the parameters of the model, and ε is a random disturbance or error term.

Now note that a regression model is a formal means of expressing the two essential ingredients of a statistical relationship:

1. A tendency of the dependent variable Y to vary with the independent variable or variables in a systematic fashion.
2. A scattering of points around the curve of a statistical relationship.

These two characteristics are included in a regression model assuming that:

1. There is a probability distribution of Y for each level of X.
2. The means of these probability distributions vary in some systematic fashion with X and it is called the regression function of Y on X (refer figure 01)

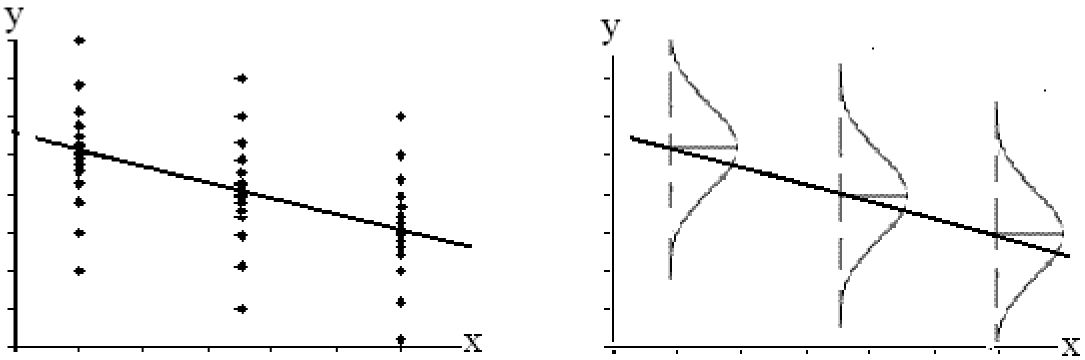

Figure 01: Regression function of Y on X.

It is assumed that this linear equation provides an acceptable approximation to the true relationship between Y and X within the considered range of observations and ε measures the discrepancy in that approximation. In particular, ε is assumed to contain no systematic information for determining Y that has not already captured in X. Note that the equation (2) can be defined as the population regression function (PRF). To find the PRF data should be collected on variables Y and X from all the units of the population of interest. It is important to note that PRF is not directly observable in general. Therefore, we estimate the sample regression function (SRF) based on a randomly observed sample of size n from the corresponding population.

Each observation in the sample can be written in the form; $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; $i = 1, 2, \dots, n$.

Here, y_i – i$^{\text{th}}$ observation of the response variable Y in the sample

x_i – i$^{\text{th}}$ observation of the explanatory variable X in the sample

ε_i – error term associated with the i$^{\text{th}}$ observation y_i.

Next step is to estimate the unknown parameters; β_0 and β_1, in the regression model. This process is called fitting the model to the data.

Parameter estimation

As stated before, a relationship between two variables can be presented in a scatter plot. Bye eye inspection, there are many ways of drawing a line that one may think would fit the data. One may join the two extreme points or would use two mid points to draw the straight line. Thus, there could be many straight lines that describe the same data set. However, there is only ***one best line.*** We use least squares method to find the best fitted line. Here, the best line is obtained by minimizing the sum of the squares of the errors (SSE) of n observations. This ***best line is called the least squares regression line.***

For convenience, let us denote SSE by S. Now the requirement is to minimize

$$S = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \;.$$

Thus, we use; $\dfrac{\partial S}{\partial \beta_0} = 0$ ………….. (3) and $\dfrac{\partial S}{\partial \beta_1} = 0$ …………… (4)

Equation (3) gives $$2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0 \text{ …………… (5)}$$

Equation (4) gives $$2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0 \text{ …………… (6)}$$

The above equations are called ***normal equations*** and the least squares estimates of β_0 and β_1 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{XY}}{S_{XX}} \text{ and } \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \text{, respectively.}$$

Hence, the least square regression line is; $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

For each observation we can find $\hat{y}_i$ using this equation, and $\hat{y}_i$ are called the fitted values. The estimated errors are called the ordinary least squares residuals and they are estimated as $e_i = y_i - \hat{y}_i$ for each $i = 1, 2, 3, \dots, n$. Figure 02 illustrates the least square criterion.

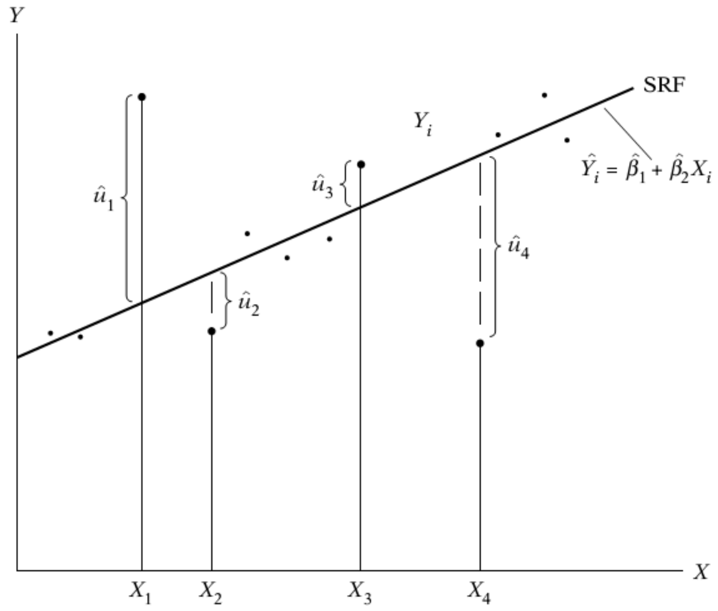

Figure 02: Least squares criterion

Deriving the OLS (ordinary least squares) estimators of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$

Step 1: Deriving an expression for $\boldsymbol{\widehat{\beta}_0}$

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$ divide equation (5) by -2

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{\beta}_0 - \sum_{i=1}^{n}\hat{\beta}_1 x_i = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \sum_{i=1}^{n}\hat{\beta}_1 x_i = 0$$ divide both sides by n

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$$

Obtain an expression for $\hat{\beta}_0$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Step 2: Deriving an expression for $\boldsymbol{\widehat{\beta}_1}$

Re-arrange equation (5)

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \qquad \text{……………} (7)$$

Solve and re-arrange equation (6)

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \qquad \text{……………} (8)$$

Now multiply equation (7) by the sum of x_i and equation (8) by n. Subsequently,

$$\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = n \overset{\wedge}{\beta}_0 \sum_{i=1}^{n} x_i + \overset{\wedge}{\beta}_1 \left[\sum_{i=1}^{n} x_i\right]^2 \quad \text{…………… (9)}$$

$$n \sum_{i=1}^{n} x_i y_i = n \overset{\wedge}{\beta}_0 \sum_{i=1}^{n} x_i + n \overset{\wedge}{\beta}_1 \sum_{i=1}^{n} x_i^2 \quad \text{…………… (10)}$$

Subtract equation (9) from equation (10).

$$n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = n \overset{\wedge}{\beta}_1 \sum_{i=1}^{n} x_i^2 - \overset{\wedge}{\beta}_1 \left[\sum_{i=1}^{n} x_i\right]^2 \quad \text{…………… (11)}$$

Solving equation (11) yields the OLS estimate of $\hat{\beta}_1$

$$\overset{\wedge}{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2} \quad \text{…………… (12)}$$

Recall from above that $S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Exercise: Prove $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{S_{XY}}{S_{XX}}$ using equation (12).

So far in our estimation, we have assumed that X and Y are linearly related. We should validate this assumption by using a scatter plot and the line superimposed on it. Apart from the assumption of linearity there are other required assumptions that must be satisfied before meaningful conclusions are drawn from the fitted line.

Standard regression assumptions

- The response Y and the explanatory variable(s) are linearly related.
- ε_i is a normally distributed random variable with mean zero $(i.e.\ E[\varepsilon_i] = 0)$ and constant variance σ^2 $(i.e.\ Var[\varepsilon_i] = \sigma^2)$ for all $i = 1, 2, 3, \dots, n$.
- ε_i and ε_j are uncorrelated so that the covariance between ε_i and ε_j is zero $\left(i.e.\ Cov\left[\varepsilon_i\,,\ \varepsilon_j\ \right] = 0\right)$ for all $ij; i \neq j$.