Linear Regression

Regression analysis is a mathematical measure of average relationship between two or more variables in terms of the original units of data. The curve around which the scatter diagram of the two variables cluster around is called the curve of regression. If this curve looks like a straight line, it is known as line of regression.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. This line of best fit is obtained by the principle of least squares. Let $(x_i, y_i); i = 1, 2, \cdots, n$ be the given data, where Y is the dependent variable and X is the independent variable. Let the line of regression of Y on X be

$$Y = a + bX. \tag{1}$$

Here, (1) represents a family of straight lines for different values of the arbitrary constants a and b. The problem is to determine the values of a and b so that the line (1) gives the best fit to the given data. For this, we use the principle of least squares. Let (x_i, y_i) be an observation of the given data. As per our assumption, we have $y_i = a + bx_i$. Hence, the error of the estimate or the residue for y_i is given by

$$\epsilon_i = y_i - a - bx_i.$$

Therefore, the total error sum of squares is given by

$$E = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{2}$$

According to the principle of least squares, we have to determine a and b so that the total error sum of squares E is minimum. From the principle of maxima and minima, the partial derivatives of E with respect to a and b are given should vanish separately, i.e.,

$$\frac{\partial E}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0,$$

which gives

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i. \tag{3}$$

$$\frac{\partial E}{\partial b} = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0,$$

which gives

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2. \tag{4}$$

(5) and (6) are known as normal equations for estimating a and b. All the quantities in (5) and (6) can be obtained form the given set of data except a and b and hence these equations can be solved for a and b. With these values of a and b so obtained, (1) is the line of best fit for the given data set $(x_i, y_i); i = 1, 2, \cdots, n$.

Now, dividing (3) by n we get

$$\bar{y} = a + b\bar{x} \tag{5}$$

Thus, the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$. Therefore, (4) implies

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = b\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

which implies

$$b = b_{yx} = \frac{cov(X,Y)}{var(X)}. \tag{6}$$

Here, b_{yx} is the coefficient of regression of Y on X and it is also the slope of the line of regression Y on X. Since the line of regression passes through $(\bar{x}, \bar{y})$, its equation can also be written as

$$Y - \bar{y} = b_{yx}(X - \bar{x}) = \frac{cov(X,Y)}{var(X)}(X - \bar{x}),$$

which gives the line of regression Y on X as

$$Y - \bar{y} = \rho\frac{\sigma_y}{\sigma_x}(X - \bar{x}). \tag{7}$$

where ρ is the Karl Pearson's coefficient of correlation

Exercise 1. *1. Derive the expression for line of regression X on Y.*

2. Obtain the equations of two lines of regression for the following data. Also obtain the estimate of X for $Y = 70$.

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71