Curve fitting Techniques

Method of Least Squares

thod of least squares can be used to determine the line of best fit in such cases. It determines the line of best fit for given observed data by minimizing the sum of the uares of the vertical deviations from each data point to the line.

METHOD OF LEAST SQUARES

In most of the cases, the data points do not fall on a straight line (not highly correlated), thus leading to a possibility of depicting the relationship between the two variables using several different lines. Selection of each line may lead to a situation where the line will be closer to some points and farther from other points. We cannot decide which line can provide best fit to the data.

Method of least squares can be used to determine the line of best fit in such cases. It determines the line of best fit for given observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line.

1. Method of Least Squares

To obtain the estimates of the coefficients 'a' and 'b', the least squares method minimizes the sum of squares of residuals. The residual for the i^{th} data point e_i is defined as the difference between the observed value of the response variable, y_i, and the estimate of the response variable, $\hat{y}_i$, and is identified as the error associated with the data. *i.e.*, $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$.

The method of least squares helps us to find the values of unknowns 'a' and 'b' in such a way that the following two conditions are satisfied:

Sum of the residuals is zero. That is $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$

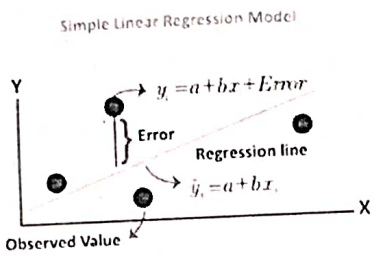
Sum of the squares of the residuals $E(a, b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the least

2. Fitting of Simple Linear Regression Equation

The method of least squares can be applied to determine the estimates of 'a' and 'b' in the simple linear regression equation using the given data (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) by minimizing

$$E(a,b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

i.e., $$E(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$



Here, $\hat{y}i = a + bx_i$ is the expected (estimated) value of the response variable for given x_i.

It is obvious that if the expected value ($\hat{y}_i$) is close to the observed value (y_i), the residual will be small. Since the magnitude of the residual is determined by the values of 'a' and 'b', estimates of these coefficients are obtained by minimizing the sum of the squared residuals, $E(a,b)$.

Differentiation of $E(a,b)$ with respect to 'a' and 'b' and equating them to zero constitute a set of two equations as described below:

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

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

These give

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

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

These equations are popularly known as **normal equations.** Solving these equations for 'a' and 'b' yield the estimates $\hat{a}$ and $\hat{b}$.

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

and

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

It may be seen that in the estimate of 'b', the numerator and denominator are respectively the sample covariance between X and Y, and the sample variance of X. Hence, the estimate of 'b' may be expressed as

$$\hat{b} = \frac{Cov(X,Y)}{V(X)}$$

Further, it may be noted that for notational convenience the denominator of $\hat{b}$ above is mentioned as variance of X. But, the definition of sample variance remains valid as defined in Chapter I, that is,

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}^2\right)$$

From Chapter 4, the above estimate can be expressed using, r_{XY}, Pearson's coefficient of the simple correlation between X and Y, as

$$\hat{b} = r_{XY}\frac{SD(Y)}{SD(X)}.$$

Important Considerations in the Use of Regression Equation:

1. Regression equation exhibits only the relationship between the respective two variables. Cause and effect study shall not be carried out using regression analysis.

2. The regression equation is fitted to the given values of the independent variable. Hence, the fitted equation can be used for prediction purpose corresponding to the values of the regressor within its range. Interpolation of values of the response variable may be done corresponding to the values of the regressor from its range only. The results obtained from extrapolation work could not be interpreted.

Example 5.1

Construct the simple linear regression equation of Y on X if

$$n = 7, \sum_{i=1}^{n} x_i = 113, \sum_{i=1}^{n} x_i^2 = 1983,$$

$$\sum_{i=1}^{n} y_i = 182 \text{ and } \sum_{i=1}^{n} x_i y_i = 3186.$$

Solution:

The simple linear regression equation of Y on X to be fitted for given data is of the form

$$\hat{Y} = a + bx \qquad \ldots\ldots(1)$$

The values of 'a' and 'b' have to be estimated from the sample data solving the following normal equations.

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (2)$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad (3)$$

Substituting the given sample information in (2) and (3), the above equations can be expressed as

$7a + 113b = 182$ (4)

$113a + 1983b = 3186$ (5)

(4) $\times 113 \Rightarrow 791a + 12769b = 20566$

(5) $\times 7 \Rightarrow 791a + 13881b = 22302$

$$\begin{array}{rl} & (4) \\ 7a + 113b = 182 & \\ 113a + 1983b = 3186 & (5) \\ (4) \times 113 \Rightarrow 791a + 12769b = 20566 & \\ (5) \quad \times 7 \Rightarrow 791a + 13881b = 22302 & \\ \hline (-) \quad (-) \quad (-) & \\ -1112b = -1736 & \end{array}$$

$$\Rightarrow b = \frac{1736}{1112} = 1.56$$

$$b = 1.56$$

Substituting this in (4) it follows that,

$7a + 113 \times 1.56 = 182$

$7a + 176.28 = 182$

$7a = 182 - 176.28$

$= 5.72$

Hence, $a = 0.82$

Example 5.2

Number of man-hours and the corresponding productivity (in units) are furnished below. Fit a simple linear regression equation $\hat{Y} = a + bx$ applying the method of least squares.

Man-hours	3.6	4.8	7.2	6.9	10.7	6.1	7.9	9.5	5.4
Productivity (in units)	9.3	10.2	11.5	12	18.6	13.2	10.8	22.7	12.7

Solution:

The simple linear regression equation to be fitted for the given data is

$\hat{Y} = a + bx$

Here, the estimates of a and b can be calculated using their least squares estimates

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{b}\frac{1}{n}\sum_{i=1}^{n} x_i$$

i.e.,

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

or equivalently $\hat{b} = \dfrac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$

From the given data, the following calculations are made with n=9

Man-hours x_i	Productivity y_i	x_i^2	$x_i y_i$
3.6	9.3	12.96	33.48
4.8	10.2	23.04	48.96
7.2	11.5	51.84	82.8
6.9	12	47.61	82.8
10.7	18.6	114.49	199.02
6.1	13.2	37.21	80.52
7.9	10.8	62.41	85.32
9.5	22.7	90.25	215.65
5.4	12.7	29.16	66.42
$\sum_{i=1}^{9} x_i = 62.1$	$\sum_{i=1}^{9} y_i = 121$	$\sum_{i=1}^{9} x_i^2 = 468.97$	$\sum_{i=1}^{9} x_i y_i = 894.97$

Substituting the column totals in the respective places in the of the estimates $a^{\wedge}$ and $b^{\wedge}$, their values can be calculated as follows:

$$\hat{b} = \frac{(9 \times 894.97) - (62.1 \times 121)}{(9 \times 468.97) - (62.1)^2}$$

$$= \frac{8054.73 - 7514}{4220.73 - 3856.41}$$

$$= \frac{540.73}{364.32}$$

Thus, $b^{\wedge} = 1.48$.

Now $a^{\wedge}$ can be calculated using $b^{\wedge}$ as

$a^{\wedge} = 121/9 - (1.48 \times 62.1/9)$

$= 13.40 - 10.21$

Hence, $a^{\wedge} = 3.19$

Therefore, the required simple linear regression equation fitted to the given data is

$\hat{y} = 3.19 + 1.48x$

It should be noted that the value of Y can be estimated using the above fitted equation for the values of x in its range *i.e.*, 3.6 to 10.7.

In the estimated simple linear regression equation of Y on X

$\hat{y} = a^{\wedge} + {}^{\wedge}bx$

we can substitute the estimate $a^{\wedge} = \bar{y} - b^{\wedge}\bar{x}$. Then, the regression equation will become as

$$\hat{Y} = \bar{y} - \hat{b}\bar{x} + \hat{b}x$$

$$\hat{y} - \bar{y} = \hat{b}(x - \bar{x})$$

It shows that the simple linear regression equation of Y on X has the slope $\hat{b}$ and the corresponding straight line passes through the point of averages ($\bar{x}$, $\bar{y}$). The above representation of straight line is popularly known in the field of Coordinate Geometry as 'Slope-Point form'. The above form can be applied in fitting the regression equation for given regression coefficient $\hat{b}$ and the averages $\bar{x}$ and $\bar{y}$.

As mentioned in Section 5.3, there may be two simple linear regression equations for each X and Y. Since the regression coefficients of these regression equations are different, it is essential to distinguish the coefficients with different symbols. The regression coefficient of the simple linear regression equation of Y on X may be denoted as b_{YX} and the regression coefficient of the simple linear regression equation of X on Y may be denoted as b_{XY}.

Using the same argument for fitting the regression equation of Y on X, we have the simple linear regression equation of X on Y with best fit as

$$\hat{X} = \hat{c} + b_{XY}\, y$$

where $\hat{c} = \bar{x} - b_{XY}\,\bar{y}$

$$b_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2}$$

The slope-point form of this equation is

$$\hat{X} - \bar{x} = b_{XY}(y - \bar{y}).$$

Also, the relationship between the Karl Pearson's coefficient of correlation and the regression coefficient are

$$b_{XX} = r_{XY}\frac{SD(X)}{SD(Y)} \text{ and } b_{YX} = r_{XY}\frac{SD(Y)}{SD(X)}.$$