Linear Regression :

— Dr. Arti Singh
USAR, GGSIPU

Regression analysis is statistical method used to estimate relationship between two or more random variables.

In regression models, there is one dependent random variable (response variable), and one or more independent variables (predictors).

r.v.s | $Y \rightarrow$ Response variable
$X^{(1)}, X^{(2)}, \ldots, X^{(k)} \rightarrow$ Predictors.

Linear regression model

$$Y = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \ldots + \beta_K X^{(k)} + \underbrace{\varepsilon}_{\text{Error}}$$

Here ε is a random error variable with mean zero and unknown variance σ^2.

Consider the case when only one predictor is there —

$$\boxed{Y = \beta_0 + \beta_1 X + \varepsilon}$$

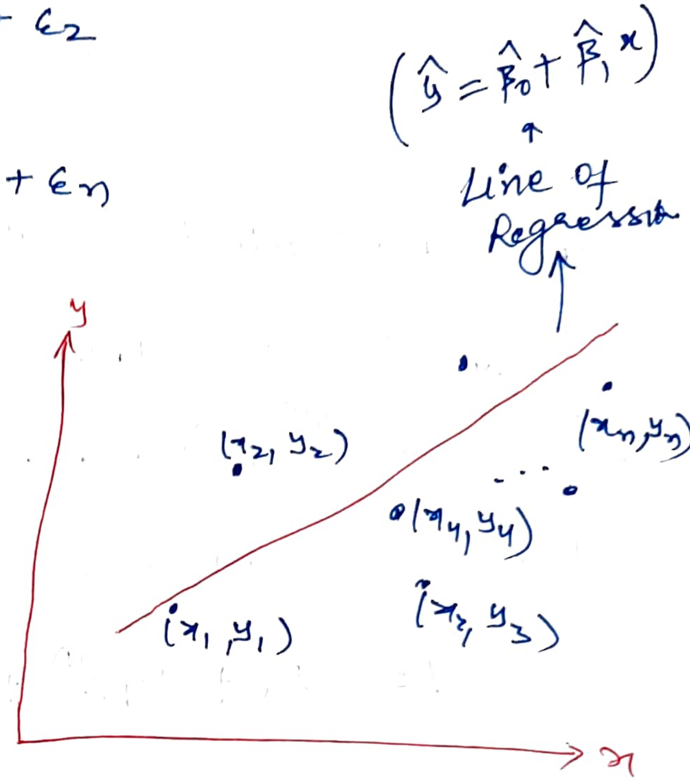
This is called Univariate Linear Regression model.

Suppose that we have n pairs of observations

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Then we have –

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_2 + \epsilon_2$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n$$



Consider following sum of squares of error terms –

$$L = \sum_{i=1}^{n} \epsilon_i^2$$ → [It is also sum of squares of the deviations of observations from the true regression Line.]

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Least square linear regression models:

To find the estimates of slope ($\hat{\beta}_1$) and intercept ($\hat{\beta}_0$) is the fitted line of regre-ssion ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$), we minimize functi-on L with respect to β_0 & β_1.

From theory of calculus, we get

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\beta_0 = \hat{\beta}_0} = 0 \Rightarrow -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\beta_1 = \hat{\beta}_1} = 0 \Rightarrow -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Simplifying above two equations:

$$\sum_{i=1}^{n} y_i - \hat{\beta}_0 \sum_{i=1}^{n} 1 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \quad \text{—— (1)}$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \quad \text{—— (2)}$$

From (1)

$$\sum y_i - \hat{\beta}_0 n - \hat{\beta}_1 \sum x_i = 0$$

$$\Rightarrow \hat{\beta}_0 = \frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum x_i}{n}$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}} \quad \text{—— (A)}$$

From (2)

$$\Rightarrow \sum x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\sum x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) n \bar{x} - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

Also we can write

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \quad \text{— (3)}$$

So fitted line of regression is —

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x .$$

where

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{where} \quad \bar{y} = \frac{\sum y_i}{n}$$

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\& \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Eg Do eg 11.1 of ~~Reference~~ Text book.

Sample Correlation Coefficient (R):

Consider observations (x_i, y_i), $i = 1, 2, \ldots, n$ corresponding to the r.v.s X & Y.

The sample correlation coefficient (R) is given as follows —

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

We can see that

$$\hat{\beta}_1 = \sqrt{\frac{\Sigma (y_i - \bar{y})^2}{\Sigma (x_i - \bar{x})^2}} \times R$$

$$\hat{\beta}_1 = \sqrt{\frac{\frac{1}{n-1} \Sigma (y_i - \bar{y})^2}{\frac{1}{n-1} \Sigma (x_i - \bar{x})^2}} \, R$$

$$\hat{\beta}_1 = \frac{S_y}{S_x} R$$

where S_y & S_x are sample standard deviations corresponding to r.v.s Y & X, respectively.

Multiple Linear Regression Model –

6

A regression model that contains more than one regressor variable is called a multiple linear regression model.

For eg:–

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon}$$

Logistic Regression –

A regression model in which response variable (Y) takes on only two possible values, 0 and 1, is called as Logistic regression model.