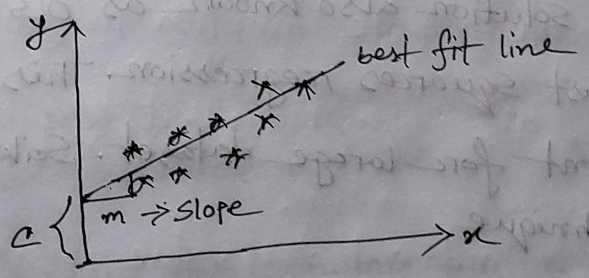
Linear Regression

Simple linear Regression:- In simple linear regression we work one feature and one target column.

When to apply:- Linear Regression basically used when given data is sort of linear. And our main goal is to find out a relationship between features and target column. OR we can say we have to find out a best fit line so that it pass trough all the data point. See the below diagram:-



⇒ We know the straight line formula that is:-

$$y = mx + c$$

Here, m = slope
c = intercept.

And our goal is to find out the value of 'm' and 'c'.

⇒ $y = mx + c$

Here, 'm' also known as weightage. Here, 'm' denote the significance of 'x' that means if we find out 'y' then it how much depend on 'x' features.

⊞ There are two way to find out 'm' and 'c':

① closed form solution [Direct formula]

② Non-colsed from Solution [Using Gradient Descent]

⇒ closed from solution also known as OLS or ordinary Least squares regression. This technique is not efficient for large dataset. Scikit-learn use this technique.

⇒ Non-closed form solution used a optimization algorithm that is gradient descent. It is useful for large dataset.

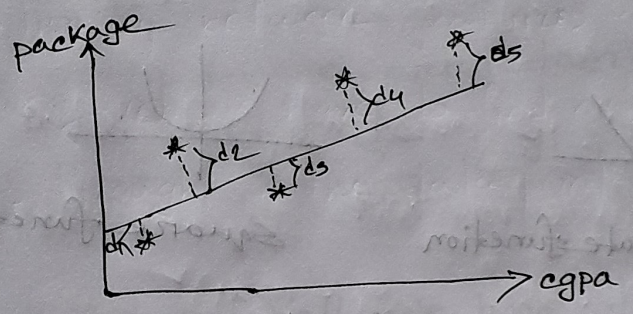
Formula :-

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

Here,
$\bar{x}$, $\bar{y}$ > mean value.

② $$c = \bar{y} - m\bar{x}$$

Let's explore how this formula came :-



Note :- Errore is nothing but the difference between actual value and predicted value.

If we want to findout total number of errore then we can write above picture:

$$E = d_1 + d_2 + d_3 + \dots + d_n$$ [where, n is total rows]

⇒ Errore above the line would be positive and below errore would be Negative. So we can take the square of the all errore:

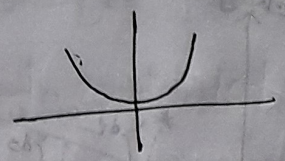
So, $$E = d_1^2 + d_2^2 + d_3^2 + \dots + d_n^2$$

⇒ We can also take absolute value like $|d_i|$ but here is a problem that is :

(i) We not use this so that we can panalize the outlier.

(ii) We can not deffierentiation the absolute value function at any point. But we can do it with square function.

Absolute function

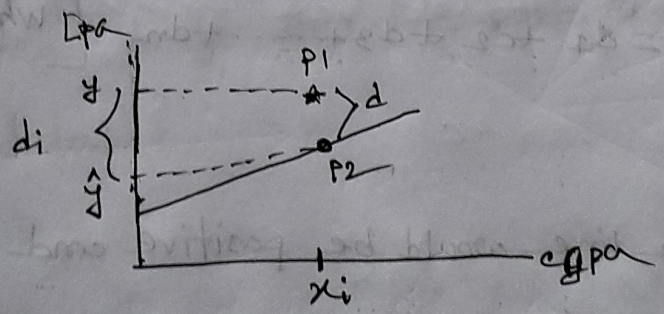


Square function

Now, we can write error function in generalize way;

$$E = \sum_{i=1}^{n} d_i^2 \rightarrow \text{Also called loss function.}$$

Now, we have to find out the value of 'm' and 'c' so that our error function will be minimum.



For a particular point the P_1 is actual point but our model saying the point should be P_2 position. Now, can see for P_1 and P_2 point value of 'x' is same.

The change is happening in 'y' axis. So, if the distance between 'y' and '$\hat{y}$' is 'd_i' then we can write:

For every points, $d_i = y_i - \hat{y}_i$ [where, y_i = actual, $\hat{y}_i$ = predicted]

Now, I can re-write the equation of total error:

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

If it is mean square error then:

$$E = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

⇒ We can say in nutshell we have to find out a line or 'm' and 'c' value so that our error function will be minimum.

⇒ But as we can see in our error function there no 'm' and 'b' term present. It has in hidden form.

We know, $\hat{y}_i = mx_i + c$

So, we can put it in our error function.

Now, $$E(m, c) = \sum_{i=1}^{n} (y_i - mx_i - c)$$

Here, $E(m, c)$ denote the error function depends on two value that are 'm' and 'c'. If we change 'm' and 'c' value then we can reduce the error or loss.

Let's suppose, $y = f(x)$. Now what is the meaning of this function. It means 'y' is a function of 'x'. If we change the value of 'x' then 'y' will be change. Now we will apply this technique on our equation. We will see if we change the value of 'm' and 'c' individually then what impect occurs on the error function.

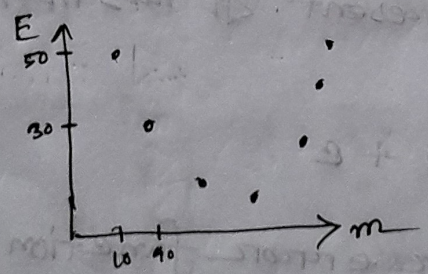
Let's say 'c' is constant and let, $c = 0$.

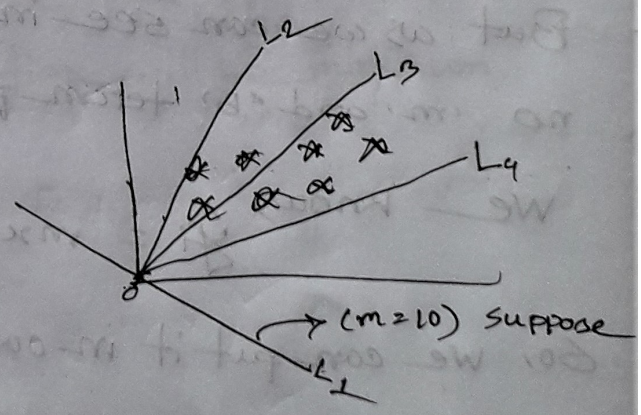
Now, error function will,

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

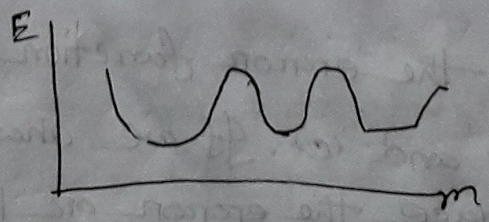
⇒ Now, If we change the 'm' value what impect occur in error function.

let's assume,





This curve could be anything like:



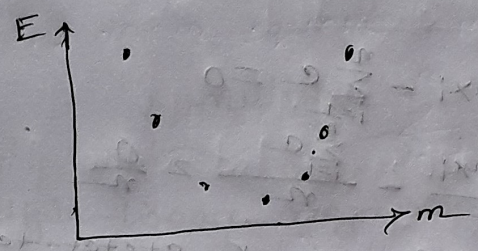
Now, Let's say 'm' is constant and let, $m = 1$.

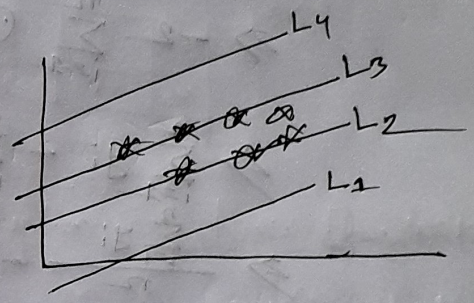
Now, error function will be,

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

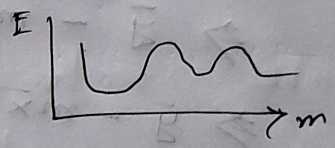
⇒ Now, If we chang the 'c' value what impect occurs in error function.

Let's assume,





This curve also can anything:



⇒ We clearly see for certain value of 'm' and 'c' the error ~~with~~ is maximum and for certain value error is minimum. And our goal is to find out those minimum value.

To minimize the loss or error function we will use partial derivatives, individually 'm' and 'c' and it will be equal to zero.

So, $\frac{\partial E}{\partial m} = 0$, $\frac{\partial E}{\partial c} = 0$.

$\left\langle \frac{\partial E}{\partial c} \rightarrow \text{derivative of } E \text{ with respect to } c \right\rangle$

Now,

$$\frac{\partial E}{\partial c} = \frac{\partial}{\partial c} \sum_{i=1}^{n} (y_i - mx_i - c)^2 = 0$$

$\frac{d}{dx}(x^n) = nx^{n-1}$

$$\Rightarrow \sum_{i=1}^{n} \frac{\partial}{\partial c} (y_i - mx_i - c)^2 = 0$$

$$\Rightarrow 2 \sum_{i=1}^{n} (y_i - mx_i - c) \cdot \frac{\partial}{\partial c} (y_i - mx_i - c) = 0$$

$$\Rightarrow 2 \sum_{i=1}^{n} (y_i - mx_i - c) \cdot (-1) = 0$$

$$\Rightarrow -2 \sum_{i=1}^{n} (y_i - mx_i - c) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} mx_i - \sum_{i=1}^{n} c = 0$$

$$\Rightarrow \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} mx_i}{n} - \frac{\sum_{i=1}^{n} c}{n} = \frac{0}{n}$$

$$\Rightarrow \bar{y} - m\bar{x} - \frac{nc}{n} = 0 \quad \left[\because \frac{c+c+c+\dots+c}{n} = nc\right]$$

$$\Rightarrow \bar{y} - m\bar{x} - c = 0$$

$$\Rightarrow c = \bar{y} - m\bar{x}$$

Again,

$$E(m) = \sum_{i=1}^{n} (y_i - mx_i - c)^2$$

$$\Rightarrow E(m) = \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})^2 \quad [\text{put 'c' value}]$$

$$\Rightarrow \frac{\partial E}{\partial m} \Rightarrow \frac{\partial}{\partial m} \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{\partial}{\partial m} (y_i - mx_i - \bar{y} + m\bar{x})^2 = 0$$

$$\Rightarrow 2 \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x}) \cdot (-x_i + \bar{x}) = 0$$

$$\Rightarrow -2 \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})(x_i - \bar{x}) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})(x_i - \bar{x}) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \left[(y_i - \bar{y}) - m(x_i - \bar{x}) \right] . (x_i - \bar{x}) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - m \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$$

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

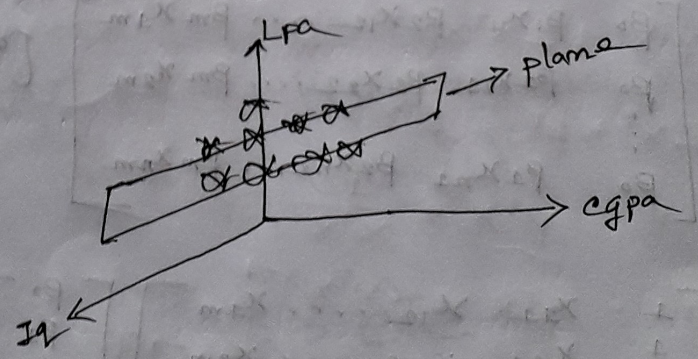
Multiple Linear Regression

⇒ In multiple linear regression the number of features more than one.

Ex:

cgpa	IQ	Lpa
–	–	–
–	–	–

↓



4D – Hyperplane

Now, for multiple linear regression Straight line equation will be

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

or, $\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$

Here, $\beta_1, \beta_2 \ldots \beta_n \rightarrow$ slope
$\beta_0 \rightarrow$ Intercept.

Let's suppose we ~~hon~~ have 'n' number of rows and 'm' number of columns. So, the dataframe shape will be (n, m).

⇒ In multiple linear regression our goal is to find out the value of $\beta_0, \beta_1, \beta_2 \ldots \beta_m$

Ex:-

	gender	cgpa	iq	lpa
	x_0	x_1	x_2	y (actual)

Now, I want to calculate all predicted value for all rows.

Let make a vector that will contain all the predected '$\hat{y}_i$' and make another matrix that will contain all slope and input values:

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \beta_0 & \beta_1 x_{11} & \beta_2 x_{12} & \ldots & \beta_m x_{1m} \\ \beta_0 & \beta_1 x_{21} & \beta_2 x_{22} & \ldots & \beta_m x_{2m} \\ \vdots & & & & \\ \beta_0 & \beta_1 x_{n1} & \beta_2 x_{n2} & \ldots & \beta_m x_{nm} \end{bmatrix}$$

[where $X_{row\ col}$]

$$= \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

If we write more general form:

$$\hat{Y} = X\beta$$

Here, $\hat{Y}$ = predection of all the input rows.

X = all the input values as a matrix.

β = Matrix of all coefficient.

$\Rightarrow$ Again, let's creat a matrix that will contain all actual or target value y_i

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Now, we write,

$$e = Y - \hat{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

$$\Rightarrow e = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

We know, for simple linear regression error function:

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

If we write, $E = e^T e$

Let's prove it,

$$E = \left[(y_1 - \hat{y}_1)(y_2 - \hat{y}_2) \dots (y_n - \hat{y}_n) \right] \cdot \begin{bmatrix} (y_1 - \hat{y}_1) \\ (y_2 - \hat{y}_2) \\ (y_n - \hat{y}_n) \end{bmatrix}$$

Now, if we apply matrix multiplication then we will get the above $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ equation.

Formula
$(A+B)^T = A^T + B^T$
$(A-B)^T = A^T - B^T$
$(A^T)^T = A$

Now, our error function,

$$E = e^T e$$

$$= (Y - \hat{Y})^T \cdot (Y - \hat{Y})$$

$$= \left[Y^T - (X\beta)^T\right] (Y - X\beta) \quad [\because \hat{Y} = X\beta]$$

$$= Y^T Y - \underline{Y^T X\beta} - \underline{(X\beta)^T Y} + (X\beta)^T \cdot X\beta$$

$$= Y^T Y - 2Y^T X\beta + X^T \beta^T X\beta$$

Now, $\frac{\partial E}{\partial \beta} \Rightarrow \frac{\partial}{\partial \beta}\left[Y^T Y - 2Y^T X\beta + X^T \beta^T X\beta\right] = 0$

$$\Rightarrow 0 - 2Y^T X + \frac{\partial}{\partial \beta}\left[\beta^T X^T \cdot X\beta\right] = 0$$

based on derivatives of matrix formula

$$\Rightarrow -2Y^T X + 2X^T X\beta^T = 0$$

$$\Rightarrow \cancel{2}X^T X\beta^T = \cancel{2}Y^T X$$

$$\Rightarrow \beta^T = Y^T X * \frac{1}{X^T X}$$

$$\Rightarrow \beta^T = Y^T X * (X^T X)^{-1}$$

$$\Rightarrow (\beta^T)^T = \left[Y^T X * (X^T X)^{-1}\right]^T$$

$$\Rightarrow \beta = \left[(X^T X)^{-1}\right]^T (Y^T X)^T$$

$$\therefore \beta = (X^T X)^{-1} X^T Y$$

Both side shape will be same

$$\Rightarrow \beta = (X^T X)^{-1} X^T Y$$

$$(m+1, 1) = [(m+1, n)(n, m+1)][(m+1, n)][(n, 1)]$$

$$[(m+1), (m+1)][(m+1, n)][(n, 1)]$$

$$(m+1, n)\ (n, 1)$$

$$\Rightarrow (m+1, 1) = (m+1, 1)$$

$X \rightarrow (n, m+1)$
$X^T \rightarrow (m+1, n)$
$Y \rightarrow (n, 1)$

Let's prove, $Y^T X\beta = (X\beta)^T Y$

$$\therefore A^T B = B^T A \quad \text{—— (i)}$$

let,
$Y = A$
$X\beta = B$

Let's assume another equation

$$(A^T B)^T$$
$$= B^T A \quad \text{—— (ii)} \qquad [\because (A^T)^T = A]$$

So, we can write

$$A^T B = (A^T B)^T$$
$$\Rightarrow c = c^T$$

let,
$A^T B = c$

Here, c is $A^T B$ and A is y^T and B is $X\beta$, so we can write,

$$c = A^T B$$
$$\Rightarrow c = Y^T \cdot X\beta$$

Now, we get,

$$c = c^T$$
$$\Rightarrow Y^T X\beta = (Y^T X\beta)^T$$
$$\Rightarrow Y^T \cdot X\beta = (X\beta)^T Y$$