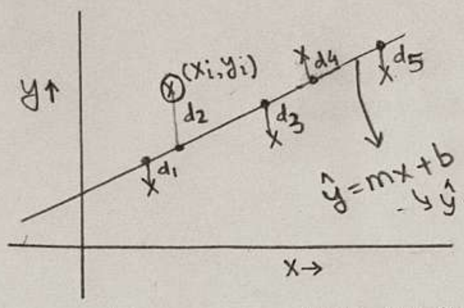
Jeet Banerjee

Mathematical Formulation (Simple Linear Regression)



x → actual data points

• → predicted output (o/p)

$d_1, d_2, \ldots, d_5$ → difference between actual and predicted o/p and this difference is absolute difference

$\hat{y} = mx+b$ - $\hat{y}$ = predicted o/p

m = slope of LR line

b = intercept of LR line

Here x → input, y → output, (x_i, y_i) → for i^{th} i/p, the given o/p is y_i → i^{th} given data point

x (i/p)	y (o/p)
x_1	y_1
x_2	y_2
x_3	y_3
⋮	⋮
x_n	y_n

Simple Linear Regression (SLR)
↓
Single i/p and single o/p

Here, for SLR, the loss function can be represented as:

L = squared/absolute difference between actual o/p and predicted o/p

$\Rightarrow L = d_1^2 + d_2^2 + d_3^2 + \ldots + d_n^2$; where n = no. of data points given/ no. of rows in data

$\Rightarrow L = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + \ldots + (y_n - \hat{y}_n)^2$

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

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

$\Rightarrow \boxed{L(m,b) = \sum_{i=1}^{n} (y_i - mx_i - b)^2}$ — (1) ⇒ Loss function is a function of slope and intercept

To minimize the error b/w actual and predicted o/p, we need to minimize the loss function. Minimizing loss-function w.r.t slope (m) and intercept (b)

Hence, $\boxed{\frac{\partial L}{\partial m} = 0, \frac{\partial L}{\partial b} = 0}$ ⇒ Minimizing — (2)

Now, $\frac{\partial L}{\partial b} = \sum_{i=1}^{n} 2(y_i - mx_i - b)(-1) = 0$

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

$\Rightarrow \sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} 1 = 0$

$\Rightarrow n\bar{y} - mn\bar{x} - bn = 0$ [where, $\bar{x}, \bar{y}$ ⇒ mean of x and mean of y, respectively]

Jeet Banerjee

$\Rightarrow n(\bar{y}-m\bar{x}-b)=0$

$\Rightarrow \bar{y}-m\bar{x}-b=0 \; [\because n\neq 0]$

$\Rightarrow \boxed{b=\bar{y}-m\bar{x}}$ —③

Hence, the loss function $L(m,b)=\sum_{i=1}^{n}(y_i-mx_i-b)^2$

replacing the value of b from eq. (3),

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

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

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

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

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

$\Rightarrow \boxed{m=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$ —④

Hence, using the ordinary least square method, the values of slope (m) and intercept, we got:

$$\boxed{\begin{aligned} m&=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ b&=\bar{y}-m\bar{x}\end{aligned}}$$

— From eq. (3) and eq. (4)

— Only dependent on the given datapoints

—⑤

And, corresponding equation of the simple linear regression line:

$\boxed{\hat{y}=mx+b}$; where (m,b) can be calculated using eq. (5)

└⑥