Algorithms

1) SLR (Simple linear regression)

→ Models the relationship between a single independent Variable X (Predictor) and dependent Variable Y (response).

→ Model Formulation

Expressed as : $(y_i = \beta_0 + \beta_1 x_i + \varepsilon_i)$

- y_i - The observed value for i-th observation
- x_i - value of independent Var
- β_0 - The intercept (the value of y when $x = 0$)
- β_1 - Slope (change in y for unit increase in x).
- ε_i - Error term representing difference b/w (observed) y_i, (predicted value) and $(\beta_0 + \beta_1 x_i)$.

→ Goal :- To Estimate β_0 and β_1 to best fit the data

→ Example Problem :

Q. Suppose we want to predict a student's test score (y) based on the number of hours they studied (x).

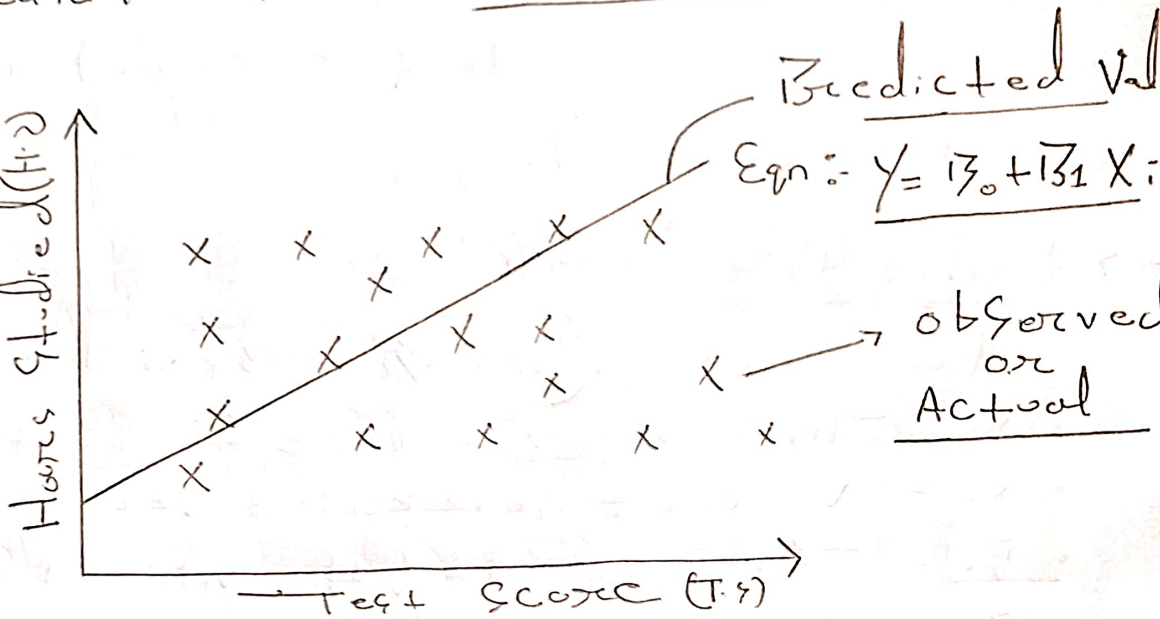
① stat → Data Set

(x_i) Hours Studied	Test Score (y_i)
1	50
2	55
3	65
4	70

Here (x_i) is no. of hours student studied and (y_i) observed value for every (x_i).

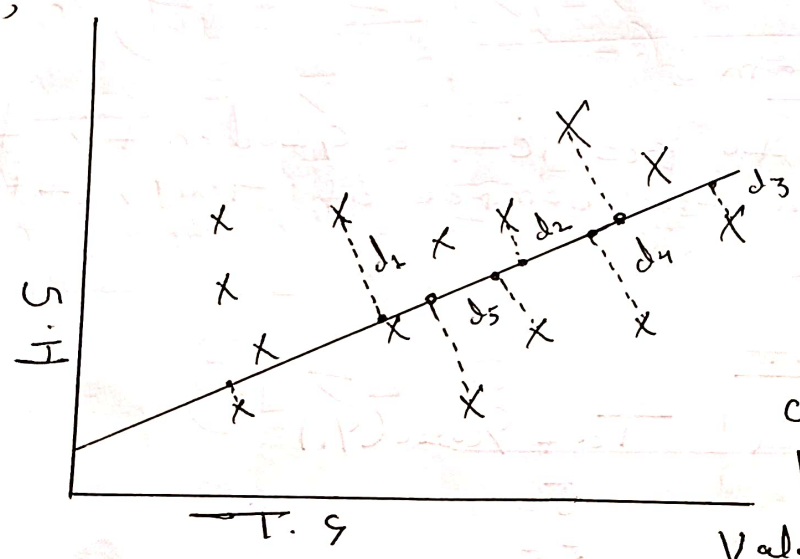
So we will show this Dataset to our M.L model or we will train our model, then it will predict answer for new data.

Graph :-



→ We want to plot a line that passes very close to observed values then only that line will be best fit line.

So,



d_1, d_2, d_3, d_4 are distance between actual and predicted values.

$d_1 \rightarrow (Y_i - \hat{Y}_i)$

but we need +ve values so $(Y_i - \hat{Y}_i)^2$.

Now for all points or $(X_i$'s)

① $E = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ ∴ Y_i = Actual observed

∴ $\hat{Y}_i$ = Predicted by model.

Let - Error fonct (loss function)

Now we have to minimize this distance, As I said line needs to be close to observed values.

We use OLS (ordinary least square) method, which minimizes the sum of squared residuals

(E) SSE $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$

Let

where $(\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i)$ Predicted value.

Now, upon taking partial derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, setting them to zero, and solving yields:

$= \left(\frac{\partial E}{\partial \hat{\beta}_0} = 0\right)$ and $\left(\frac{\partial E}{\partial \hat{\beta}_1} = 0\right)$

we will get →

⊙ $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)}$

⊙ $(\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x})$

Note :- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$: (Mean of predictor)

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$: (Mean of response)

Now we will substitute all these findings in Eqn $[\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i]$ and will get line.

Done (GLR) Simple linear regression

Important term

Goodness of Fit

The Coefficient of determination (R^2) measures the proportion of variance in y explained by x:

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

R^2 ranges from 0 to 1, higher values indicating a better fit.

X