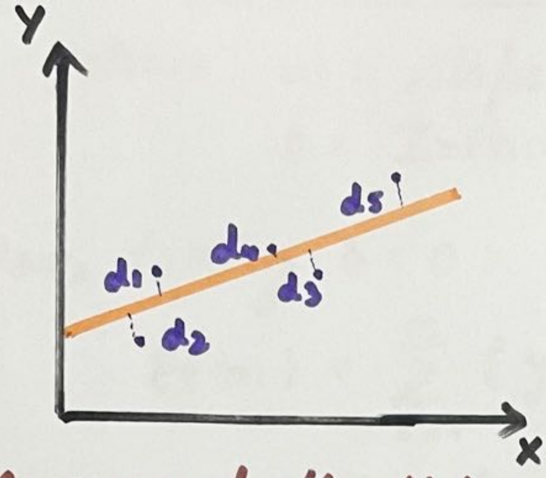
LINEAR REGRESSION



ERROR

- $E = d_1 + d_2 + d_3 + d_4 + d_5$

Squaring the distances

- $E = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2$

- We squared the distances because there might be some positive & negative values & they might cancel out each other.
- We didn't used modulus instead of squaring because squaring can penalise some points that are far & are considered as outliers. And also linear regression requires differentiation but the graph for the modulo function is continuous but not differentiable at origin.

$$\boxed{E = \sum_{i=1}^{n} d_i^2}$$

Error function
OR
Loss function

also represented as J sometimes.

Now, $d_i = y_i - \hat{y}_i$

where, y_i = actual value

$\hat{y}_i$ = predicted value

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

The main goal is to find the best line that will minimize the value of this equation.

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

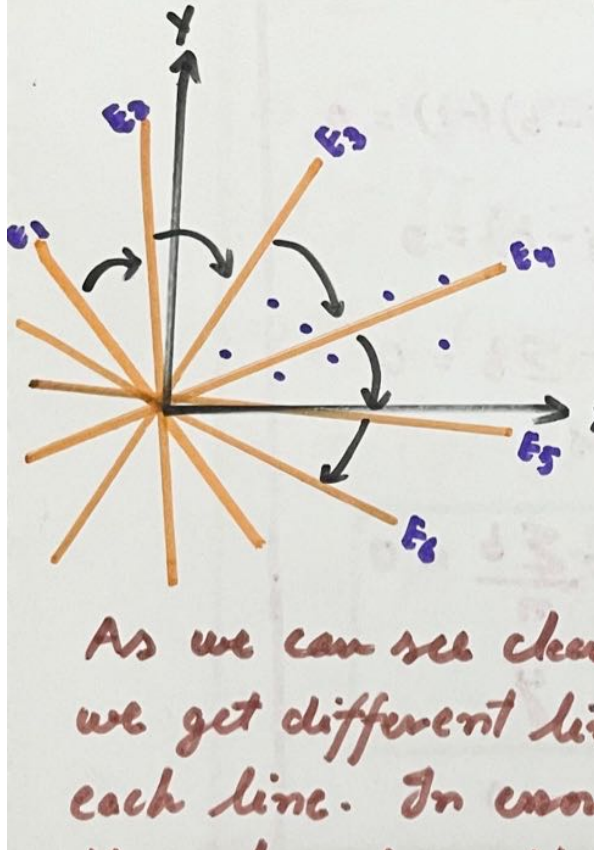
$\hat{y}_i = mx_i - b$

where, m = Slope of the line
b = Intercept of the line

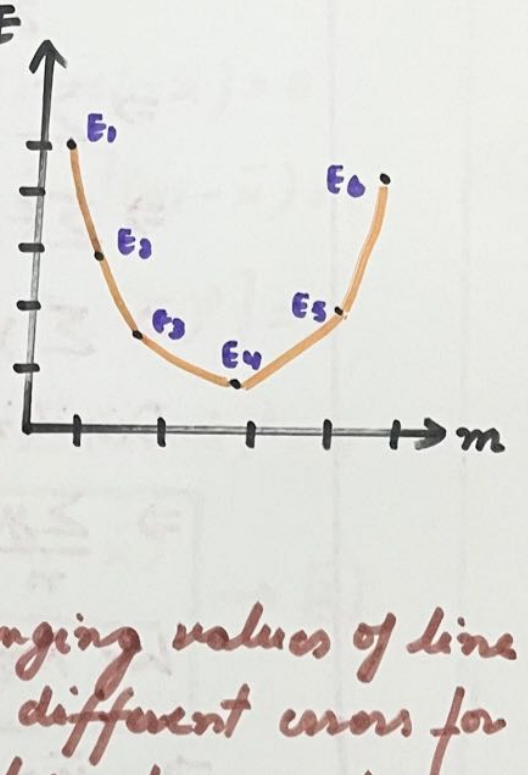
Now, taking $b = 0$:

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

Trying different values of m while the value of $b = 0$.



$\Longrightarrow$



As we can see clearly, by changing values of line we get different lines having different errors for each line. In error v/s m plot when we change the value of m, the error will start to reduce for a bit but after sometime it will again begin to increase. So our main motive is to find the global minima that will represent our best fit line.

Similarly when we plot the graph between errors in the lines & intercept taking $m = 1$ as constant, same type of curve will be obtained.

Now taking both m & b together & calculating their values to find the best fit line.

Finding derivative of E with respect to both m & b

For b:

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

$$\Rightarrow \sum \frac{\partial}{\partial b} (y_i - mx_i - b)^2 = 0$$

$$\Rightarrow \sum 2(y_i - mx_i - b)(-1) = 0$$

$$\Rightarrow \sum (-2)(y_i - mx_i - b) = 0$$

$$\Rightarrow \sum y_i - \sum mx_i - \sum b = 0$$

Dividing by n b/s

$$\Rightarrow \frac{\sum y_i}{n} - \frac{\sum mx_i}{n} - \frac{\sum b}{n} = 0$$

We know, $\frac{\sum y_i}{n} = \bar{y}$

$$\Rightarrow \bar{y} - m\bar{x} - \frac{\cancel{n} b}{\cancel{n}} = 0$$

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

$$\Rightarrow \boxed{b = \bar{y} - m\bar{x}} - \boxed{1}$$

For M:

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

Putting value of b obtained

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

$$\frac{\partial E}{\partial m} = \sum \frac{\partial}{\partial m} (y_i - m x_i - \bar{y} + m\bar{x})^2 = 0$$

$$\Rightarrow \sum 2(y_i - m x_i - \bar{y} + m\bar{x})(-x_i + \bar{x}) = 0$$

$$\Rightarrow \sum -2(y_i - m x_i - \bar{y} + m\bar{x})(x_i - \bar{x}) = 0$$

$$\Rightarrow \sum (y_i - m x_i - \bar{y} + m\bar{x})(x_i - \bar{x}) = 0$$

$$\Rightarrow \sum \left[(y_i - \bar{y}) - m(x_i - \bar{x})\right](x_i - \bar{x}) = 0$$

$$\Rightarrow \sum \left[(y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2\right] = 0$$

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

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