Date
Page

Linear Regression :—

1. check if data is linear or sort of linear.
2. in 99.99% case real world data is sort of linear (not perfectly straight)
3. Linear Regression makes a best fit straight line on this sort of linear data.
4. This line is chosen to minimize the error (difference between predicted and actual values)

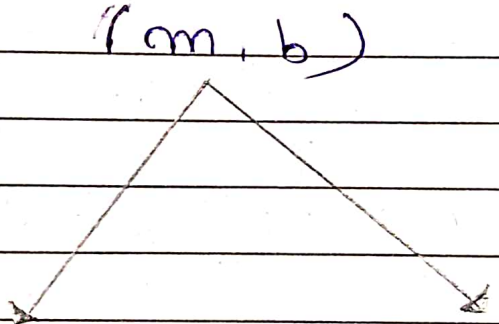
✗ Assumption of Linear Regression.

1. Linear Relationship :—
input and output should have a straight-line relationship.
2. No multicollinearity :—
input features should not be highly related to each other.
3. Normality of Residuals :—
The prediction errors should follow a normal (bell-shaped) distribution.

4. Homoscedasticity :—
Error should have constant spread across all values of input.

5. No autocorrelation :—
Errors shouldn't follow a pattern they should be random.

* Simple Linear Regression Mathematical formulation :—



Closed form solution
(direct formula)

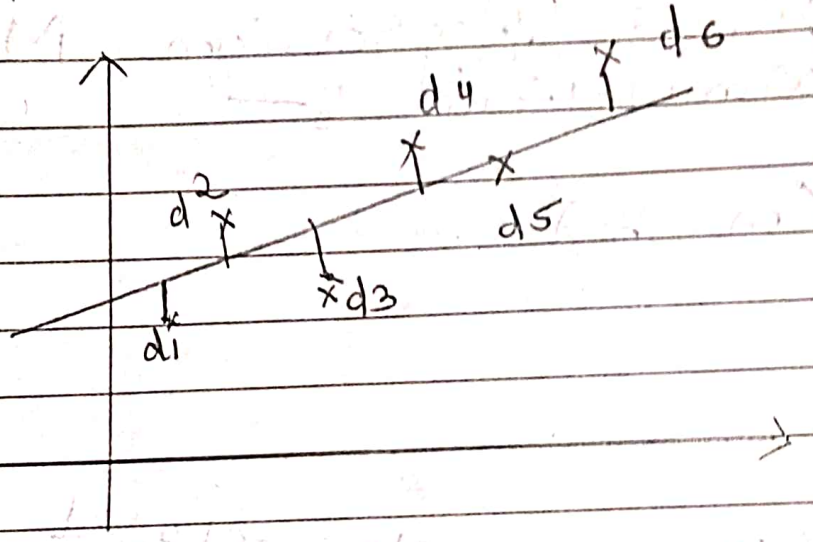
OLD
(ordinary least square)
(Low dimension)

Non-closed form solution
[Gradient descent]
(High dimension)

$$\bar{y} = mx + b$$

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

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



$$E = d_1 + d_2 + d_3 + \cdots\cdots d_n$$

Positive and negative error can cancel out each other so here we use square of error.

Why modulus is not use instead of square.?

→ Because modulus is not differentiable at origin and square is differentiable at all points.

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

$$\boxed{E = \sum_{i=1}^{n} d_i^2} \quad \dots\dots (1)$$ Error function

$$d_i = y_i - \hat{y}_i \quad \dots (2) \quad \because \text{(actual - predicted)}$$

Put (2) in (1)

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

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

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

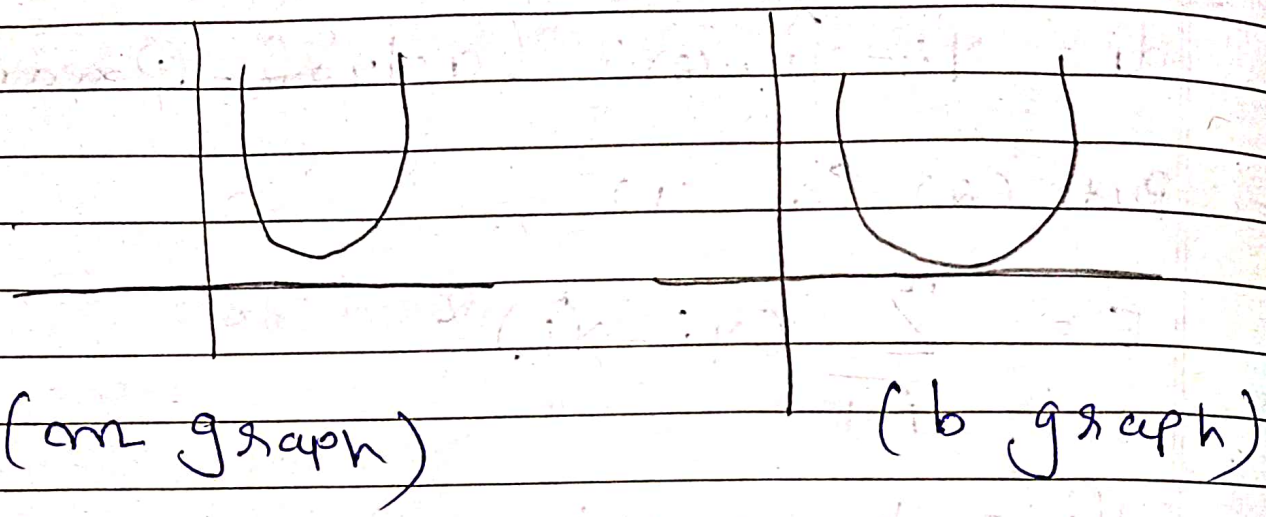
$$E(m, b) = \sum_{i=1}^{n} (y_i - mx_i - b)^2$$

Let $b = 0$

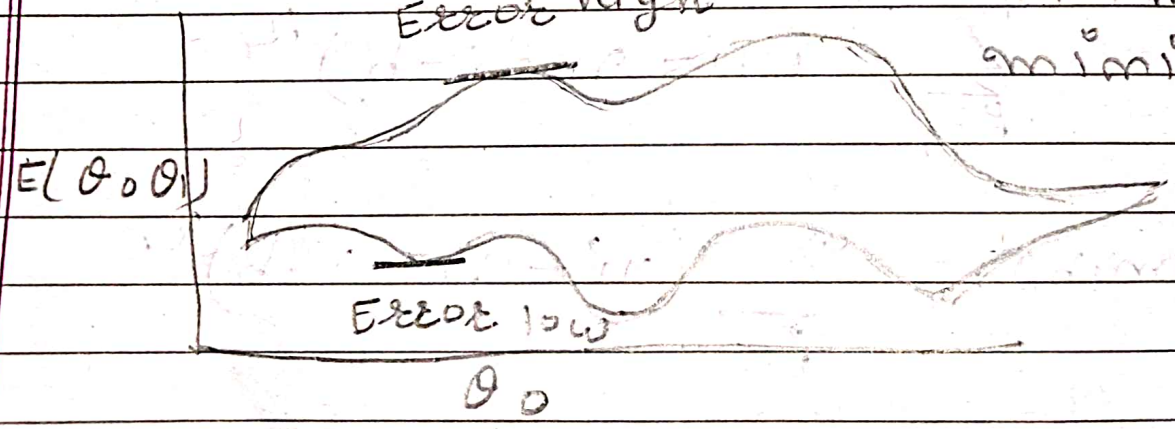
$$E(m) = \sum_{i=1}^{n} (y_i - mx_i)$$

Let $m = 1$

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



(m graph) (b graph)



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

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

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

divide both side by -2

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

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

divede both side by n

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

$$\bar{y} - m\bar{x} - \frac{nb}{n} = 0$$

$\left[\frac{y_i}{n} = \bar{y}, \frac{x_i}{n} = \bar{x}\right]$ $\because$ mean

$\left[\frac{b+b+b+\cdots\cdots+b}{n}\right]$ n times $\frac{nb}{n}$

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

$$b = \bar{y} - m\bar{x} \quad \cdots\cdots (3)$$

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

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

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

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

$$\sum [(y_i - \bar{y}) - m(x_i - \bar{x})] (x_i - \bar{x}$$

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

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

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