Linear Regression

The linear Regression are has three Types

1. Simple linear Regression
2. Multiple linear Regression
3. Polynomial linear Regression

1. Simple linear Regression
   Simple linear regression is a statistical technique used to model the relationship between two variables: a dependent variable (also called the response variable) and an independent variable (also called the predictor variable). It assumes a linear relationship between the two variables and aims to find the best-fit line that represents this relationship.

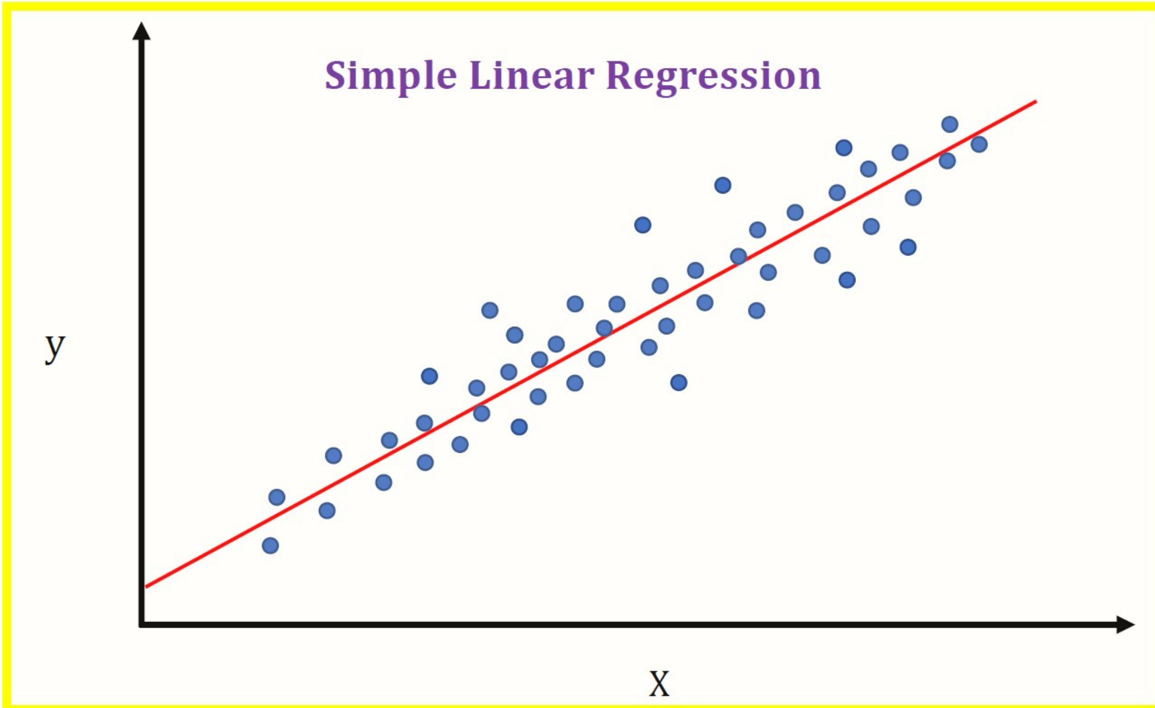
   The equation for simple linear regression can be written as:

   Equation of line is : $y = mx + b$

   Where:
   - **y** represents the dependent variable (response variable).
   - **x** represents the independent variable (predictor variable).
   - **b** is the y-intercept of the line (the value of y when x is 0).
   - **M** is the slope of the line (the change in y for a one-unit increase in x).

   The goal of simple linear regression is to estimate the values of **b and m** that minimize the sum of squared differences between the observed values of y and the predicted values based on the linear model. This process is typically done using the least squares method, as mentioned earlier.

   Once the values of **b and m** are determined, the best-fit line can be plotted, and it represents the linear relationship between the variables. This line can be used to predict the values of the dependent variable (y) based on different values of the independent variable (x).



$$(m, b)\ are\ slop\ and\ y_{intercept}$$

To find value of (m, b) we have two methods

1. Closed form expression

   In mathematics, a **closed-form expression** is a mathematical expression that is formed with constants, variables and a finite number of standard operations and functions, such as +, −, ×, ÷, *n*th root, exponentiation, logarithm, trigonometric functions, and inverse hyperbolic functions. Usually, no limits or integrals are accepted.

   Used directly formula. Sklearn used this method to find solution. Ordinary Least Squares (OLS) Regression.

2. Non closed form expression.
   A closed form solution provides an exact answer and one that is not closed form is an **approximation**, but you can get a non closed form solution as close as to a closed form solution as you want.

   Used in higher Dimension data set, Gradient Descent tech used. ML library is SGD-Regression.

1. Closed form expression (OLS)
   Equation of line is : $y = mx + b$

OLS:-

Equation of line :- $y = mx + b$

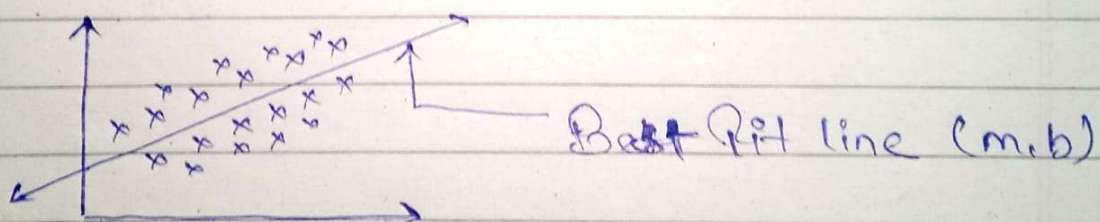
where:-

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

$\bar{x}, \bar{y}$ = Mean, x_i, y_i = current variable

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

Now, How to find m & b?

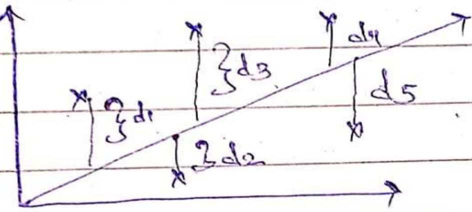


To find Best-fit line, we need to minimizing the error

To find the best fit line, we typically use a method called linear regression, which involves minimizing the error between the predicted values and the actual values of the data points. The most common approach to minimizing the error is called the least squares method.

To find Best-fit line. We need to minimizing the Errors

Best fit line.

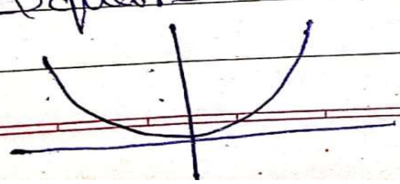


Error = $d_1 + d_2 + d_3 \dots d_n$

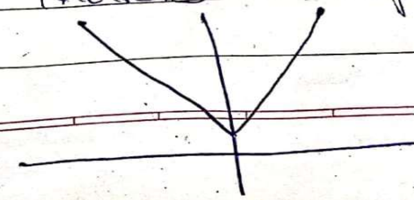
To Removing negative value

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

Square root



Model's Graph



When finding the error function in linear regression, we use the squared difference between the predicted values and the actual values rather than the absolute difference (or modulus). There are a few reasons for this choice:

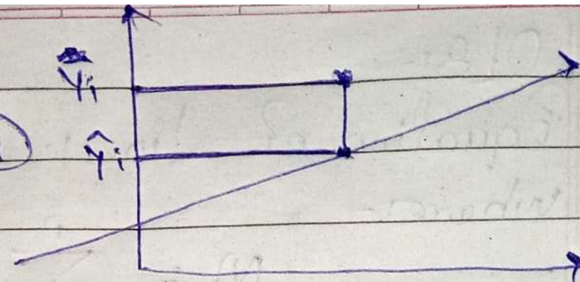
1. Mathematical Convenience: Using squared differences leads to a simpler and more convenient mathematical formulation. Squaring the differences eliminates the absolute value function, which simplifies calculations and allows us to use calculus for optimization purposes.

2. Emphasis on Larger Errors: Squaring the differences amplifies larger errors, giving them more weight in the optimization process. This behavior can be desirable when we want to prioritize reducing the impact of larger errors in our model.

3. Optimization: The squared error function is differentiable, which means we can find its minimum analytically by taking the derivative and setting it to zero. This allows us to use optimization techniques like gradient descent to efficiently find the best fit line.

Error Function

$$\boxed{E = \sum_{i=1}^{n} d_i^2} \quad \text{—} (i)$$



y_i – actual value
$\hat{y}_i$ – Predicted value

Total error

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

Average error

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

Formula of Predicted Value is :-

$$\boxed{\hat{y}_i = mx_i + b} \quad \text{———} (ii)$$

Putting equation (ii) in error Function eq (i)

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

By minimizing the error, we ensure that the line is as close as possible to the actual data points and provides a good representation of the relationship between the variables being studied.

We Required (m,b) value in which the error function is minimum.

For eq. of cgpi and package prediction date set

Y_i - predictor point in package and x_i = predictor point in cgpi

We can’t change the value of x and y, we only change the value of M and b

M = slope of line

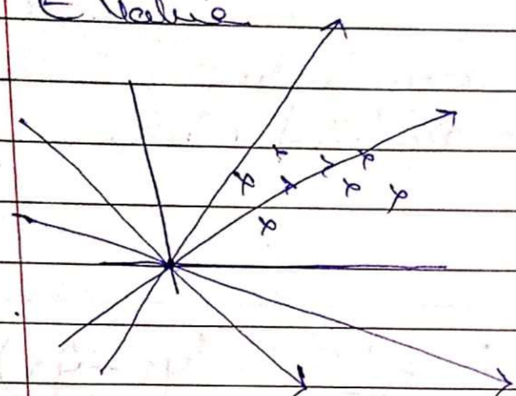
b = y_intercept

$$y = f(x)$$

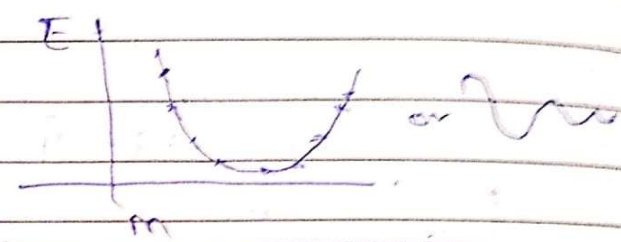
Change in x mean change in y similar

$$f(m, b)$$

let see, b is constant & we change m value & check E value



b = 0, then Y = f(m)

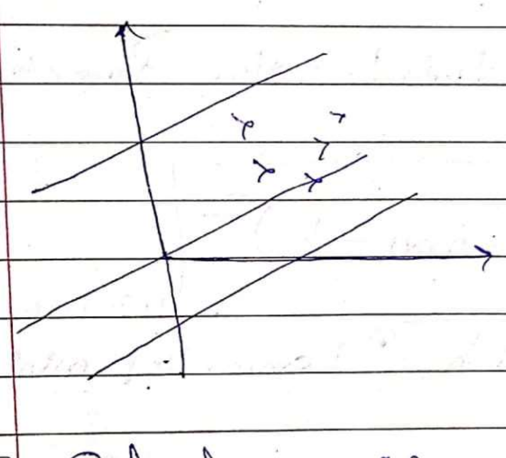


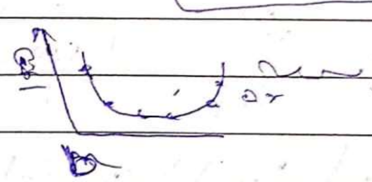
$$\boxed{E(m) = \sum_{i=1}^{n} (y_i - x_i M)^2}$$

Now M is constant & change b value

M = 1 then y = f(b)

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





Calculus - Max, min

$$E(x) = \frac{dE}{dx} = 0$$

$$E = E(x, y) = f(x, y) = f_2(m, b)$$

How do find b & m

First find b

$$E(b) = \sum (y_i - mx_i - b^2)$$

$$\frac{dE}{db} = \frac{d}{db} \sum_{i=1}^{n} (y_i - mx_i - b^2)$$

By chainrule ....

$$\frac{dE}{db} = \sum \frac{d}{db} (y_i - mx_i - b^2)$$

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

Divide -2 by both side

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

Divide n by both side.

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

$$\bar{y} - m\frac{\sum x_i}{n} - \frac{\sum b}{n}$$

b is constant :-

$$\frac{b + b + b \dots b_n}{n} = \frac{\cancel{n} b}{\cancel{n}} = b$$

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

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

Find m value now.

$$E = \sum (y_i - mx_i - b)^2$$

$$\frac{dE}{dm} = \sum \frac{d}{dm} (y_i - mx_i - b)^2$$

we now the value of b

$$\frac{dE}{dm} = \sum \frac{d}{dm} (y_i - mx_i - (\bar{y} + \bar{x}m))^2$$

By chainrule:-

$$\sum -2 (y_i - mx_i - \bar{y} + xm) \cdot \frac{d}{dm} (y_i - mx_i - \bar{y} + \bar{x}m)$$

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

Divide by -2 by both side.

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

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

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

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