Linear Regression

1) What is Linear Regression.
2) What are the Types of Linear Regression.
3) Linear Regression Line.
4) Finding the Best Fit Line
5) Cost function
6) Gradient Decent
7) Model Performance
8) Assumptions of Linear Regression.

⟹ Linear Regression is one of the Easiest and most popular Machine Learning Algorithms.

→ It is a statistical method that is used for predictive analysis.

⟹ Linear Regression makes predictions for Continuous / Real or numeric Variables Such as Sales, Salary, age, product Price, etc.

⟹ Linear Regression algorithm shows a linear Relationship between a dependent (y) and one or more independent (y) Variables, hence called as linear regression.

Types of Linear Regression.

i) Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependa variable, then such a Linear Regression algorithm is called simple Linear Regression.

ii) Multiple Linear Regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a lineal Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

A Linear line showing the Relationship between the dependent and independent variables is called a Regression Line.

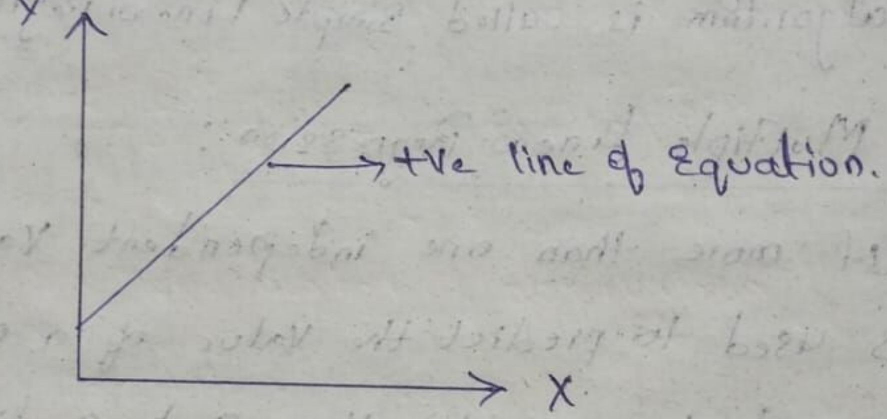
Two Types of Regression Line:-

(i) Positive Linear Relationship

(ii) Negative Lineal Relationship.

(i) Positive Linear Relationship.

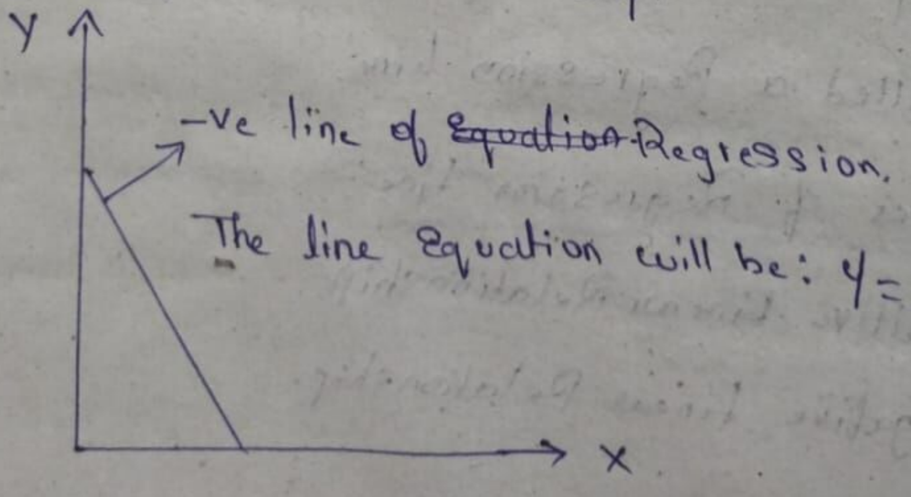
if the dependent Variable increases on the Y-axis and independent variable increases on X-axis, then such a Relationship is termed as a positive linear Relationship.



∴ The line Equation will be: $Y = a_0 + a_1X$

ii) Negative Linear Relationship:

if the dependent Variable decreases on the Y-axis and independent Variable increases on the X-axis, then such a relationship is called a Negative linear Relationship.



The line Equation will be: $y = -a_0 + a_1x$

Mean Squared Error.(MSE)

for Linear Regression, we use the (MSE) Mean Squared Error Cost function, which is the average of Squared Error occurred b/w the predicted Values and actual Values.

$$\boxed{MSE = \frac{1}{N}\sum_{i=1}^{n}\left(y_i - (a_i x_i + a_0)\right)^2}$$

N = Total number of observation

Y_i = Actual Value

$(a_1 x_i + a_0)$ = Predicted Value.

Residuals:

The distance between the actual Value and predicted Value is called Residual. If the observed points are far from the Regression line, then the residual will be high, and so cost function will high. if the Scatter points are close to the Regression line, then the residual will be ~~fam~~ Small and hence the cost function.

Gradient Decent

⟹ Gradient decent is used to minimize the MSE by calculating the gradient of the cost function,

⟹ A regression model uses gradient descent to ~~upgra~~ update the coefficients of the line by reducing the cost function.

→ it is done by random selection of values of coefficients and then iteratively update the values to reach the minimum cost function.

Model Performance

⟹ The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called optimization.

Assumptions of Linear Regression:-

⇒ **Linear relationship between the features and target:**

- Linear regression assumes the linear relationship between the dependend & independent variables

⇒ **Small or no Multicollinearity between the features:**

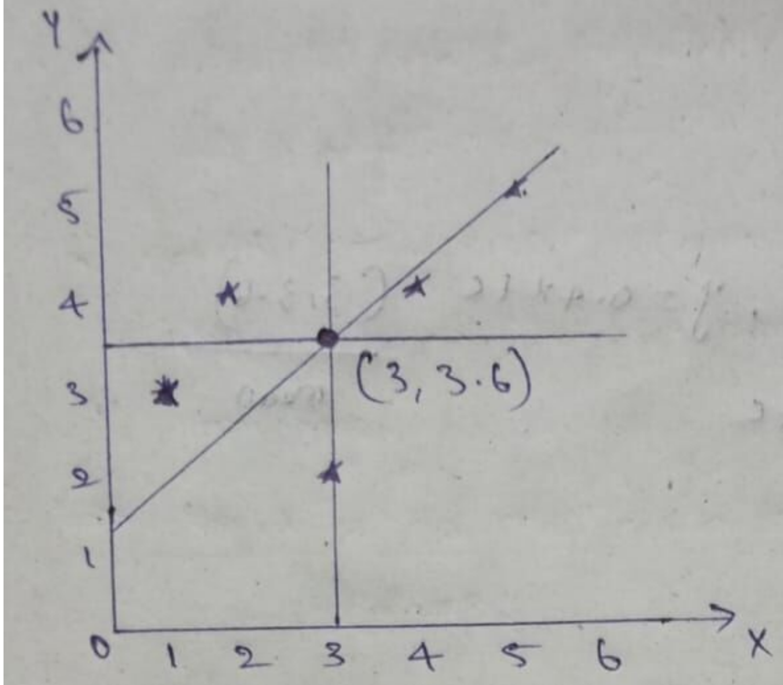
- Multicollinearity means high correlation b/w the independent variables. Due to multicollinearity, it may difficult to find the true relationship b/w the predictors and target variables.

Or we can say, it is difficult to determine which predictor variable is affecting the target variables and which is not. So, the model assumes either little or no multicollinearity b/w the features or independent variables.

⇒ **Homoscedasticity Assumption:**

Homoscedasticity is a situation when the error term is the same for all the values of independent variables. with homoscedasticity, there should be no clear pattern distibution of data in the scatter plot.

Linear Regression Example:-



X	Y
1	3
2	4
3	2
4	4
5	5

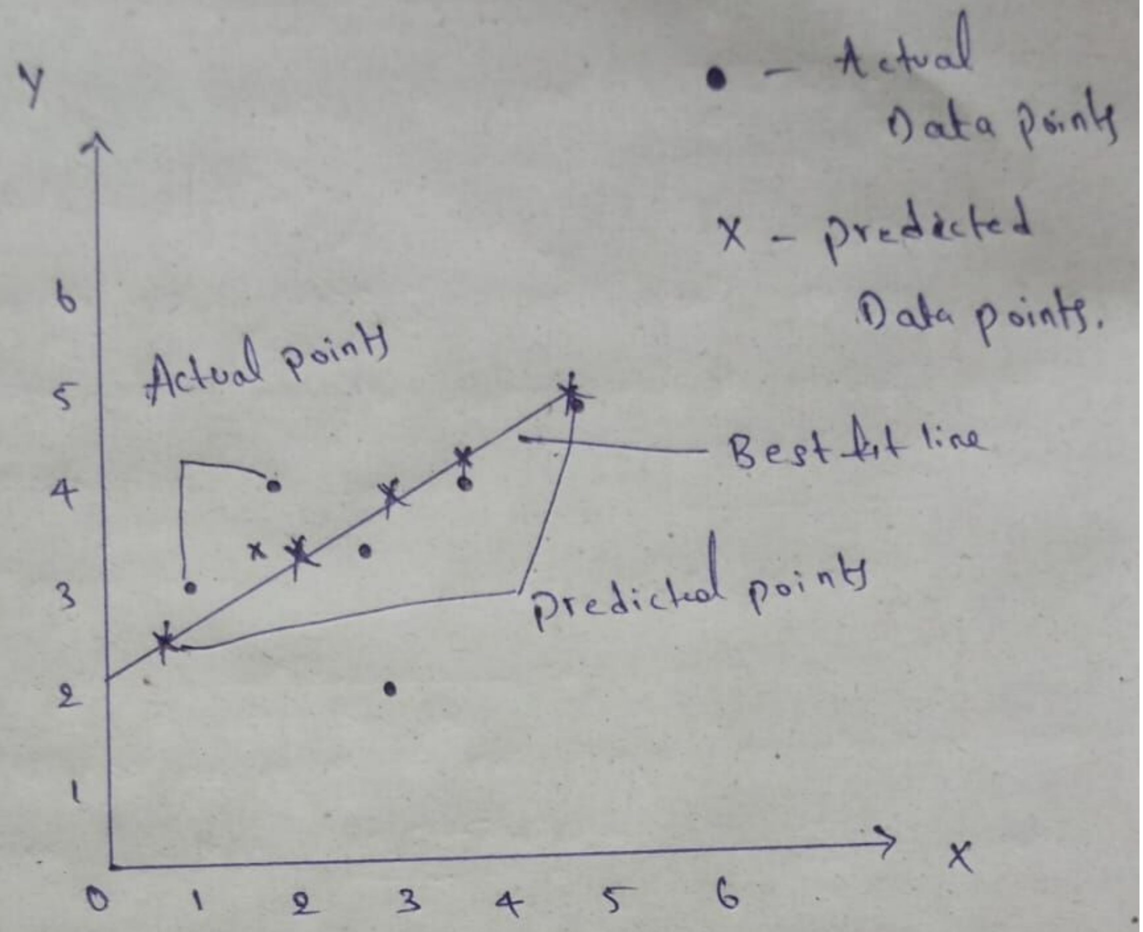
Mean of $x = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

Mean of $Y = \frac{3+4+2+4+5}{5} = \frac{18}{5} = 3.6$

$$\boxed{y = mx + c}$$

$$\boxed{m = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2}}$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4 $\Sigma=10$	2.8 $\Sigma=4$



Therefore we got Best fit line, The distance Between Actual points and predicted points is error.

$m = \frac{4}{10} = \cdot 0.4$

$m = 0.4$

$y = mx + c$

$y = \textcircled{m} \rightarrow 0.4 , \; y = 0.4x + c \qquad \underset{\text{mean}}{\underline{(3, 3.6)}}$

$3.6 = 0.4(3) + c$

$3.6 = 1.2 + c$

$3.6 - 1.2 = c$

$\underline{\underline{c = 2.4}}$

$\underline{y = 0.4x + 2.4}$ Regression line

(i) $x = 1$

$y = 0.4(1) + 2.4 = 2.8$

(ii) $x = 2$

$y = 0.4(2) + 2.4 = 3.2$

(iii) $x = 3$

$y = 0.4(3) + 2.4 = 3.6$

(iv) $y \; x = 4$

$y = 0.4(4) + 2.4 = 4.0$

(v) $x = 5$

$y = 0.4(5) + 2.4 = 4.4.$

⟹ Normal distibution of error terms:

- Linear regression assumes that the error term should follow the normal distibution pattern. if error terms are not normally distibuted, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding Coefficients.

It can be checked using the q-q plot. If the plot shows a straight line without any deviation, which means the error is normally distributed.

⟹ No Autocorrelations:

- The linear regression model assumes no autocorrelation in error terms. if there will be any correlation in the error term, then it will drastically reduce the accuracy of the ~~#~~ model. Autocorelation usually occurs if there is a dependency between residual errors.

R-Squared method

→ R-Squared is a Statistical method that determines the goodness of fit.

→ It measures the steength of the relationship between the dependent and independent Variables on a scale of 0 - 100%.

⇒ The high Value of R-Square determines the less difference between the predicted Values and actual Values and hence represents a good model.

⇒ It is also called a coefficient of determination or Coefficient of multiple determination for multiple regression.

$$\boxed{\text{R-Squared} = \frac{\text{Explained Variation}}{\text{Total Variation}}}$$

Finding the Best fit line

* Our main goal is to find the Best fit line it is the Error between predicted Values and actual Values Should be minimized.

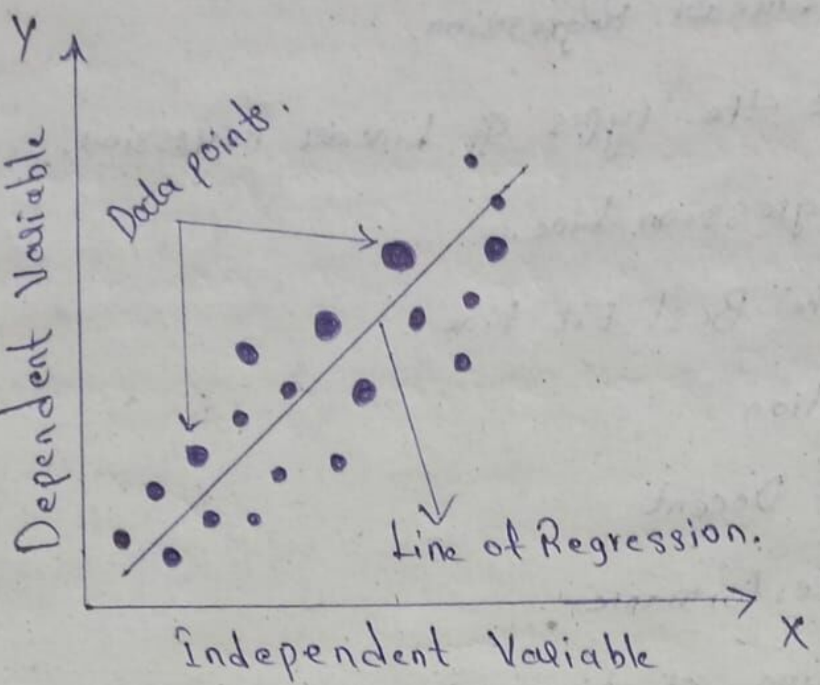
* The Best fit line will have the least Error.

* The different Value for weights or the Coefficient of lines, (a_0, a_1) gives a different line of regression, so we need to calculate the Best Values for a_0 & a_1 to find the Best fit line, so to calculate this we use cost function.

Cost function.

⇒ The cost function is used to find the accuracy of mapping function, which maps the input Variable to the output Variable.

This mapping function is also known as Hypothesis function.



Mathematical Representation.

$$\boxed{y = a_0 + a_1 x + \varepsilon}$$

- y = Dependent Variable (Target variable)
- x = Independent Variable (Predictor Variable)
- $a0$ = intercept of the line
  (Gives an additional degree of freedom)
- a_1 = Linear regression Coefficient
  (Scale factor to each input value)
- ε = Random Error.
  (ε = Epsulain)