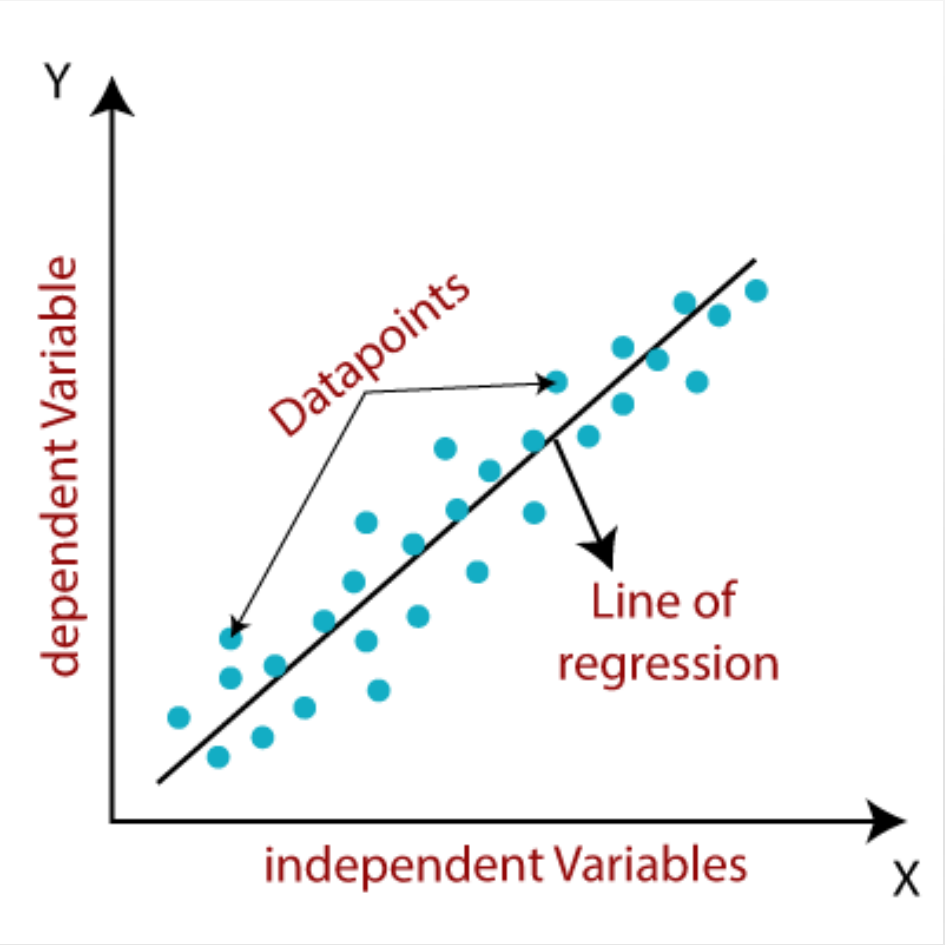
Linear Regression

Linear Regression

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.

provides a sloped straight line representing the relationship between the variables. Consider the image:



- Mathematically

$$y = a_0 + a_1 x + \varepsilon$$

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

Types of Linear Regression

Simple Linear Regression:
If a single independent variable is used to predict the value of a numerical dependent variable

Multiple Linear Regression:
If more than one independent variable is used to predict the value of a numerical dependent variable

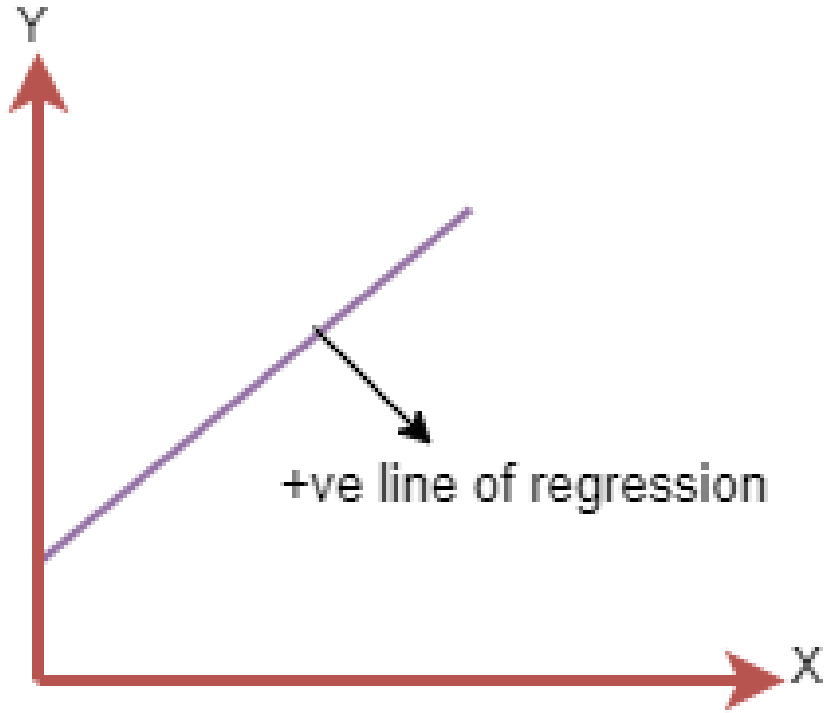
Linear Regression Line

- Linear line showing the relationship between the dependent and independent variables is called a **regression line**.

- It can show two types of relationship:

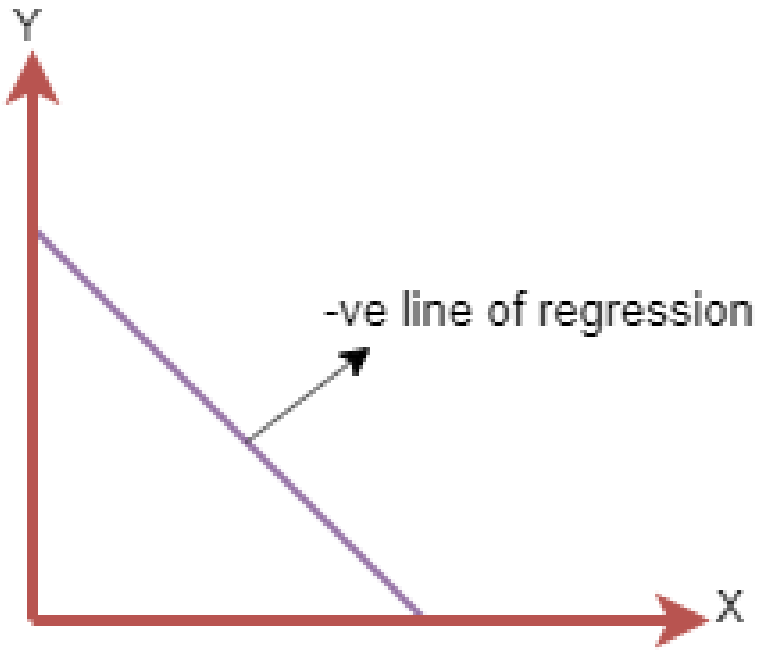
1. **Positive Linear Relationship**
2. **Negative Linear Relationship**

Positive Linear Regression



The line equation will be: $Y = a_0 + a_1 x$

Negative Linear Regression



The line of equation will be: $\mathbf{Y= -a_0+a_1x}$

Finding the best line

- The error between predicted values and actual values should be minimized. The best fit line will have the least error.
- The different values 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 and a_1 to find the best fit line, so to calculate this we use cost function.

Cost Function

- The cost function is used to estimate the values of the coefficient for the best fit line.
- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
- Also used to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.

MSE

Mean Squared Error (MSE) cost function:

- The average of squared error occurred between the predicted values and actual values.

$$MSE = 1\frac{1}{N}\sum_{i=1}^{n}(y_i - (a_1 x_i + a_0))^2$$

N=Total number of observation
Yi = Actual value
$(a1x_i+a_0)$= Predicted value.

Residuals:

The distance between the actual value and predicted values 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 small and hence the cost function.

Gradient Descent

- It is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.

Model Performance

- How the line of regression fits the set of observations. This is called Optimization.
- It can be achieved by **R-Square** method.

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression

Assumptions of Linear Regression

- Below are some important assumptions of Linear Regression.

✓ **Linear relationship between the features and target**

✓ **Small or no multicollinearity between the features**

✓ **Homoscedasticity Assumption**

✓ **Normal distribution of error terms**

✓ **No autocorrelations**