**LINEAR REGRESSION:**

**DEFINITION:**

Linear regression finds out the relationship between a dependent variable and an independent variable. It performs the task to predict a dependent variable(‘**y’**=**output**) based on a given independent variable(**‘x’**=**input**). So, this regression technique finds out a linear relationship between x and y. Hence, the name is Linear Regression.

There are two types of liner regressions:

1: Simple linear regression

2: multiple linear regression

**SIMPLE LINEAR REGRESSION:**

In simple linear regression, we aim to reveal the relationship between a single independent variable or you can say input, and a corresponding dependent variable or output. We can discuss this in a simple line as

It can be described as a method of statistical analysis that can be used to study the relationship between two quantitative variables.

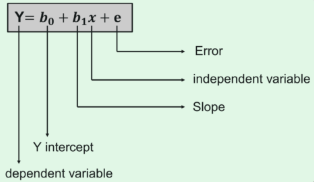
There are two things which can be found out by using the method of simple linear regression:

1. **Strength of the relationship between two variables.**

**Example:** The relationship between global warming and the melting of glaciers)

1. **How much the value of the dependent variable is at a given value of the independent variable.**

**Example:** the amount of melting of a glacier at a certain level of temperature)

****

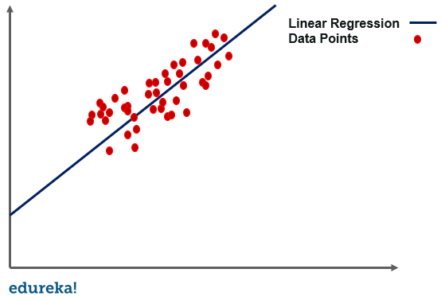
**MULTIPLE LINEAR REGRESSION:**

Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

To be precise:

In Simple linear regression it allows to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables, an independent variable and a dependent variable. A multiple regression model extends to several explanatory variables.

Multiple linear regression is used to determine a mathematical relationship among several random variables. It examines how multiple independent variables are related to one dependent variable. Once each of the independent factors has been determined to predict the dependent variable, the information on the multiple variables can be used to create an accurate prediction on the level of effect they have on the outcome variable. The model creates a relationship in the form of a straight line (linear) that best approximates all the individual data points.



**Linear regression itself is a part of several regression types:**

**Types of regressions are:**

1: linear regression

2: logistic regression

3: Ridge regression

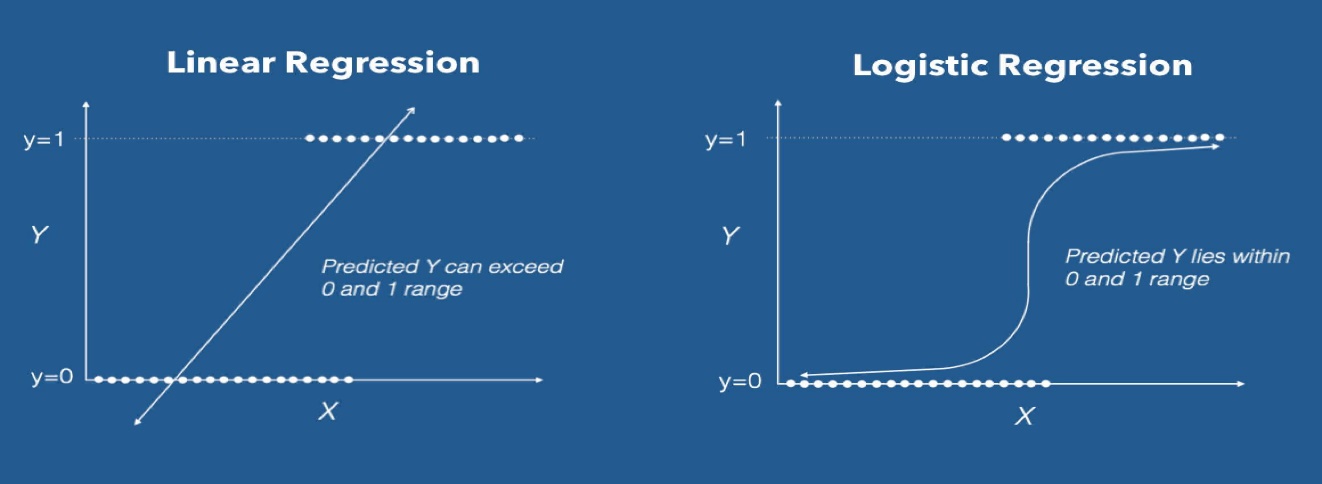
4: Lasso regression

5: Polynomial regression

**Logistic regression:**

Logistic regression uses a sigmoid curve to show the relationship between the target and independent variables. logistic regression works best with large data sets that have an almost equal occurrence of values in target variables

Let’s see the difference between linear regression and logistic regression.

****

Here we have taken two variables to find the predicted y, in linear regression the predicted y can be exceeded out of the range as it is a straight line. But in logistic regression because of taking curved line, it lines within range.

**Ridge regression:**

Ridge regression is used when there is a high correlation between independent variables. It is known as a regularization technique, and is used to reduce the complexity of the model. It introduces a small amount of bias which, using a bias matrix, makes the model less susceptible to overfitting.

**Lasso regression:**

Just like ridge regression, lasso regression is another regularization technique that reduces the model’s complexity. It does so by prohibiting the absolute size of the regression coefficient. This causes the coefficient value to become closer to zero, which does not happen with ridge regression.

**Polynomial regression:**

Polynomial regression models a non-linear dataset using a linear model. It is the equivalent of making a square peg fit into a round hole. It works in a similar way to multiple linear regression, but uses a non-linear curve. It is used when data points are present in a non-linear fashion.

The model transforms these data points into polynomial features of a given degree, and models them using a linear model. This involves best fitting them using a polynomial line, which is curved, rather than the straight line seen in linear regression.

There are more types of regression analysis than those listed here, but these five are probably the most commonly used. Make sure you pick the right one, and it can unlock the full potential of your data, setting you on the path to greater insights.

# 

# **Assumptions of Linear Regression**

To conduct a simple linear regression, one has to make certain assumptions about the data. This is because it is a parametric test. The assumptions used while performing a simple linear regression are as follows:

Let us take to variables x and y one is independent and another is dependent

1. **Linear relationship:**

There exists a linear relationship between the independent variable, x, and the dependent variable, y.

1. **Independence:**

The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

1. **Homoscedasticity:**

The residuals have constant variance at every level of x.

1. **Normality:**

The residuals of the model are normally distributed. If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.

**Advantages of linear regression:**

* Linear Regression is simple to implement and easier to interpret the output coefficients.
* When you know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because of it is less complexity to compared to other algorithms.
* Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization techniques and cross-validation.
* Easier to implement, interpret and efficient to train

**Disadvantages of Linear gradient:**

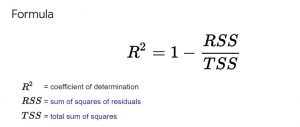
* In linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique.
* Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes.
* But then linear regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.
* Linear regression is quite sensitive to outliers

**R-Squared;**

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

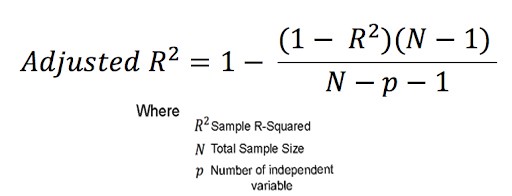
**Formula for R-Squared is:**

R^2= 1-Un explained variation/Total variation



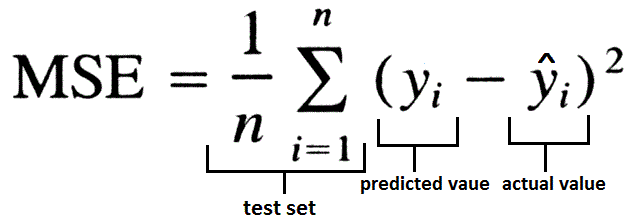
**Adjusted R-Squared:**

Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.



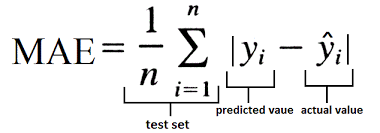
**Mean Squared Error:**

The mean squared error (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. It’s called the mean squared error as you’re finding the average of a set of errors



**Mean Absolute Error:**

Mean Absolute Error is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set.



**Root Mean Squared Error:**

Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. Formally it is defined as follows:



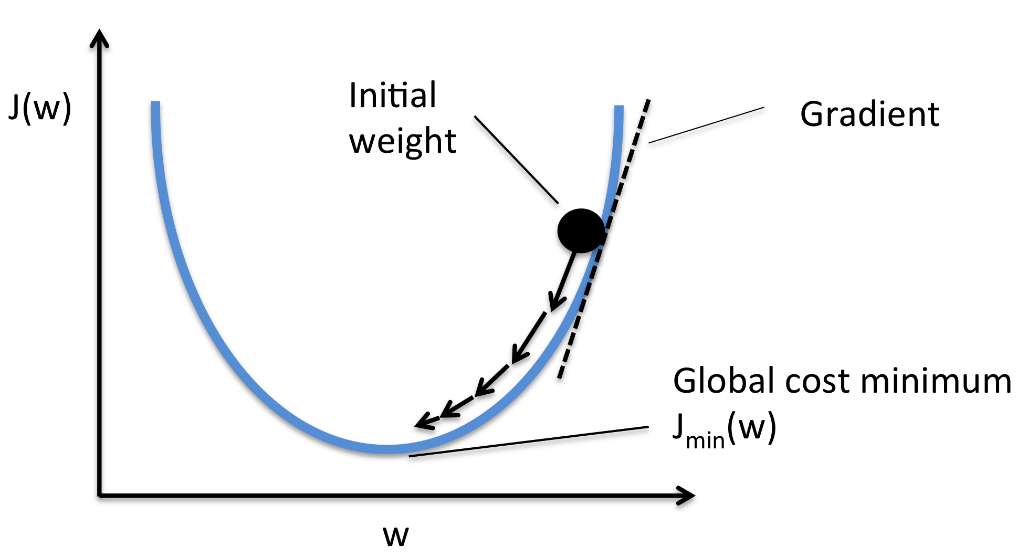
**Gradient Descent:**

Let’s know what is a gradient.

A gradient measure how much the output of a function changes if you change the inputs a little bit.

A gradient simply measures the change in all weights with regard to the change in error. You can also think of a gradient as the slope of a function.

Gradient Descent is a simple optimization technique for finding minimum of any function, in this case we want to find the minima of our MSE function.



You can visualize the function as a valley, and you are standing on some random point. Your aim is to reach the bottom most point of the valley and Gradient Descent help you in reaching to the bottom most point of the valley.

You will see and make out where the slope of the valley is going down, then you start taking steps towards the down going slope. Once you see that every contour around you is higher than where you are standing you claim that you have reached the bottom most point.

It has to know where the slope of the valley is going (and it doesn’t have eyes as you do) so it takes the help of mathematics here.

To know the slope of a function at any point you differentiate that point with respect to its parameters, thus Gradient Descent differentiates the above Cost function and comes to know the slope of that point.

To go to the bottom most point it has to go in the opposite direction of the slope, i.e., where the slope is decreasing.

It has to take small steps to move towards the bottom point and thus learning rate decides the length of step that gradient descent will take.

After every move it validates that the current position is global minima or not. This is validated by the slope of that point, if the slope is zero then the algorithm has reached the bottom most point.

After every step, it updates the parameter and by doing the above step repeatedly it reaches to the bottom most point.

**Mathematical & Geometrical Intuition of Linear Regression:**

**Step 1**: Plot of the Independent & Dependent Variables. Draw the best fit line (Approx.).

Best Fit Line is determined based on the sum of errors to be minimum.

**Step 2**: Calculate the individual errors.

Error is defined based on the actual & the predicted value. Below is the formula to calculate the error for individual points.

In our Sample Example, we have 3 points which are not in line with our best fit line. sSo, we need to calculate the error function.

**Error=y-y^**

The calculation of error can conclude the status of selecting the best fit line.

Step 3: Calculating the minimum sum of squares of errors or Ordinary Least-squares.

As we have calculated the individual Error for all the points, we are now going to sum & consider the min value to evaluate the best fit line by using the below formula.