- **Linear Regression** is a simple yet powerful and mostly used algorithm in data science.

Introduction

Models use machine learning algorithms, during which the machine learns from the data just like humans learn from their experiences. Machine learning models can be broadly divided into two categories based on the learning algorithm which can further be classified based on the task performed and the nature of the output.

Supervised learning methods: It contains past data with labels which are then used for building the model.

- **Regression**: **The output variable to be predicted is *continuous* in nature**, e.g. scores of a student, diamond prices, etc.
- **Classification**: **The output variable to be predicted is *categorical* in nature**, e.g.classifying incoming emails as spam or ham, Yes or No, True or False, 0 or 1.

2. **Unsupervised learning methods: It contains no predefined labels assigned to the past data.**

- **Clustering**: No predefined labels are assigned to groups/clusters formed,e.g. customer segmentation.

Linear Regression is a supervised learning algorithm in machine learning that supports finding the ***linear* correlation among variables**. The result or output of the regression problem is a real or continuous value.

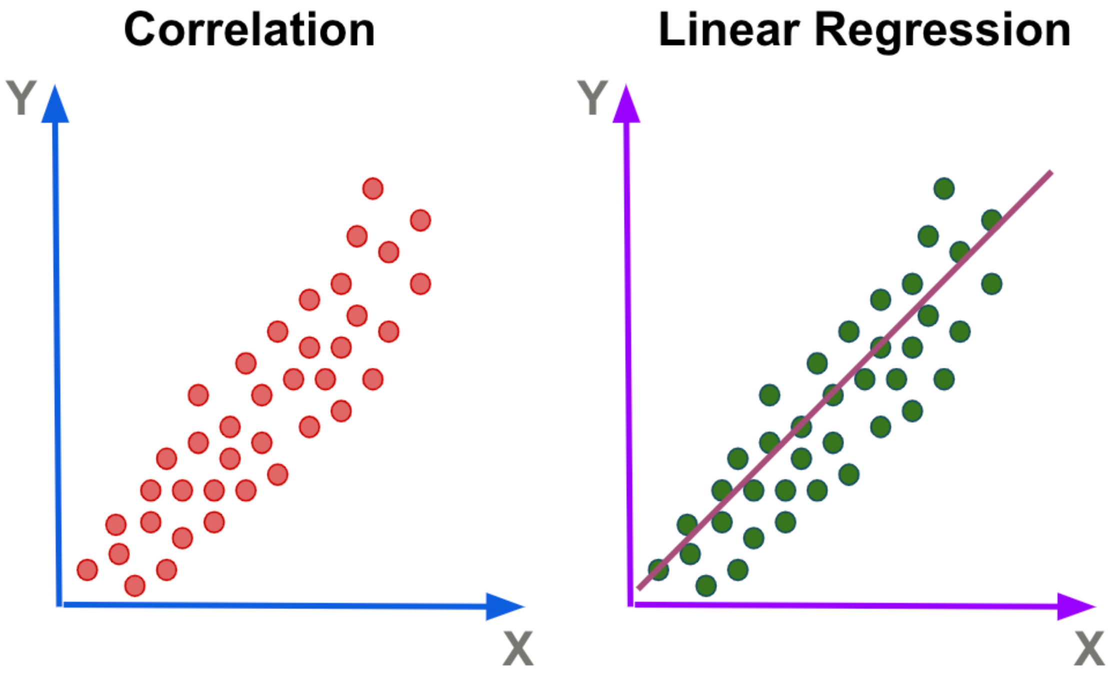
What is Linear Regression?

Linear regression is a **type of statistical analysis used to predict the relationship between two variables.** It assumes a linear relationship between the independent variable and the dependent variable, and **aims to find the best-fitting line** that describes the relationship. **The line is determined** by **minimizing the sum of the squared** differences between the predicted values and the actual values.

Simple Linear Regression

In a simple linear regression, there **is one independent variable and one dependent variable.** The model estimates **the slope and intercept of the line of best fit,** which represents the relationship between the variables. The **slope represents the change in the dependent variable for each unit change in the independent variable,** while the **intercept represents the predicted value of the dependent variable when the independent variable is zero**.

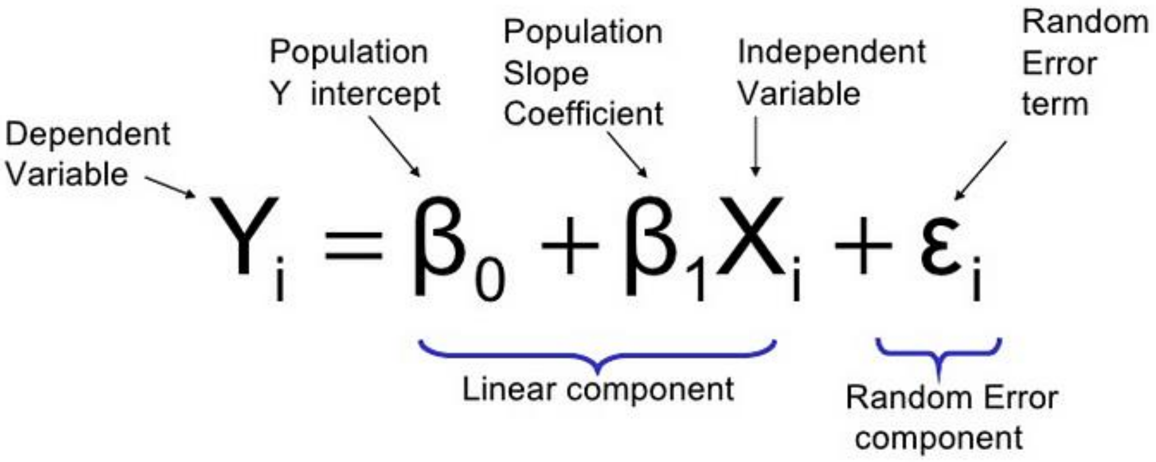
Linear regression is a quiet and the simplest statistical regression method used for predictive analysis in machine learning. Linear regression shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent(output) variable i.e. Y-axis, called linear regression. If there is a single input variable **X**(independent variable), such linear regression is called ***simple linear regression***.



To calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,

$$Yi = \beta o + \beta 1Xi$$

where Y_i = Dependent variable, $\boldsymbol{\beta_0}$ = constant/Intercept, $\boldsymbol{\beta_1}$ = Slope/Intercept, $\mathbf{X_i}$ = Independent variable.



The goal of the linear regression algorithm is to get the **best values for B_0 and B_1** to find the best fit line. The best fit line is a line **that has the least error** which means the error between predicted values and actual values should be minimum.

Random Error(Residuals)

In regression, the difference between the observed value of the dependent variable($\mathbf{y_i}$) and the predicted value(**predicted**) is called the **residuals.**

What is the best fit line?

In simple terms, the best fit line is a line that fits the given **scatter plot in the best way**. Mathematically, the best fit line is obtained by minimizing the Residual Sum of Squares(RSS).

Cost Function for Linear Regression

The cost function helps to work out the **optimal values** for B_0 and B_1, which provides the best fit line for the data points.

In Linear Regression, generally **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the $\mathbf{y_{predicted}}$ and $\mathbf{y_i}$.

We calculate MSE using simple linear equation y=mx+b:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (\mathrm{B1}x_i + \mathrm{B0}))^2$$

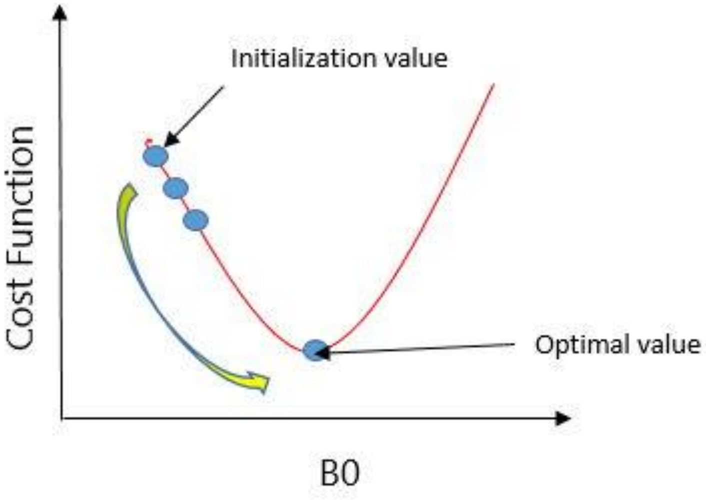
Using the MSE function, we'll update the values of B_0 and B_1 such that the MSE value settles at the minima.

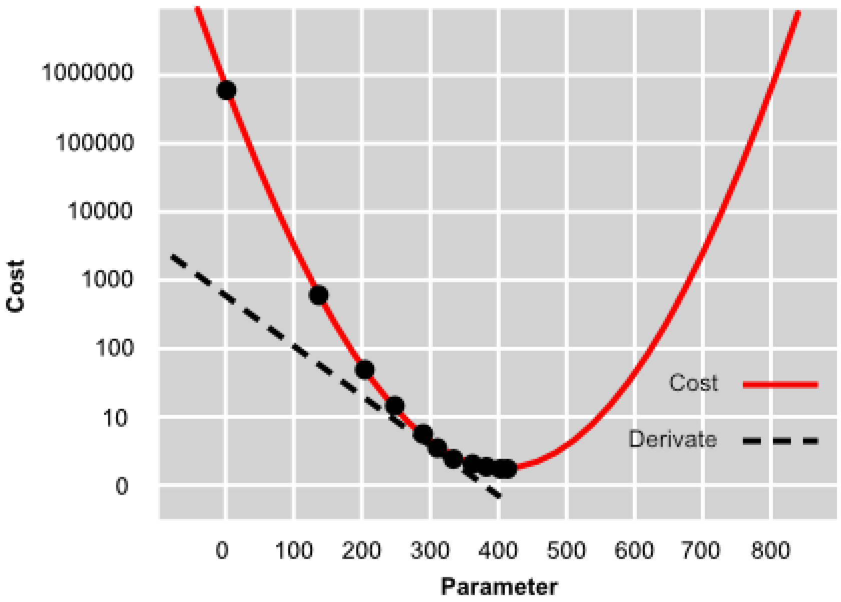
Gradient Descent for Linear Regression

Gradient Descent is one of the optimization algorithms that optimize **the cost function(objective function) to reach the optimal minimal solution.** To find the optimum solution we need to reduce the cost function(MSE) for all data points.

This is done by updating the values of B_0 and B_1 iteratively until we get an optimal solution.

A regression model optimizes the gradient descent algorithm to update the coefficients of the line by reducing the cost function by randomly selecting coefficient values and then iteratively updating the values to reach the minimum cost function.





Let’s define our Gradient Descent for Simple Linear Regression case:

First, the hypothesis expressed by the linear function:

$$h\text{o}x=\theta\text{o}+\theta1x$$

Parametrized by:

$$\theta\text{o}\theta1$$

We need to estimate the parameters for our hypothesis, with a cost function, define as:

$$J\left(\theta_0, \theta_1\right) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_0 x^i - y^i\right)^2$$

Evaluation Metrics for Linear Regression

Evaluation Metrics for Linear Regression

The **strength of any linear regression** model can be assessed using **various evaluation metrics**. These evaluation metrics usually provide a measure of how well the observed outputs are being generated by the model.

The most used metrics are,

1. Coefficient of Determination or **R-Squared (R2)**
2. **Root Mean Squared Error (RSME)** and **Residual Standard Error (RSE)**

R-Squared is a number that explains the amount of variation that is explained/captured by the developed model. **It always ranges between 0 & 1** . Overall, the higher the value of R-squared, the better the model fits the data.

Mathematically it can be represented as,

$R^2 = 1 – (RSS/TSS)$

- **Residual sum of Squares (RSS)** is defined as the **sum of squares of the residual for each data point in the plot/data.** It is the measure of the difference between the expected and the actual observed output.

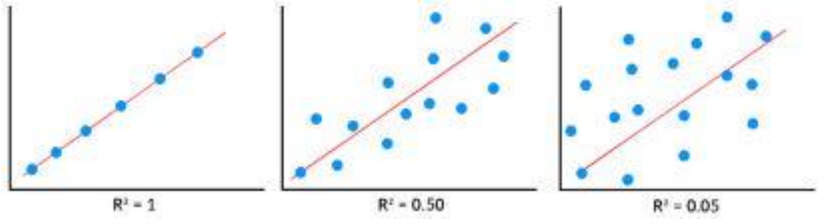
$$RSS = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

- **Total Sum of Squares (TSS)** is defined as the **sum of errors of the data points from the mean of the response variable**. Mathematically TSS is,

$$TSS = \sum (y_i - \bar{y_i})^2$$

Where **y hat is the mean of the sample data points.**

The significance of R-squared is shown by the following figures,



Root Mean Squared Error

The Root Mean Squared Error is the **square root of the variance of the residuals**. It specifies the absolute fit of the model to the data i.e. how close the observed data points are to the predicted values. Mathematically it can be represented as,

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\sum_{i=1}^{n}\left(y_i^{Actual} - y_i^{Predicted}\right)^2 \Big/ n}$$

To make this estimate unbiased, one has to divide the sum of the squared residuals by the **degrees of freedom** rather than the total number of data points in the model. This term is then called the **Residual Standard Error(RSE)**. Mathematically it can be represented as,

$$RSE = \sqrt{\frac{RSS}{df}} = \sqrt{\sum_{i=1}^{n}\left(y_i^{Actual} - y_i^{Predicted}\right)^2 \Big/ (n-2)}$$

R-squared is a better measure than RSME. Because the value of Root Mean Squared Error depends on the units of the variables (i.e. it is not a normalized measure), it can change with the change in the unit of the variables.