**Machine Learning**

The use and development of computer systems that learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data. Learning can be classified into two types:-**Supervised and Unsupervised**

**Supervised Learning**

1. Supervised learning is the learning of the model where with input variable (say, x) and an output variable (say, Y) and an algorithm to map the input to the output.  
   That is, **Y = f(X)**
2. The basic aim is to approximate the mapping function (mentioned above) that when there is a new input data (x) then the corresponding output variable can be predicted.
3. It is called supervised learning because the process of learning (from the training dataset) can be thought of as a teacher who is supervising the entire learning process. Thus, the “learning algorithm” iteratively makes predictions on the training data and is corrected by the “teacher”, and the learning stops when the algorithm achieves an acceptable level of performance.
4. **Supervised Learning methods - Regression and Classification**

**Unsupervised Learning**

1. Unsupervised learning is where only the input data (say, X) is present and no corresponding output variable is there.
2. The main aim of Unsupervised learning is to model the distribution in the data in order to learn more about the data.
3. It is called so, because there is no correct answer and there is no such teacher (unlike supervised learning). Algorithms are left to their own devises to discover and present the interesting structure in the data.
4. **Unsupervised Learning methods - Clustering and Dimensionality Reduction**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr No.** | **Parameters** | **Supervised Learning** | **Unsupervised Learning** |
| 1) | Input Data | Provided as input | Provided as input |
| 2) | Complexity | Complex | Less Complex |
| 3) | Labels/Target Feature/Output | Given | To be evaluated |
| 4) | Accuracy and Reliability | Accurate and reliable results | Less Accurate |

**Regression**

1. Regression analysis is a form of **Predictive modeling technique** which investigates the relationship between a dependent (target) and independent variable(s).
2. **Y = f(X), X= Independent variable/(s), Y= Dependent variable**
3. A regression problem is when the output variable is a real or continuous value, such as “salary”, “Weight”, ”Sales”, ”Profit”, ”Price”, ‘Revenue’ etc. Many different models can be used, the simplest is the linear regression.
4. The dependent variable is **continuous** in nature for Regression.

**Types of Regression**

1. **Simple Linear Regression**
2. **Multiple Linear Regression etc.**

**Simple Linear Regression**

1. It is a regression model that estimates the relationship between **one independent variable** and **one dependent variable** using a straight line.
2. It has an equation of the form **y = ax + b** or **y = mx + c**

Where-

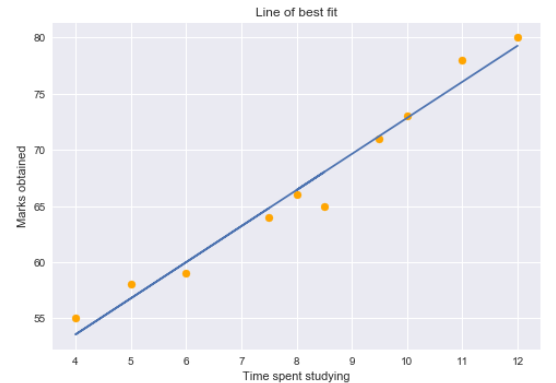
**x** = independent variable/ input feature/input variable/input column

**y** = dependent variable / output feature/target variable/ output column

**a/m** = slope / coefficient / weight / how much we expect **y** to change as **x** changes

**b/c** = intercept / constant / bias

In this graph, **x = Time spent Studying, y = Marks obtained**. The orange dots are the corresponding data points. The blue line is the best fit line for Linear regression**(y = mx +c)**



**Multiple Linear Regression**

1. **Multiple linear regression** is used to estimate the relationship between **two or more independent variables** and **one dependent variable.**
2. It has an equation of the form **y = a1x1 + a2x2 + a3x3 +…..+ anxn + b**

where:-

**x1 - xn** = independent variables/ input features

**y** = dependent variable / output feature

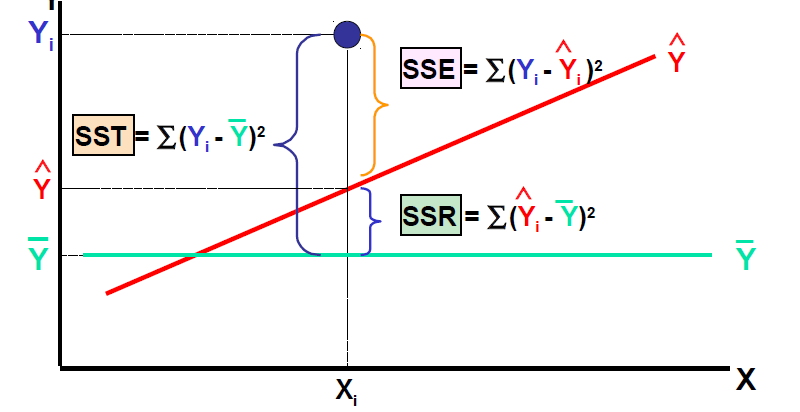
**a1- an** = coefficients/slope corresponding to **x1 - xn**

**b** = intercept / constant / bias

1. It is practically out of scope to represent multiple linear regression on a scatter chart with actual data points and a regression line as the scatter chart would have to span multiple dimensions corresponding to each independent variable and likewise the Regression line would also span across multiple dimensions

**Best Fit Line**

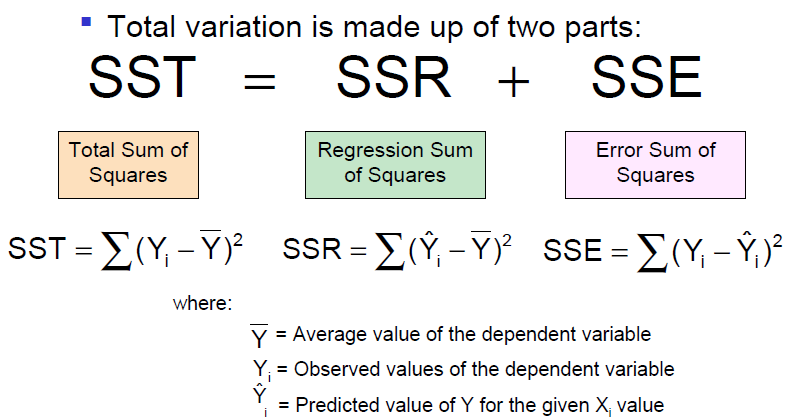
The Linear Regression model finds the best fit line that minimizes the squared error between the actual data points and its perpendicular dropped on the predicted line.



**SST = Total Sum of Squares ­**

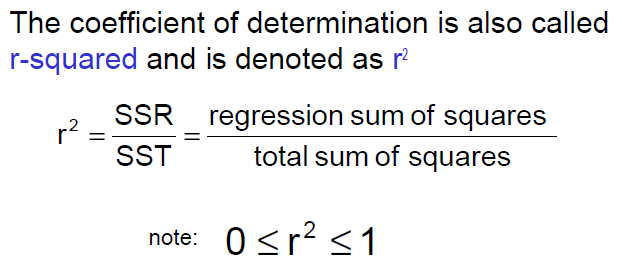
**SSE = Error/Residual Sum of squares**

**SSR = Regression sum of squares**



**R2 score (Coefficient of Determination or Goodness of Fit)**

The coefficient of determination is the portionof the total variation in the dependent variable that is explained by variation in the independent variable

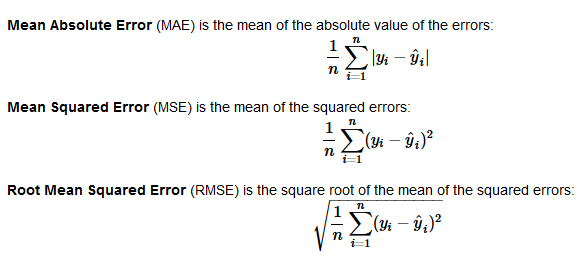
****

**SST = SSE + SSR**

**R2 score or R squared= SSR/SST = (SST- SSE)/ SST = 1 – SSE/SST**

When SSE= 0, R2 score = 1 and when SSE = SST, R2 score = 0

**Regression Metrics**

****

**Pros and Cons of Linear Regression (Simple and Multiple Linear Regression)**

**Pros**

1. Linear Regression performs well when the dataset is linearly separable
2. Linear Regression is easier to implement, interpret and efficient to train.

**Cons**

1. Main limitation of Linear Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data may not always be linearly dependent.
2. If variance of data points is not uniform throughout, these data points may become outliers and can degrade the performance of the model.