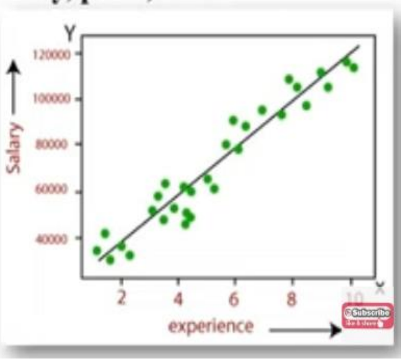
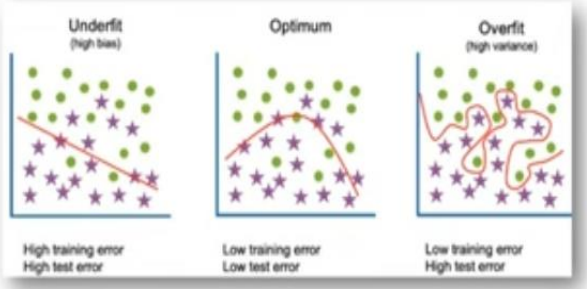
Introduction of Regression

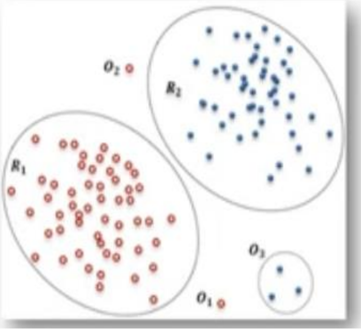
- Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
- More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.
- It predicts continuous/real values such as **temperature, age, salary, price,** etc.



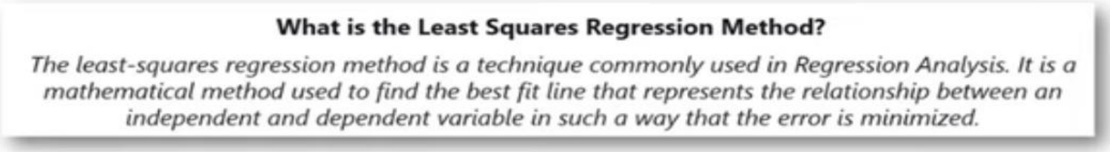
Terminologies Related to the Regression Analysis

1. **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called target variable.
2. **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a predictor.
3. **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.
4. **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called Overfitting. And if our algorithm does not perform well even with training dataset, then such problem is called underfitting.



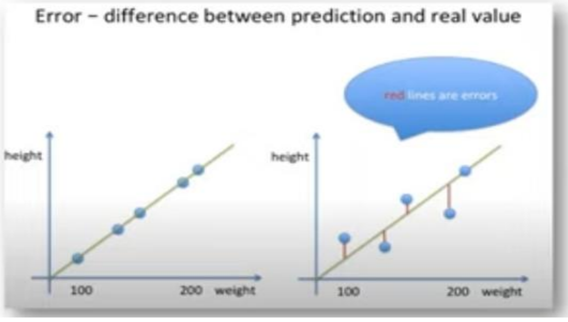


Least Square Method

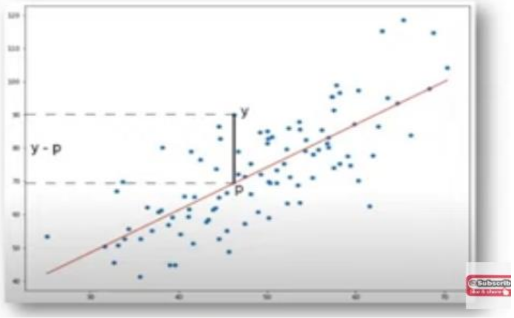


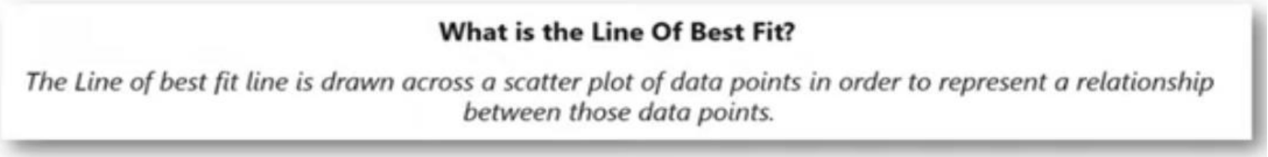
What is the Least Squares Regression Method?

The least-squares regression method is a technique commonly used in Regression Analysis. It is a mathematical method used to find the best fit line that represents the relationship between an independent and dependent variable in such a way that the error is minimized.



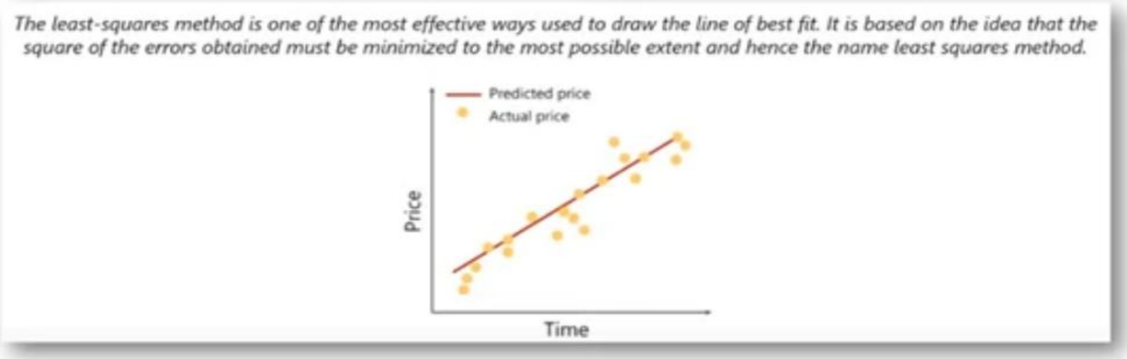
Error – difference between prediction and real value





What is the Line Of Best Fit?

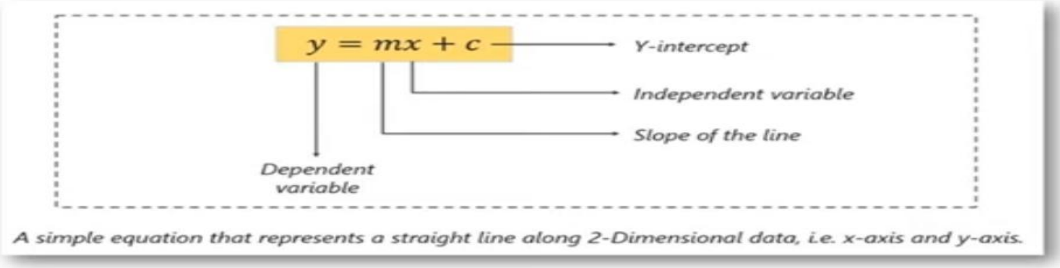
The Line of best fit line is drawn across a scatter plot of data points in order to represent a relationship between those data points.



The least-squares method is one of the most effective ways used to draw the line of best fit. It is based on the idea that the square of the errors obtained must be minimized to the most possible extent and hence the name least squares method.

Steps to calculate the Line of Best Fit

To start constructing the line that best depicts the relationship between variables in the data, we first need to get our basics right.



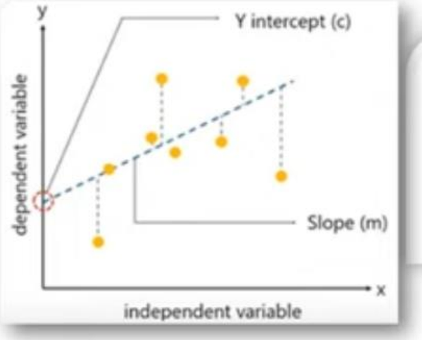
$$y = mx + c$$

A simple equation that represents a straight line along 2-Dimensional data, i.e. x-axis and y-axis.

Steps to Perform Least Square Method

Step 1: Calculate the slope 'm' of the line

The slope of a line characterizes the direction of a line. To find the slope, you divide the difference of the y-coordinates of 2 points on a line by the difference of the x-coordinates of those same 2 points .



$$m = \frac{n\sum xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

m – slope of the line

n – total number of data points

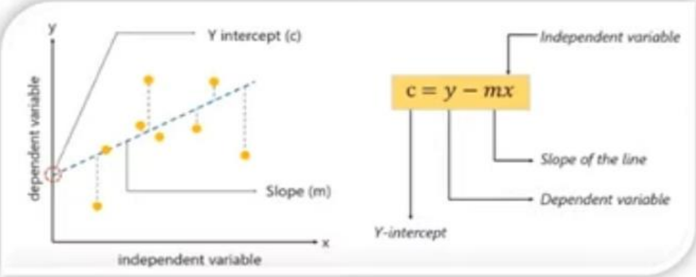
x – Independent variable

y – Dependent variable

Steps to Perform Least Square Method

Step 2: Compute the y-intercept

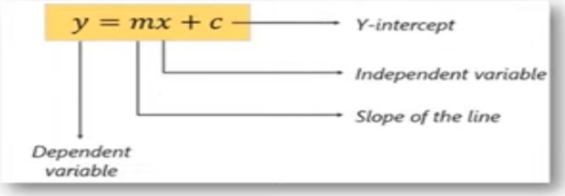
The y-intercept of a line is the value of y at the point where the line crosses the y axis.



Steps to Perform Least Square Method

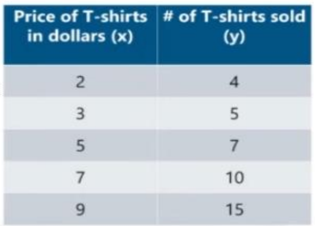
Step 3: Substitute the values in the final equation

A simple equation that represents a straight line along 2-Dimensional data, i.e. x-axis and y-axis.



Example

- **Tom who is the owner of a retail shop, found the price of different T-shirts vs the number of T-shirts sold at his shop over a period of one week.**
- Let us use the <u>concept of least squares regression to find the line of best fit for the below data.</u>



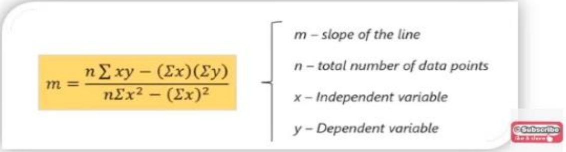
Price of T-shirts in dollars (x)	# of T-shirts sold (y)
2	4
3	5
5	7
7	10
9	15

Step 1: Calculate the **slope 'm'** x axis by using the following formula

$$m = \frac{n\sum xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

After you substitute the respective values,

m = 1.518 approximately.



$$m = \frac{n\sum xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

m – slope of the line

n – total number of data points

x – Independent variable

y – Dependent variable

Example

Step 2: Compute the **y-intercept** value (m=1.518 as per Step 1)

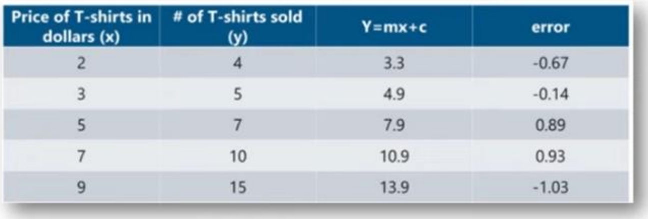
$$c = y - mx$$

After you substitute the respective values, **c = 0.305** approximately.

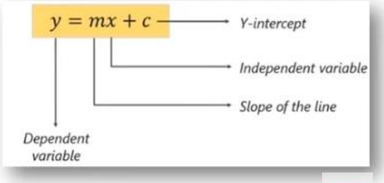
Step 3: Substitute the values in the final equation

$$y = mx + c$$

Once you substitute the values, it should look something like this



Price of T-shirts in dollars (x)	# of T-shirts sold (y)	Y=mx+c	error
2	4	3.3	-0.67
3	5	4.9	-0.14
5	7	7.9	0.89
7	10	10.9	0.93
9	15	13.9	-1.03



$y = mx + c$

Y-intercept

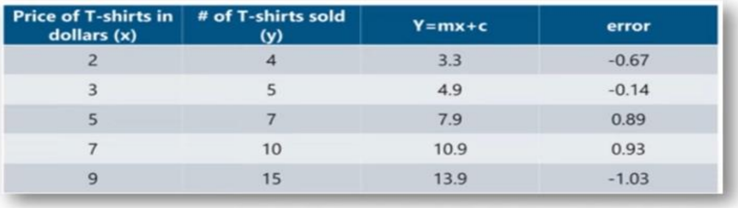
Independent variable

Slope of the line

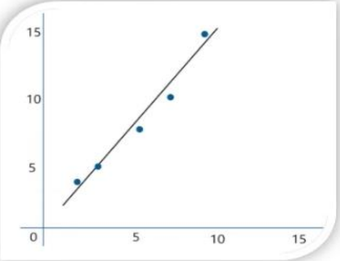
Dependent variable

Example

- Let's construct a graph that represents the **y=mx + c** line of best fit:



Price of T-shirts in dollars (x)	# of T-shirts sold (y)	Y=mx+c	error
2	4	3.3	-0.67
3	5	4.9	-0.14
5	7	7.9	0.89
7	10	10.9	0.93
9	15	13.9	-1.03



- Now **Tom can use the above equation to estimate <u>how many T-shirts of price $8 can he sell at the retail shop.</u>**

y = 1.518 x 8 + 0.305 = 12.45 T-shirts

- This comes down to 13 T-shirts!