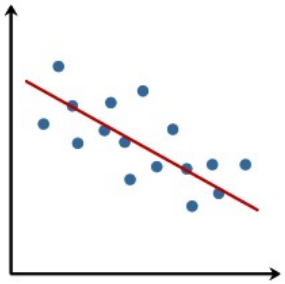
# Least Square Regression

Tushar B. Kute,
http://tusharkute.com



# Least Square Regression

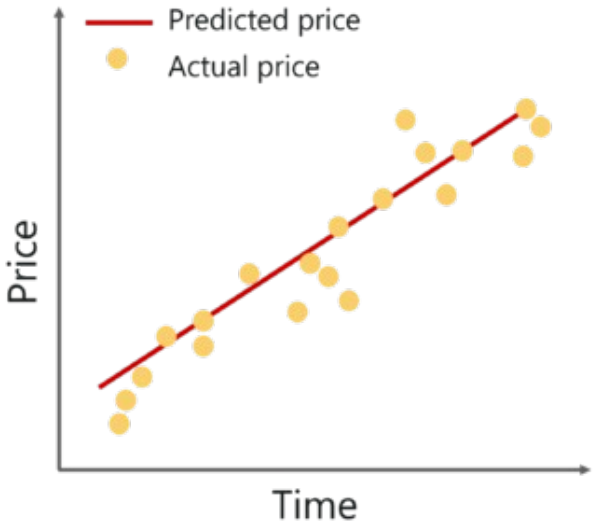
- The least-squares regression method is a technique commonly used in Regression Analysis.
- It is a mathematical method used to find the best fit line that represents the relationship between an independent and dependent variable.
- To understand the least-squares regression method lets get familiar with the concepts involved in formulating the line of best fit.

# What is line of best fit ?

- Line of best fit is drawn to represent the relationship between 2 or more variables. To be more specific, the best fit line is drawn across a scatter plot of data points in order to represent a relationship between those data points.
- Regression analysis makes use of mathematical methods such as least squares to obtain a definite relationship between the predictor variable (s) and the target variable.
- 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.

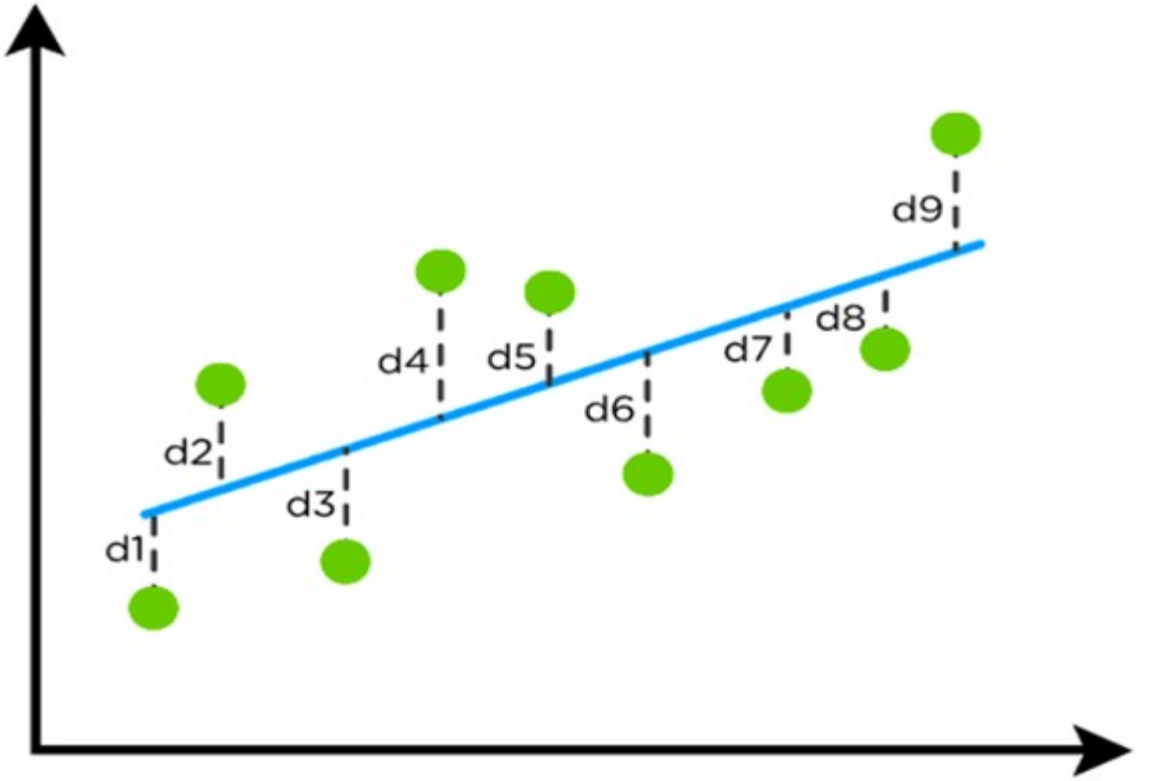
# Visualizing

- If we were to plot the best fit line that shows the depicts the sales of a company over a period of time, it would look something like this:




- Notice that the line is as close as possible to all the scattered data points. This is what an ideal best fit line looks like.

# Visualizing




$D = d1^2 + d2^2 + d3^2 + d4^2 + d5^2 + d6^2 + d7^2 + d8^2 + d9^2$

The regression line (blue) has the least value of D

# Calculate 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. Take a look at the equation below:

$$y = mx + c$$

- Surely, you've come across this equation before. It is a simple equation that represents a straight line along 2 Dimensional data, i.e. x-axis and y-axis. To better understand this, let's break down the equation:

  y: dependent variable

  m: the slope of the line

  x: independent variable

  c: y-intercept

# Calculate line of best fit

- Step 1: Calculate the slope ‘m’ by using the following formula:

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

- Step 2: Compute the y-intercept (the value of y at the point where the line crosses the y-axis):

$$c = y - mx$$

- Step 3: Substitute the values in the final equation:

$$y = mx + c$$

# Example

- Consider an 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.
- He tabulated this like shown below:

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

# Step-1

- Let us use the concept of least squares regression to find the line of best fit for the above data.
- Step 1: Calculate the slope ‘m’ 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.

# Step-2

- Step 2: Compute the y-intercept value

$$c = y - mx$$

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

# Step-3

- Step 3: Substitute the values in the final equation

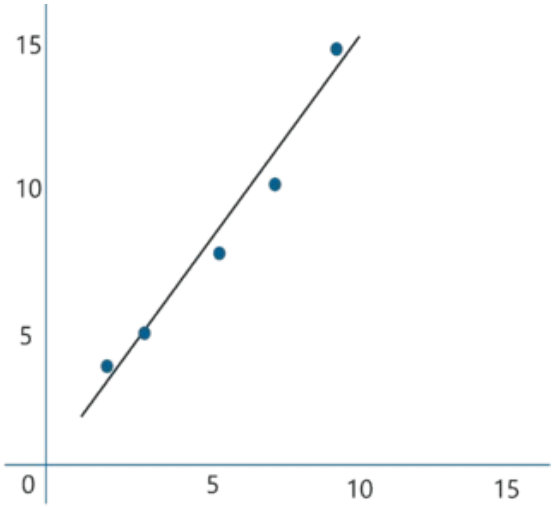
$$y = mx + c$$

- 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 |

# Predicting

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




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

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

# Summary

- The least squares regression method works by minimizing the sum of the square of the errors as small as possible, hence the name least squares. Basically the distance between the line of best fit and the error must be minimized as much as possible.
- A few things to keep in mind before implementing the least squares regression method is:
  - The data must be free of outliers because they might lead to a biased and wrongful line of best fit.
  - The line of best fit can be drawn iteratively until you get a line with the minimum possible squares of errors.
  - This method works well even with non-linear data.
  - Technically, the difference between the actual value of ‘y’ and the predicted value of ‘y’ is called the Residual (denotes the error).

# Example:

- Go practical...

# Thank you



@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

**Web Resources**
https://mitu.co.in
http://tusharkute.com

**contact@mitu.co.in**

**tushar@tusharkute.com**