- **Least Squares Methods**

Machine Learning

Dr. Jagendra Singh

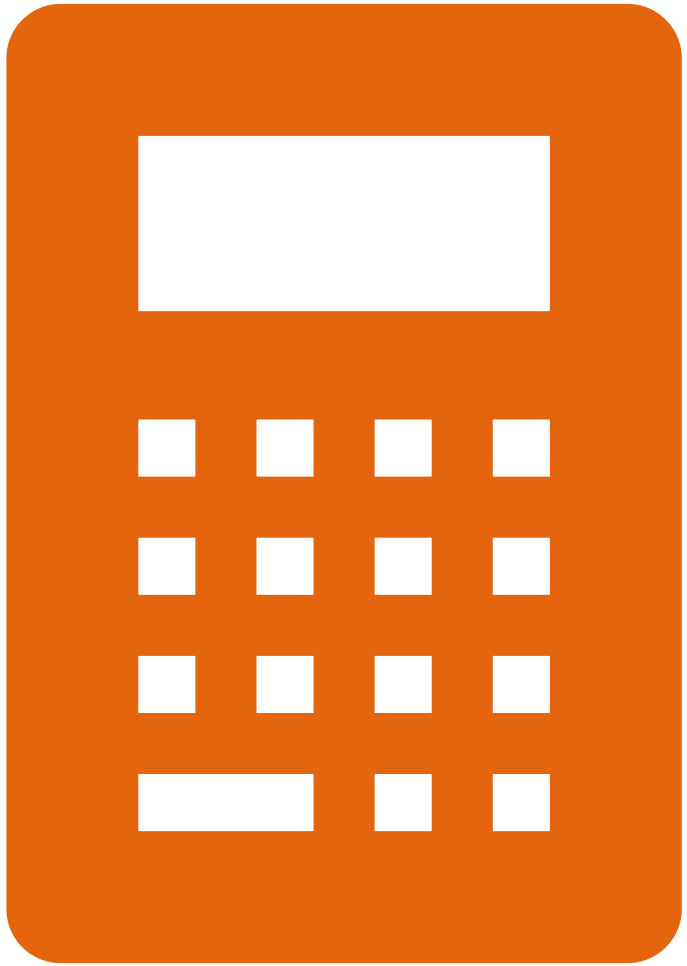
LEAST SQUARES REGRESSION METHOD

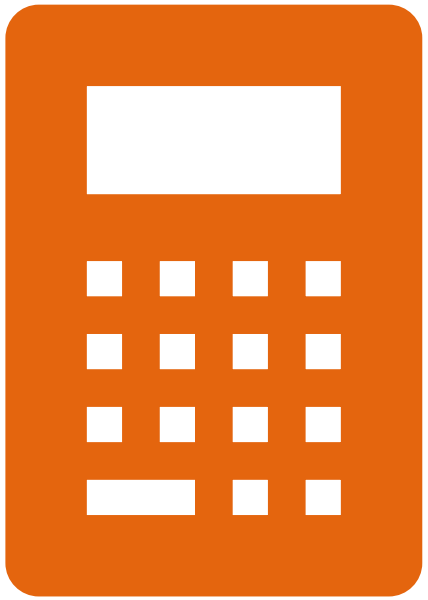
list of topics that will be covered in this session:

1. What Is the Least Squares Method?
2. Line Of Best Fit
3. Steps to Compute the Line Of Best Fit
4. The least-squares regression method with an example
5. A python program to implement Least Squares method

WHAT IS THE LEAST SQUARES 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.

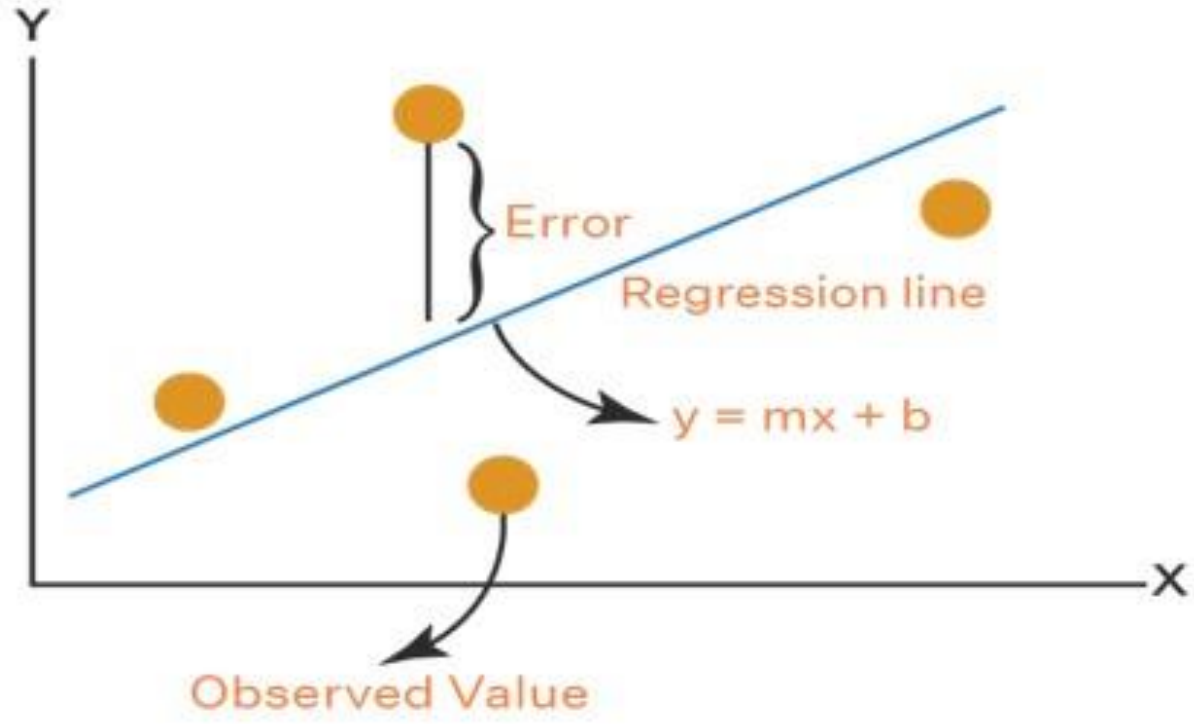




WHAT IS THE LEAST SQUARES METHOD

- The least-squares method is a statistical method used to find the line of best fit of the form of an equation such as y = mx + b to the given data.
- The curve of the equation is called the regression line. Our main objective in this method is to reduce the sum of the squares of errors as much as possible.
- This is the reason this method is called the least-squares method.
- This method is often used in data fitting where the best fit result is assumed to reduce the sum of squared errors that is considered to be the difference between the observed values and corresponding fitted value.
- The sum of squared errors helps in finding the variation in observed data. For example, we have 4 data points and using this method we arrive at the following graph.

LEAST-SQUARE METHOD GRAPH



WHAT IS THE LINE OF BEST FIT?

Line of best fit is drawn to represent the relationship between 2 or more variables.

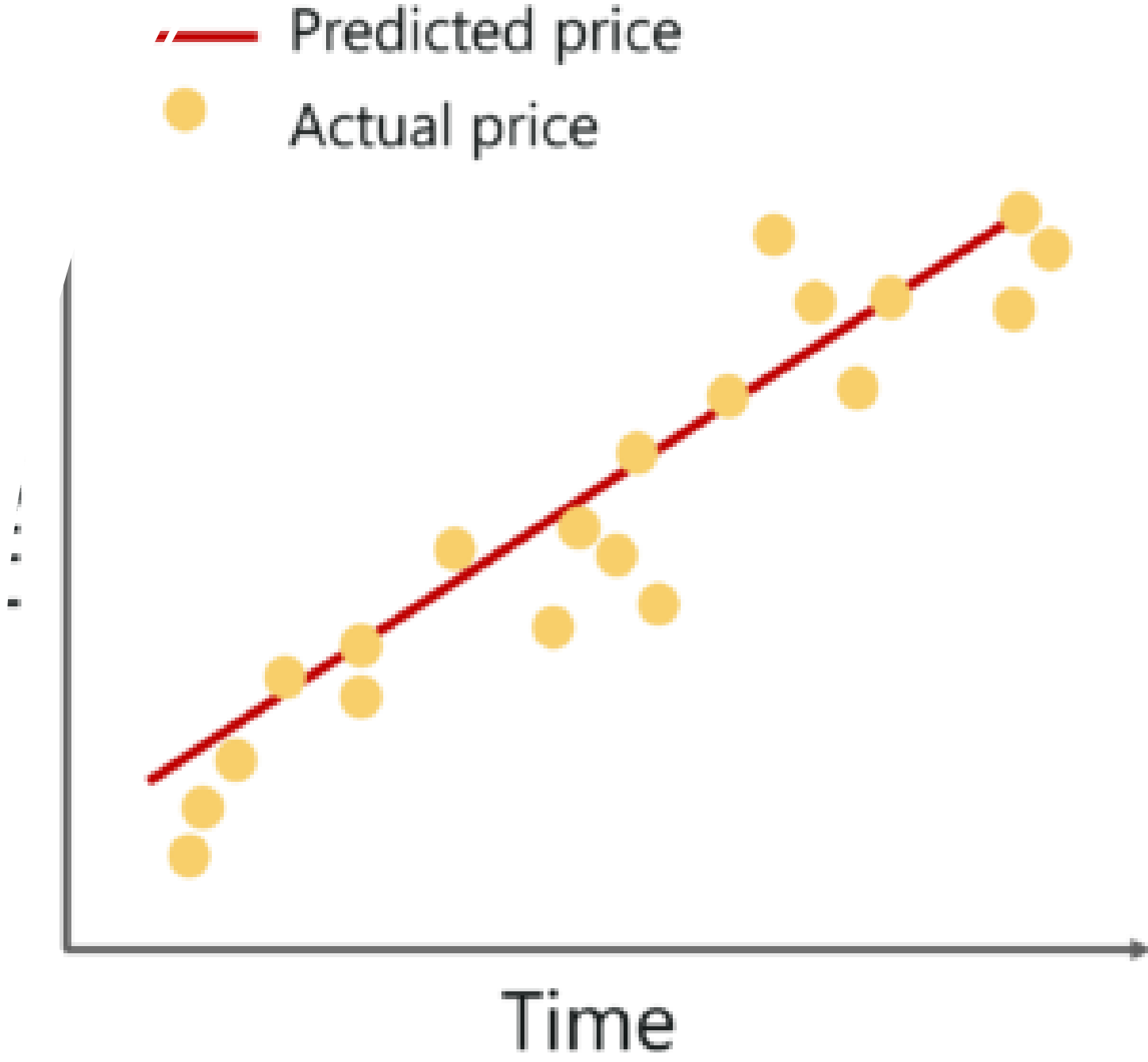
To be more specific, the best fit line is drawn across a scatter plot of data points in order to represent a relationship between those data points.

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.

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:

LINE OF BEST FIT

- 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.
- To better understand the whole process let’s see how to calculate the line using the Least Squares Regression.



LEAST SQUARE METHOD FORMULA

Least-square method is the curve that best fits a set of observations with a minimum sum of squared residuals or errors.

Let us assume that the given points of data are (x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n) in which all x's are independent variables, while all y's are dependent ones.

This method is used to find a linear line of the form $y = mx + c$, where y and x are variables, m is the slope, and c is the y-intercept.

The formula to calculate slope m and the value of c is given by:

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. Take a look at the equation below:

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: $y = mx + c$

y: dependent variable

m: the slope of the line

x: independent variable

c: y-intercept

STEPS TO CALCULATE THE LINE OF BEST FIT

- As an assumption, let's consider that there are 'n' data points.
- **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:*

$$y = mx + c$$

- Now let's look at an example and see how you can use the least-squares regression method to compute the line of best fit.

LEAST SQUARES REGRESSION 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

Let us use the concept of least squares regression to find the line of best fit for the above data.

LEAST SQUARES CALCULATION

- **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:** Compute the y-intercept value

$$c = y - mx$$

- After you substitute the respective values, c = 0.305 approximately.
- **Step 3:** Substitute the values in the final equation

$$y = mx + c$$

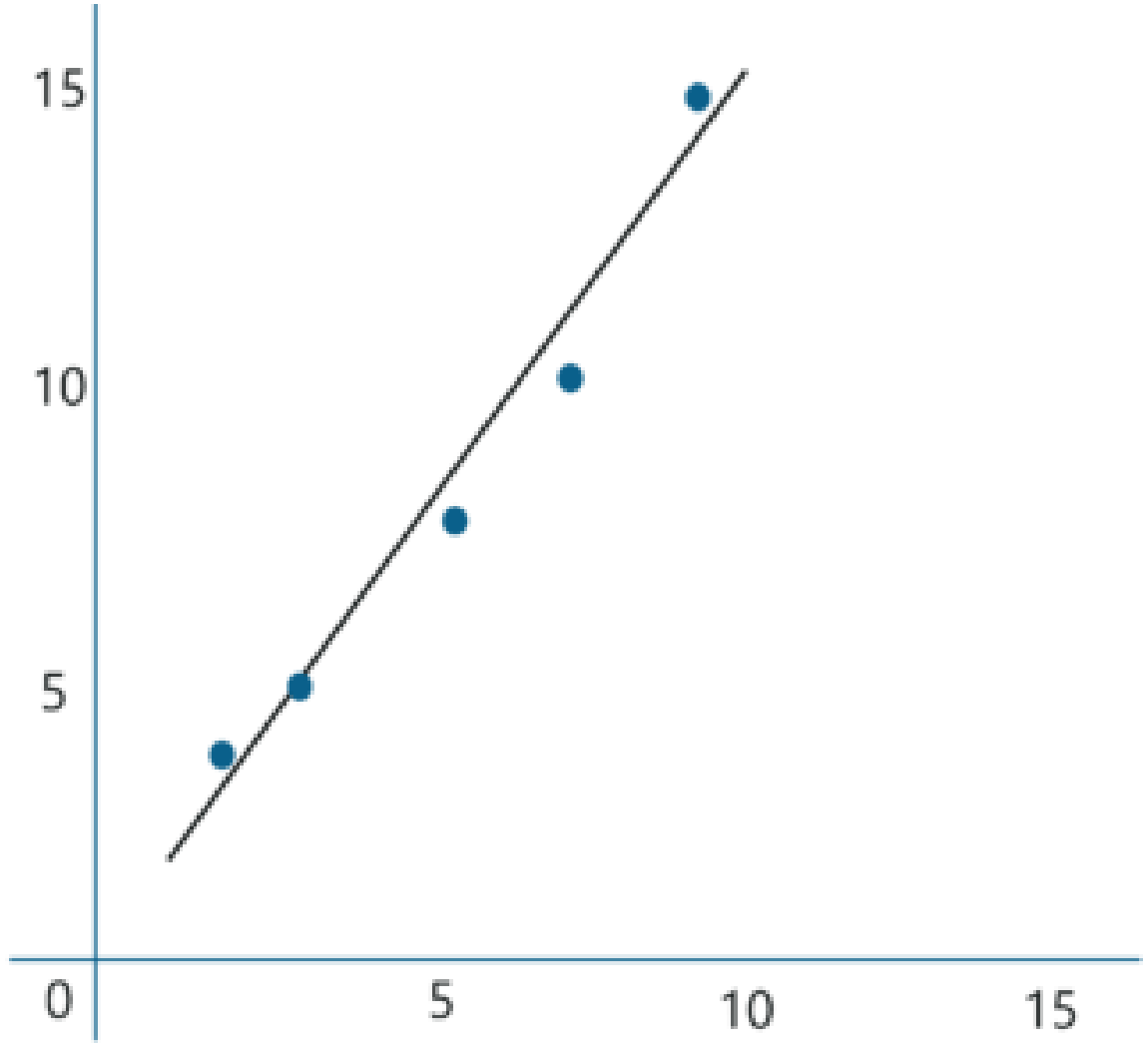
LEAST SQUARES CALCULATION

- 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

LEAST SQUARES GRAPH

- 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
- This comes down to 13 T-shirts! That's how simple it is to make predictions using Linear Regression.



LEAST SQUARES METHOD IN PYTHON

Problem Statement: To apply Linear square method and build a model that studies the relationship between the head size and the brain weight of an individual.

Data Set Description: The data set contains the following variables:

- Gender: Male or female represented as binary variables
- Age: Age of an individual
- Head size in cm^3: An individuals head size in cm^3
- Brain weight in grams: The weight of an individual's brain measured in grams

These variables need to be analyzed in order to build a model that studies the relationship between the head size and brain weight of an individual.

IMPLEMENTATION

- **Logic:** To implement Least Squares method in order to build a model that studies the relationship between an independent and dependent variable.
- The model will be evaluated by using least square regression method where RMSE and R-squared will be the model evaluation parameters.
- **Step 1: Import the required libraries**
- **Step 2: Import the data set**
- **Step 3: Assigning ‘X’ as independent variable and ‘Y’ as dependent variable**
- **Step 4: Calculate the values of the slope and y-intercept**
- **Step 5: Plotting the line of best fit**
- **Step 6: Model Evaluation**

THANK YOU