Linear Regression

Presented By:

Punit Kumar Johari

Assistant Professor

CSE & IT Department

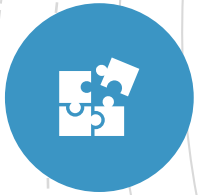
MITS, Gwalior

Content

Introduction to machine learning

Machine Learning Algorithms

Application of Linear Regression



Understanding Linear Regression

Multiple Linear Regression

Use Case – Profit Estimation of Companies

Introduction to Machine Learning

An Overview

Punit Kumar Johari

Basics

Independent and Dependent Variable

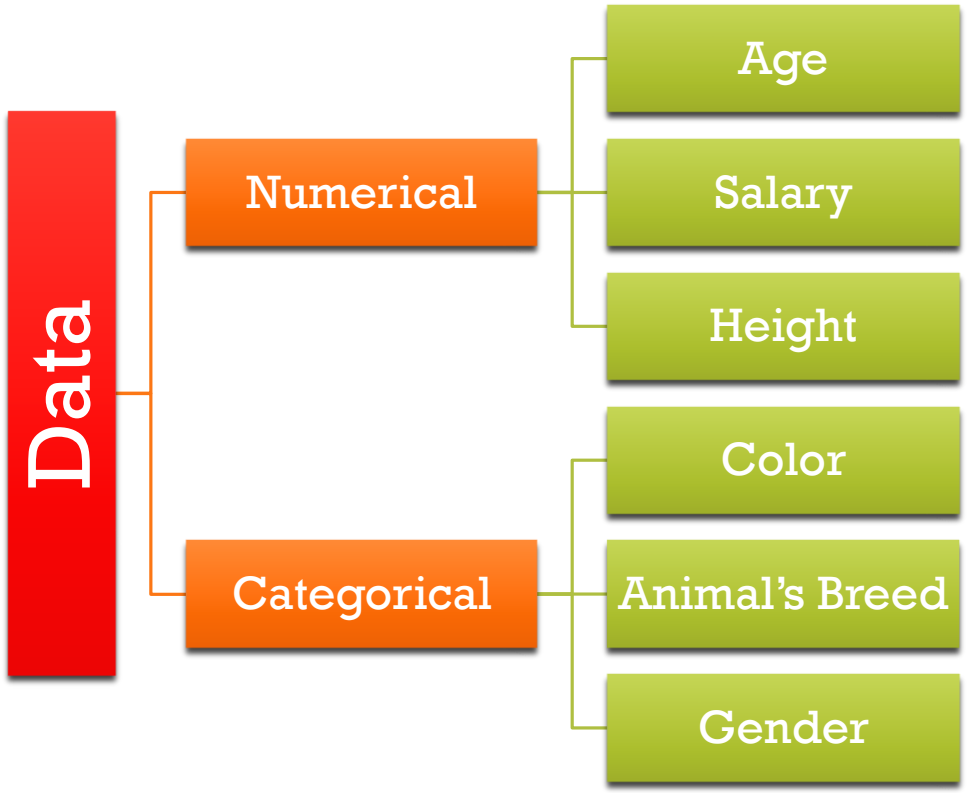
- Independent Variable - A variable whose value does not change by the effect of other variables and is used to manipulate the dependent variable. It is often denoted as X.

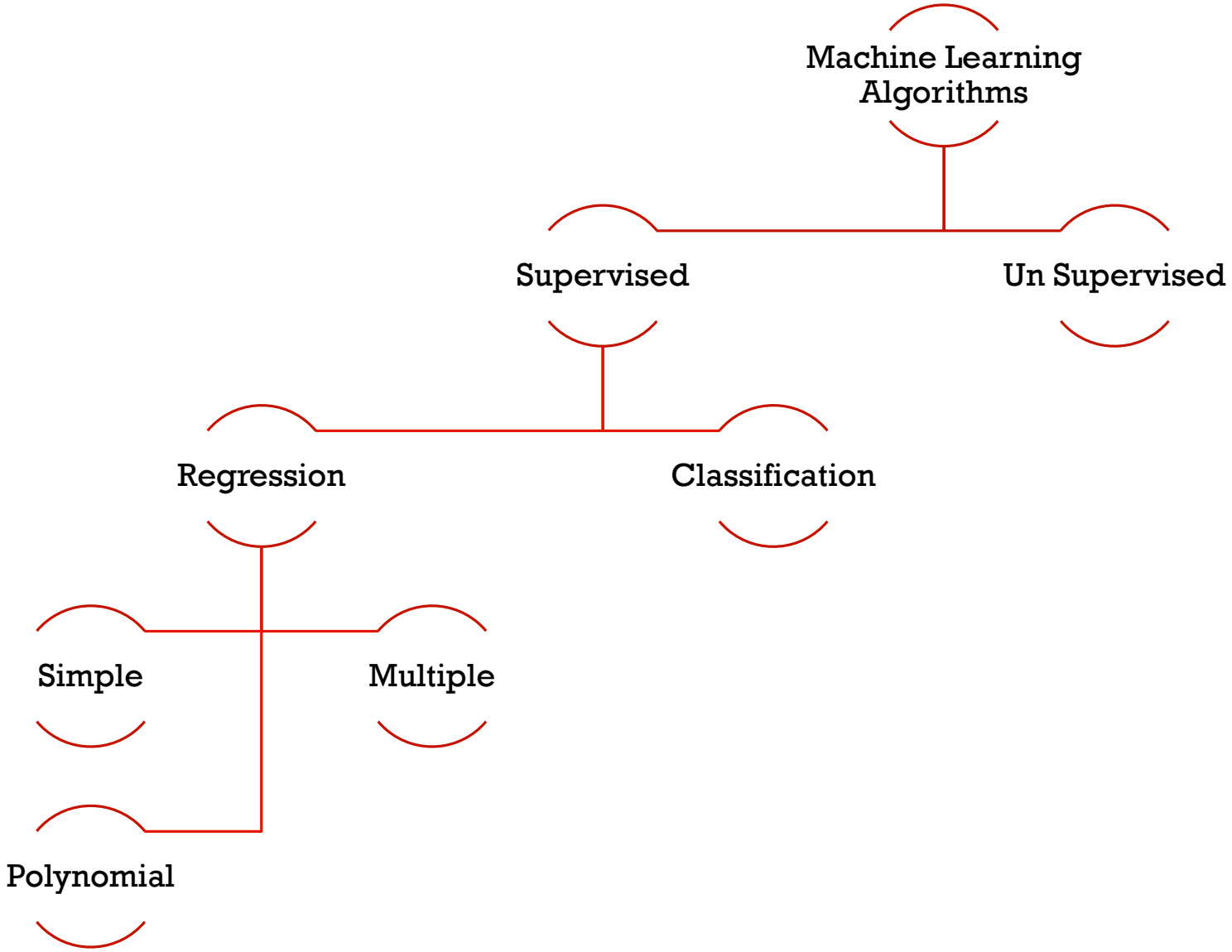
- Dependent variable - A variable whose value change when there is any manipulation in the values of independent variables. It is often denoted as Y.

Based on amount of rainfall, How much would be the crop yield?

Crop Yielded depends upon the amount of rainfall received

Numerical and Categorical values

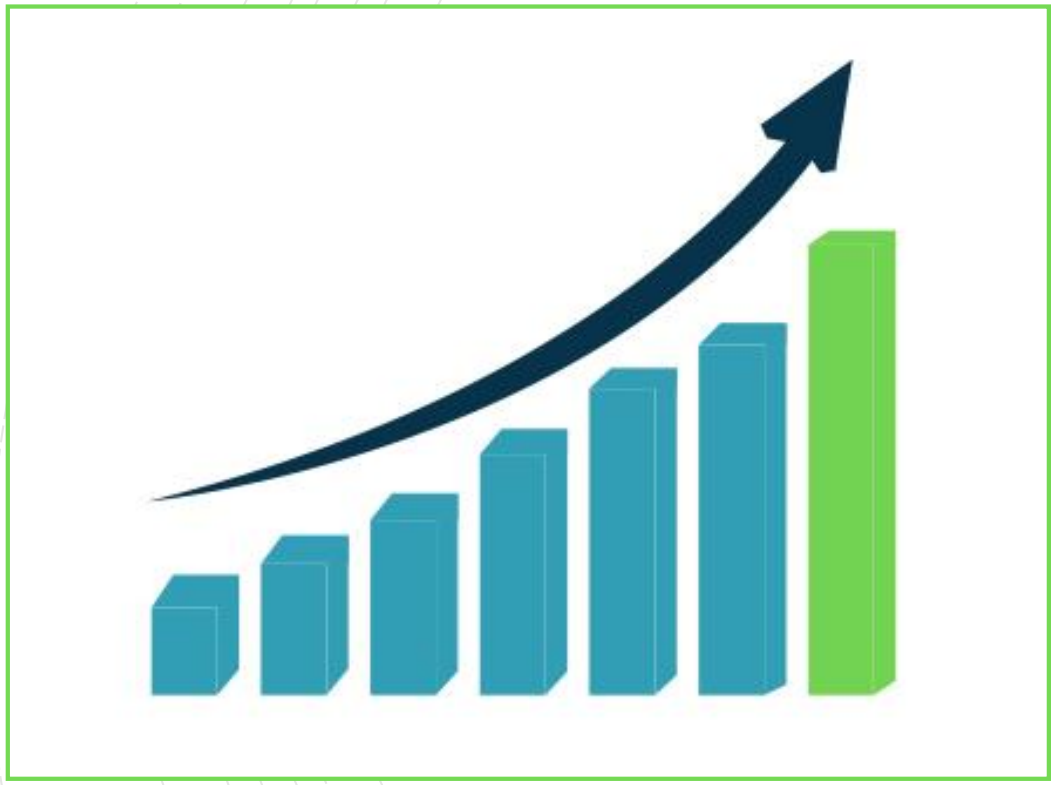



Machine Learning Algorithms
Supervised
Un Supervised
Regression
Classification
Simple
Multiple
Polynomial

Application of Linear Regression

Punit Kumar Johari

Economic Growth



- Used to determine the Economic growth of a country or a state in the coming quarter, can also be used to predict the GDP ofa country.

Product Price

- Can be used to predict what would be the price of the product in the future.

Housing Sales

- To estimate the number of houses a builder would sell and at what price in the coming months.

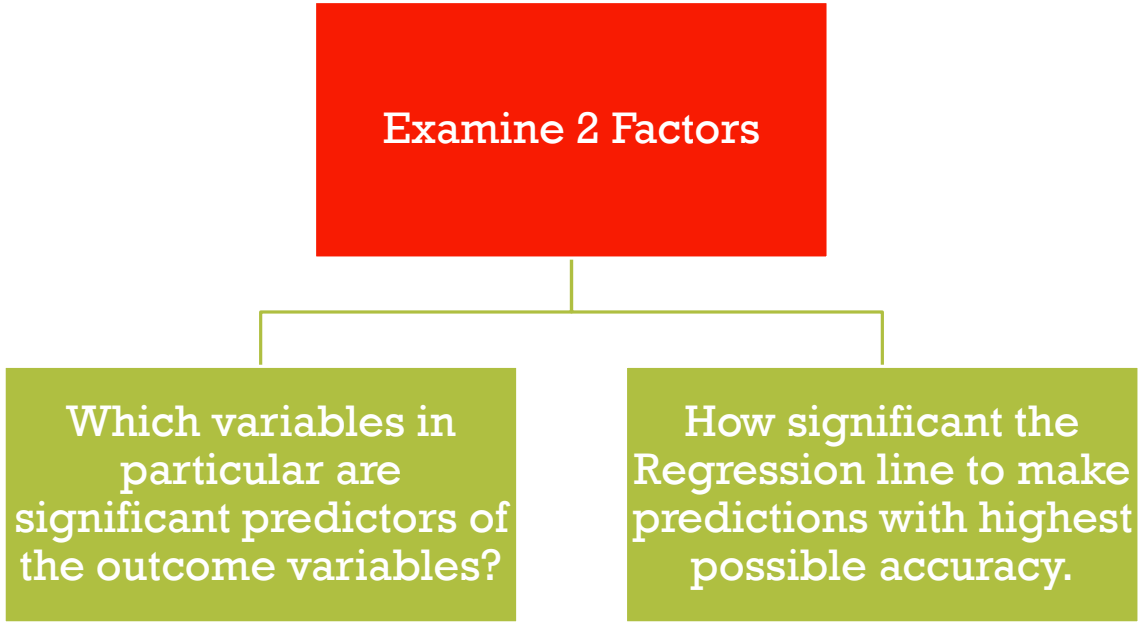
Score Prediction

- To predict the number of runs a player would score in the coming matches based on previous performance.

Understanding Linear Regression

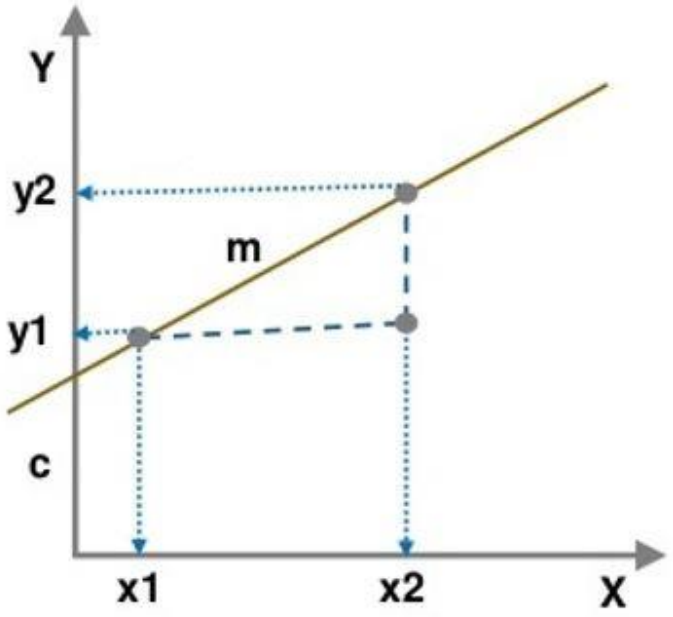
Punit Kumar Johari

Linear Regression is a statistical model used to predict the relationship between independent and dependent variables.



Regression Equation

The simplest form of a simple linear regression equation with one dependent and one independent variable is represented by:



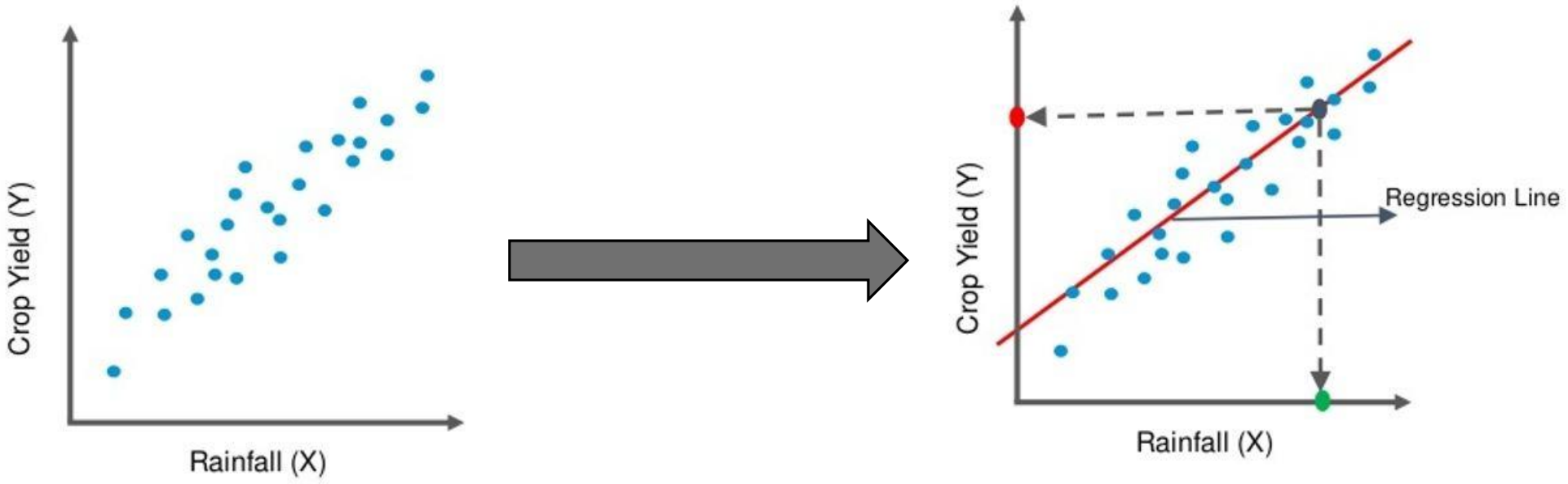
$$y = mx + c$$

- y → Dependent Variables
- x → Independent Variables
- m → Slope

$$\frac{y_2 - y_1}{x_2 - x_1}$$

- C → Coefficient of the Line

Prediction Using the Regression Line.



Plotting the amount of Crop Yield based on the amount of Rainfall

The Red point on the y axis is the amount of Crop Yield you can expect for some amount of Rainfall (X) represented by Green dot.

Intuition behind the regression Line

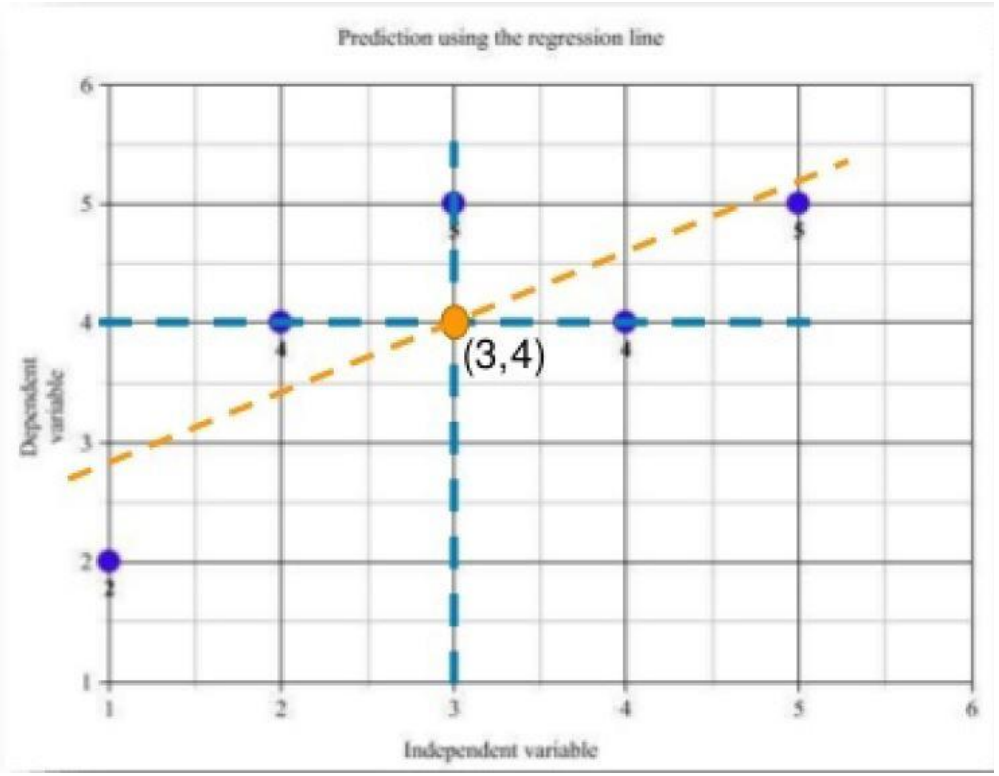
Regression line should ideally pass through the mean of X and Y.

Independent Variable X	Dependent Variable Y
1	2
2	4
3	5
4	4
5	5

Mean

3

4



Regression Line

Intuition behind the regression Line

Drawing the Equation of the Regression line

X	Y	$\dot{x}^2$	y^2	$x * y$
1	2	1	4	2
2	4	4	16	8
3	5	9	25	15
4	4	16	16	16
5	5	25	25	25
$\Sigma = 15$	$\Sigma = 20$	$\Sigma = 25$	$\Sigma = 86$	$\Sigma = 66$

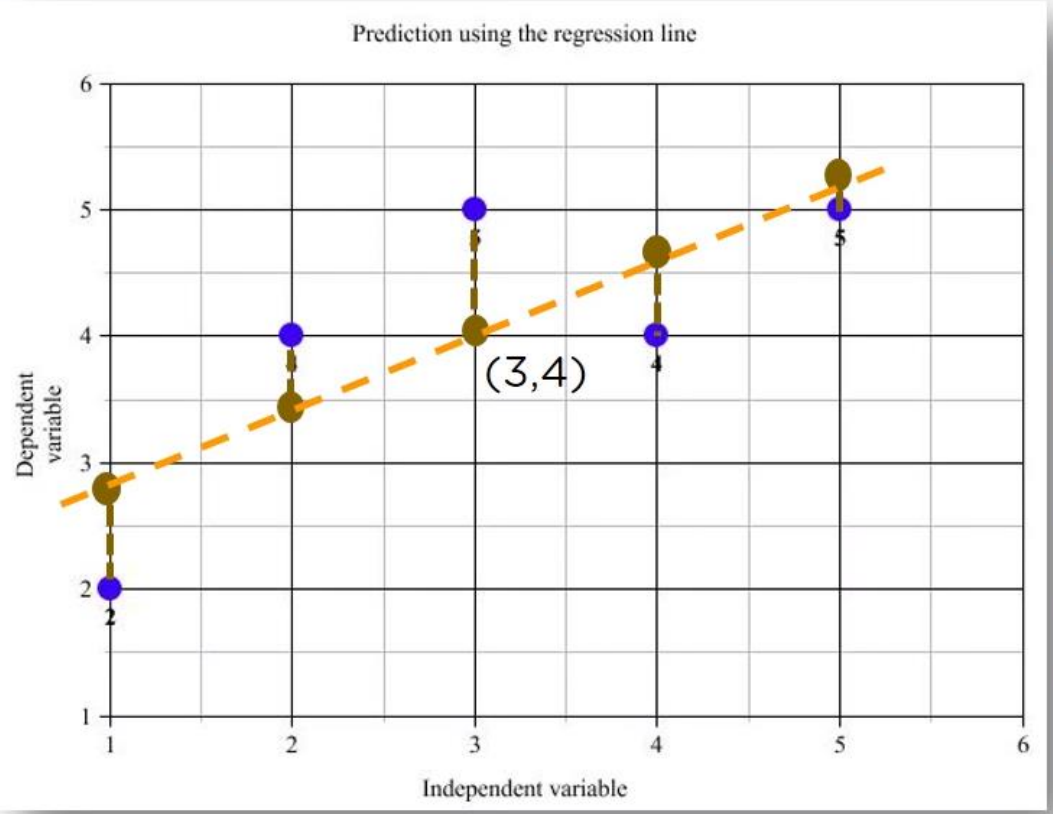
$$y = mx + c$$
$$y = 0.6 * 3 + 2.2$$
$$= 4$$

$$m = \frac{(n * \sum xy) - (\sum x * \sum y)}{(n * \sum x^2) - (\sum x)^2} = \frac{(5 * 66) - (15 * 20)}{(5 * 55) - 225} = 0.6$$

$$c = \frac{(\sum y * \sum x^2) - (\sum x * \sum xy)}{(n * \sum x^2) - (\sum x)^2} = 2.2$$

Intuition behind the regression Line

Lets find out the predicted values of Y for corresponding values of X using the linear equation where m-0.6 and c=2.2

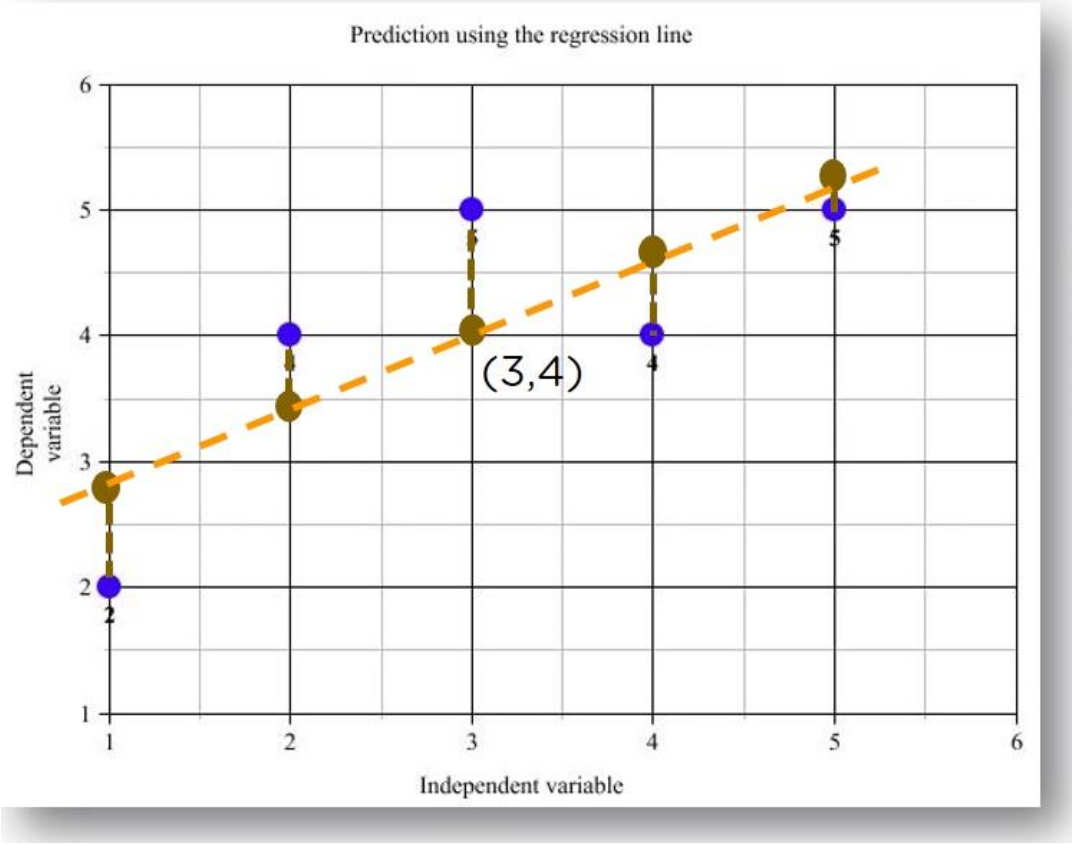


Y_{pred}
$Y=0.6*1+2.2=2.8$
$Y=0.6*2+2.2=3.4$
$Y=0.6*3+2.2=4$
$Y=0.6*4+2.2=4.6$
$Y=0.6*5+2.2=5.2$

Here the blue points represent the **actual Y values** and the brown points represent the **predicted Y values**. The distance between the actual and predicted values are known as *residuals or errors*. The best fit line should have the least sum of squares of these errors also known as *e square.*

Intuition behind the regression Line

Lets find out the predicted values of Y for corresponding values of X using the linear equation where m=0.6 and c=2.2



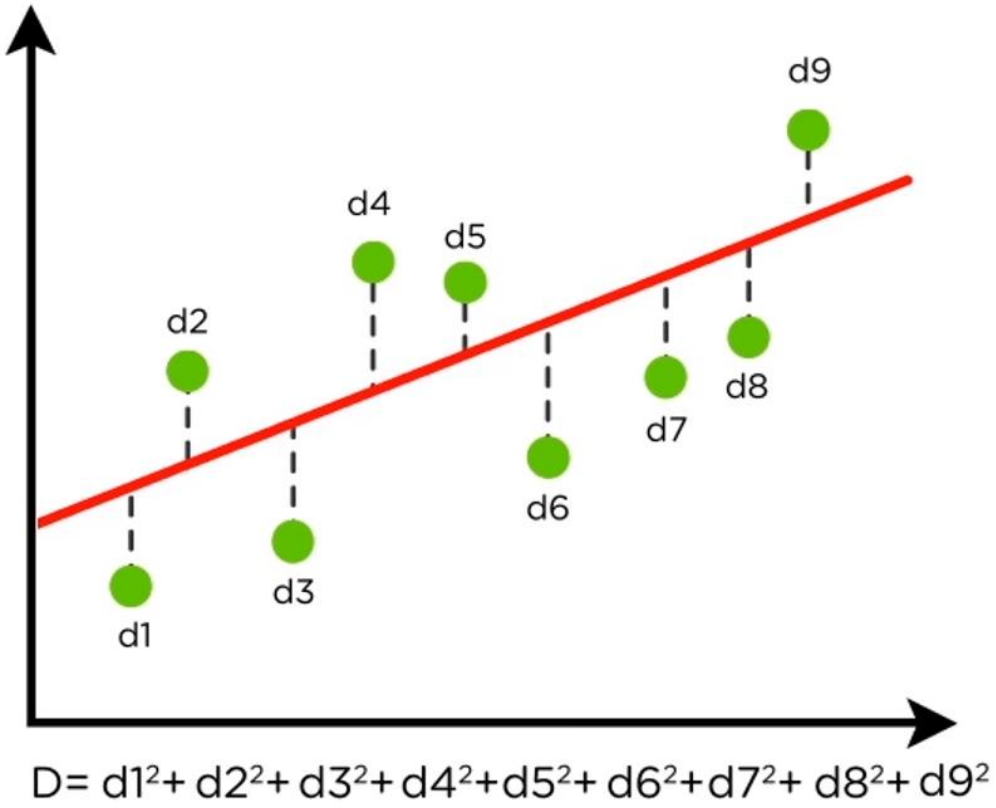
X	Y	Y_{pred}	$(Y-Y_{pred})$	$(Y-Y_{pred})^2$
1	2	2.8	-0.8	0.64
2	4	3.4	0.6	0.36
3	5	4	1	1
4	4	4.6	-0.6	0.36
5	5	5.2	-0.2	0.04

$\sum = 2.4$

The sum of squared errors for this regression line is 2.4. We check this error for each line and conclude the best fit line having the least e square value.

Finding the Best Fit Line

Minimizing the Distance: There are lots of ways to minimize the distance between the line and the data points like Sum of Squared errors, Sum of Absolute errors, Root Mean Square error etc.

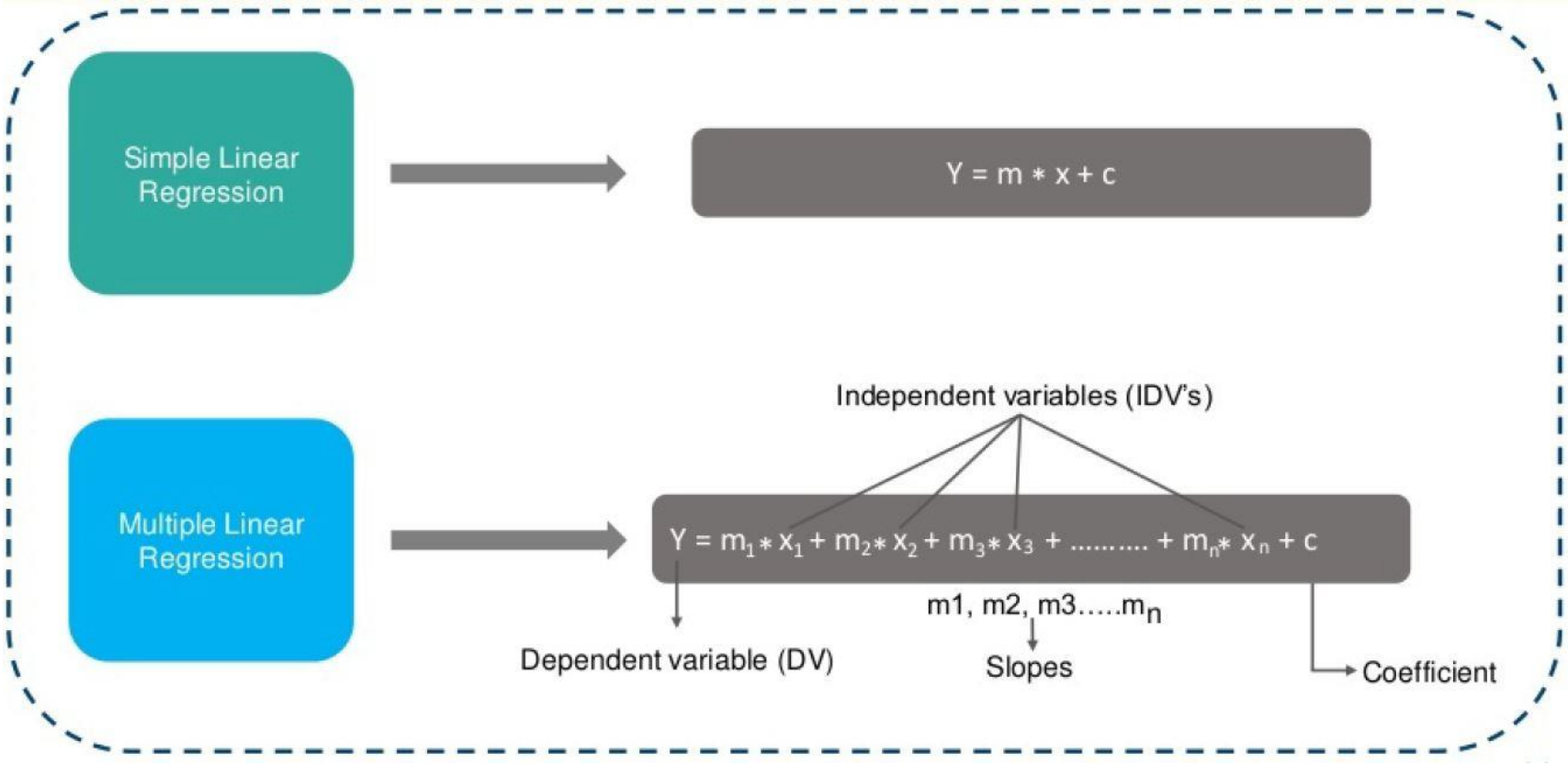


We keep moving this line through the data points to make sure the Best fit line has the least square distance between the data points and the regression line

Multiple Linear Regression

Punit Kumar Johari

Multiple Linear Regression



Implementation of Linear Regression

- **Use Case**

Let's Understand how the Linear Regression Works by implementing it in Python

- For instance we will use Python 3.7 Version for Windows platform.

You may use any of the API depending on the configuration compatible with your system

JUPYTER NOTEBOOK

GOOGLE COLAB

Presented By:

Punit Kumar Johari

Assistant Professor

CSE & IT Department

MITS, Gwalior

Email: pkjohari@mitsgwalior.in