Python for Beginners

Archer Infotech , PUNE

Python – Regression

Regression Analysis

- **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.

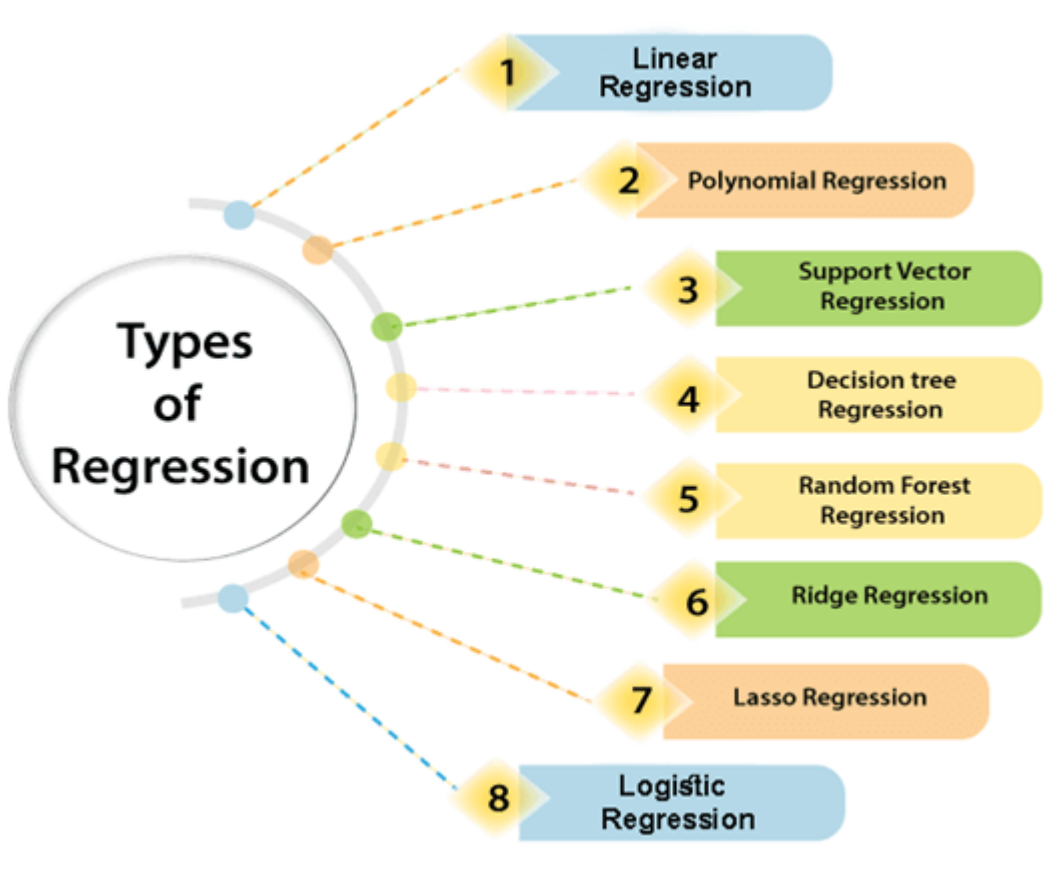
- It predicts continuous/real values such as **temperature, age, salary, price,** etc.

Regression Analysis

- Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Types of Regressions

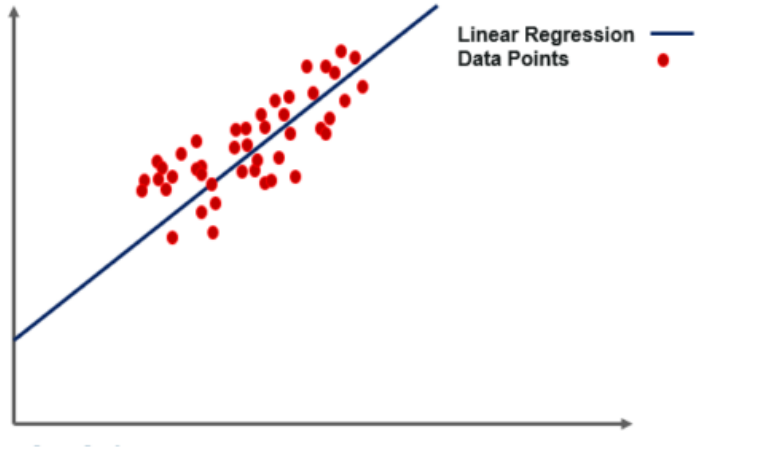


Single Linear Regression

- Linear regression is a statistical regression method which is used for predictive analysis
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression

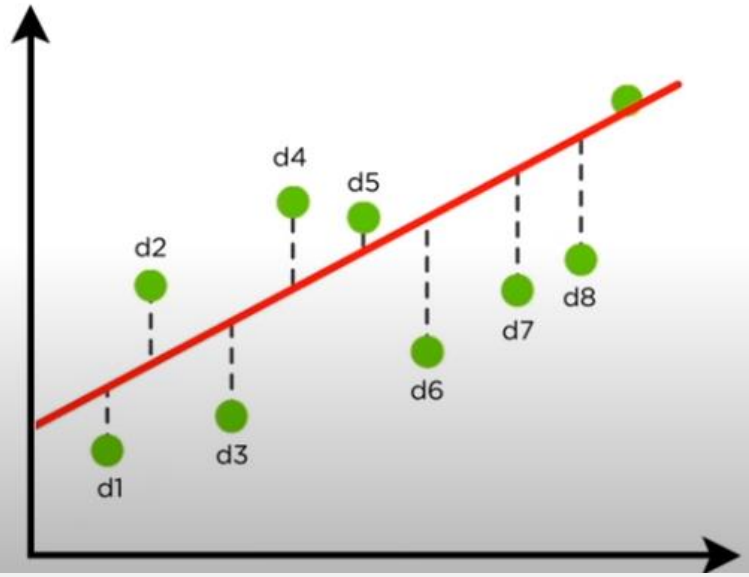
The straight line in the diagram is **the best fit line**.

The main goal of the simple linear regression is to consider the given data points and plot the best fit line to fit the model in the best way possible.



Regression Equation – 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.

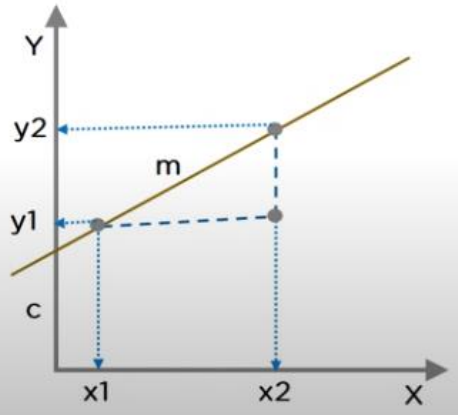


We keep moving this line through the data points to make sure the Best fit line has the least square distance between the data points and the regression line

Regression Equation – Cost Function

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

$$y = m * x + c$$



y ---> Dependent Variable

x ---> Independent Variable

m ---> Slope of the line

$$m = \frac{y2 - y1}{x2 - x1}$$

c ---> Coefficient of the line

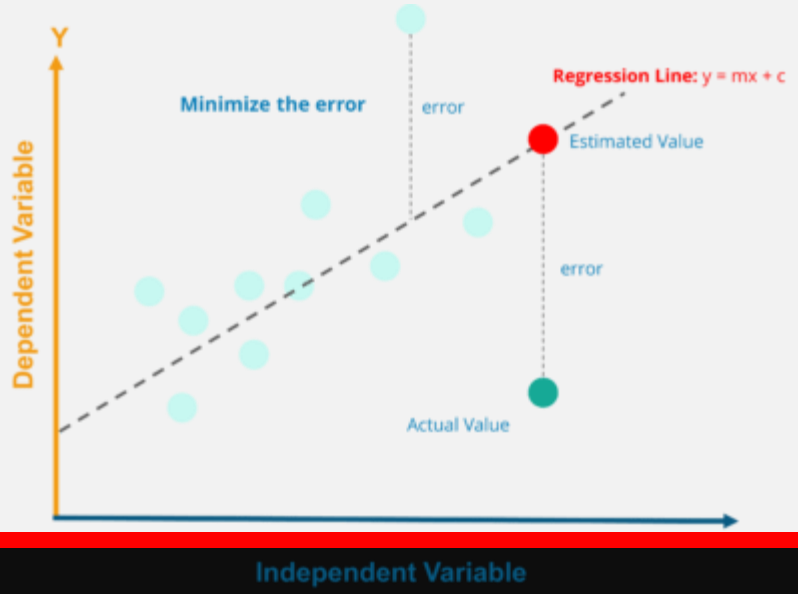
Least Square Method – Finding the best fit line

Least squares is a statistical method used to determine the best fit line or the regression line by minimizing the sum of squares created by a mathematical function. The “square” here refers to squaring the distance between a data point and the regression line. The line with the minimum value of the sum of square is the best-fit regression line.

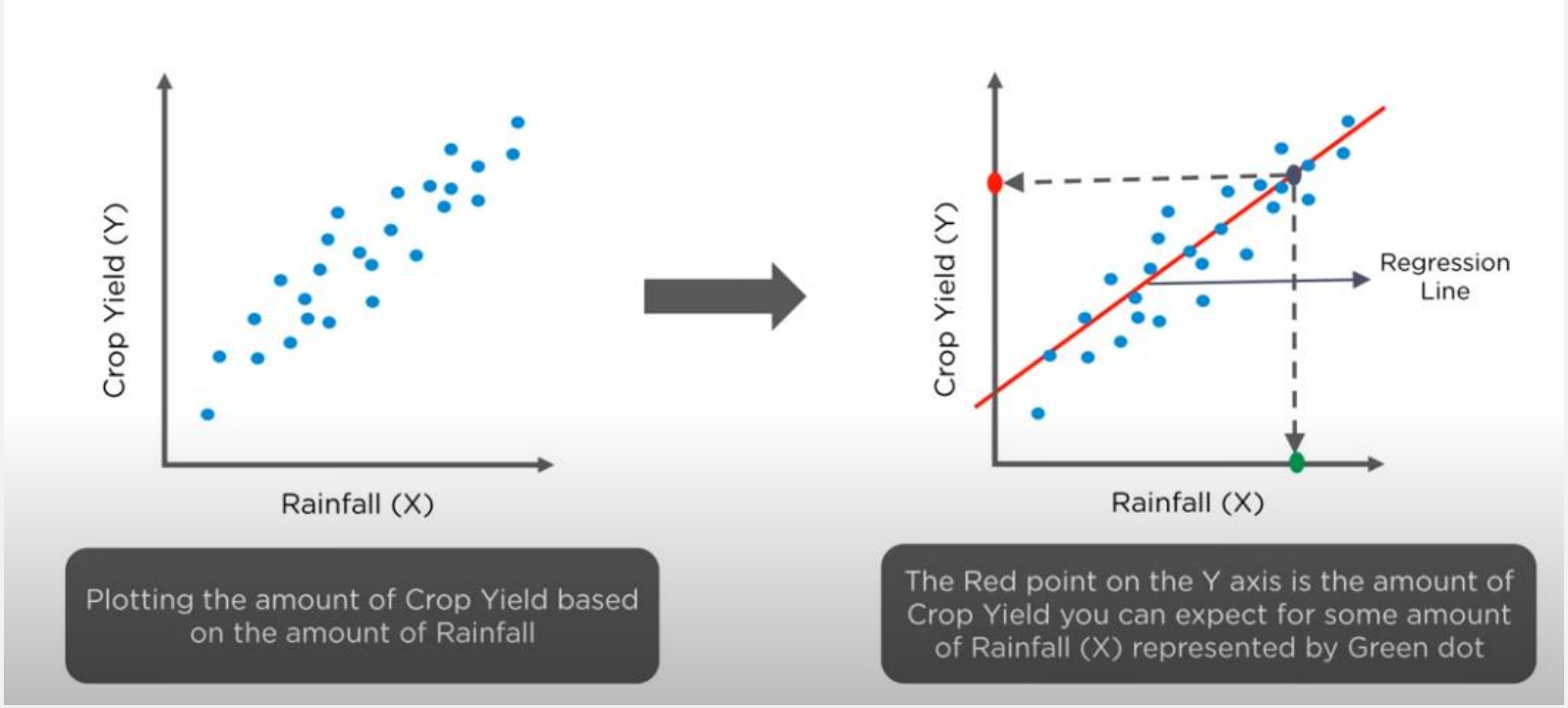
Regression Line, y = mx+c where,

$$m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

y = Dependent Variable
x= Independent Variable
c = y-Intercept



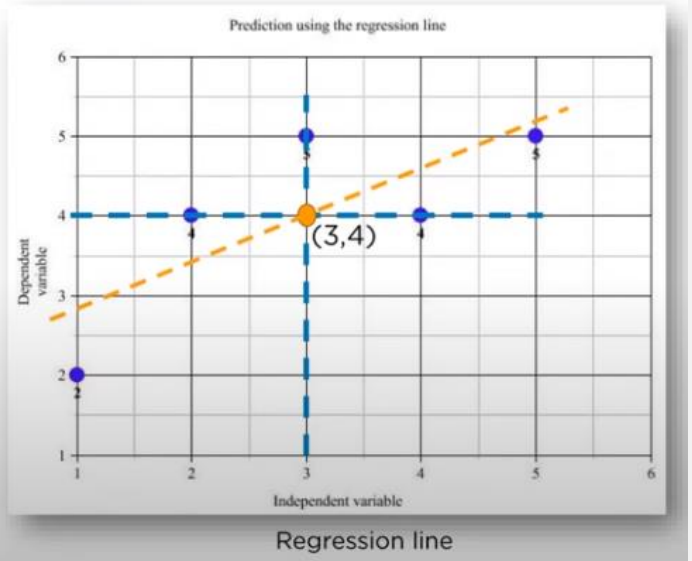
Prediction Using Regression Line



Regression Equation

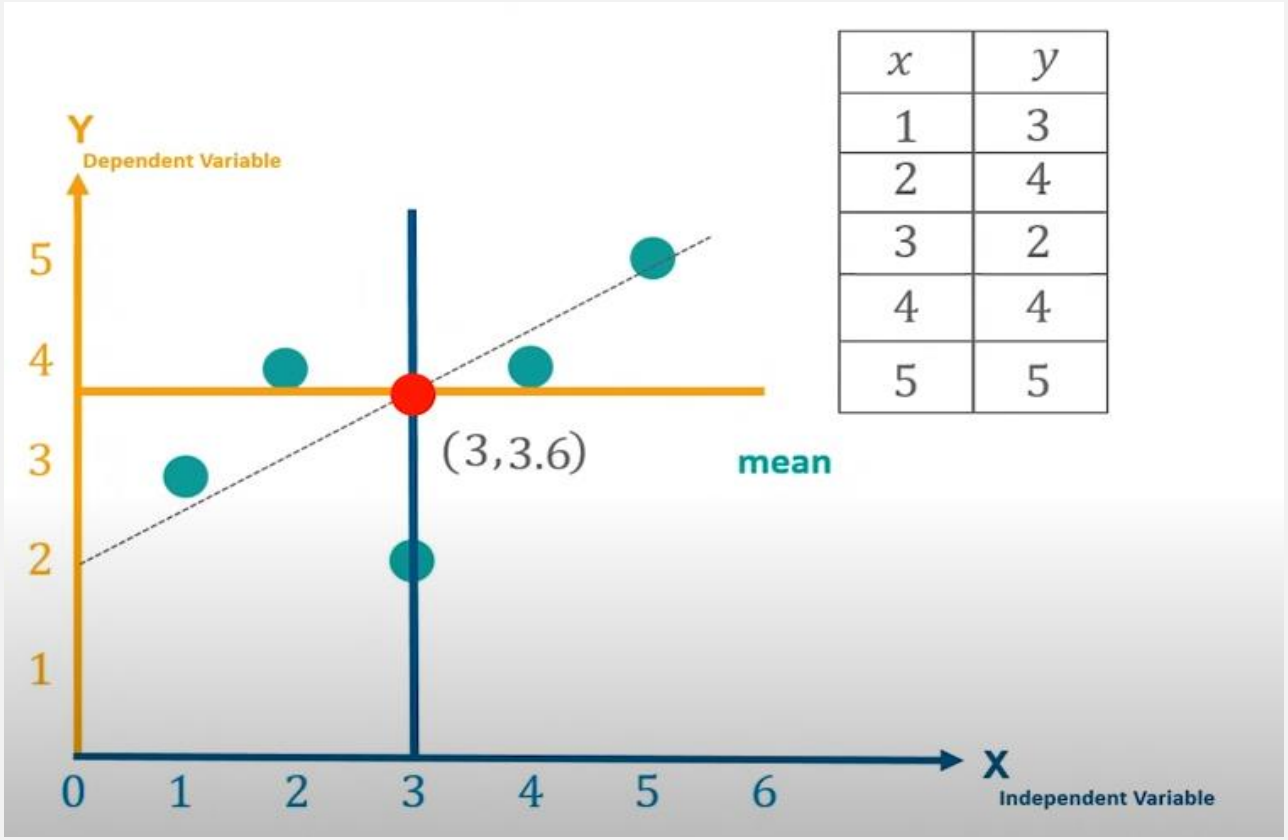
Regression line should ideally pass through the mean of X and Y

Independent variable	Dependent variable
X	Y
1	2
2	4
3	5
4	4
5	5
Mean	3	4



Regression line

Regression Equation

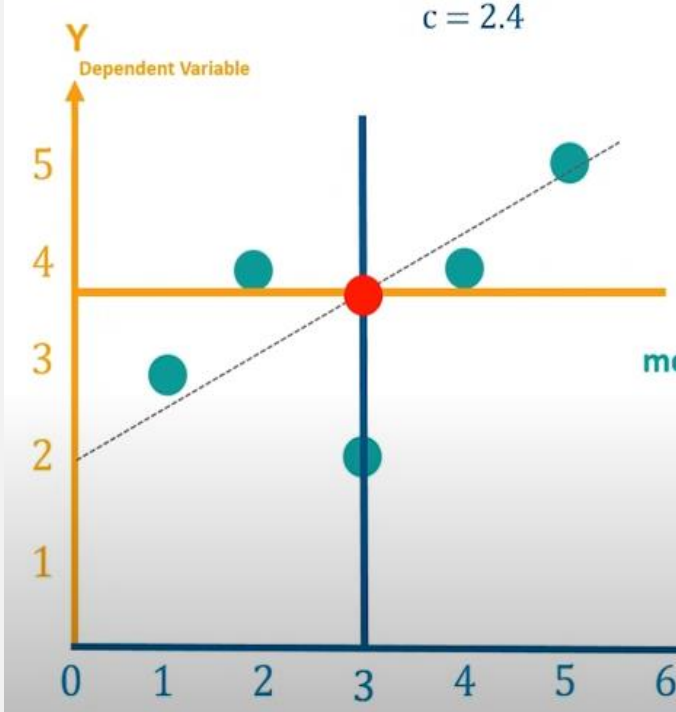


x	y
1	3
2	4
3	2
4	4
5	5

Regression Equation

$y = mx+c$

$c = 2.4$



x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
mean	3	3.6		$\Sigma = 10$	$\Sigma = 4$

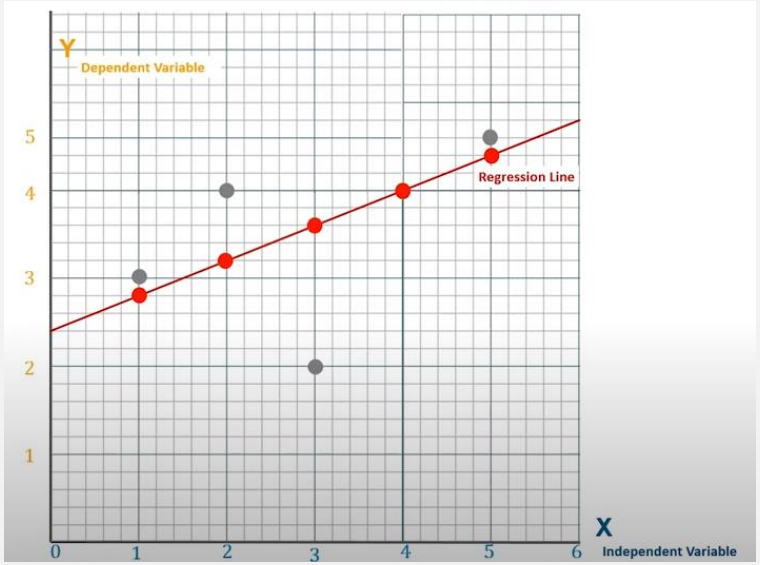
$$m = \sum \frac{(x - \bar{x})(y - \bar{y})}{(x - \bar{x})^2} = \frac{4}{10}$$

$m = 0.4$

$c = 2.4$

$y = 0.4x + 2.4$

Regression Equation

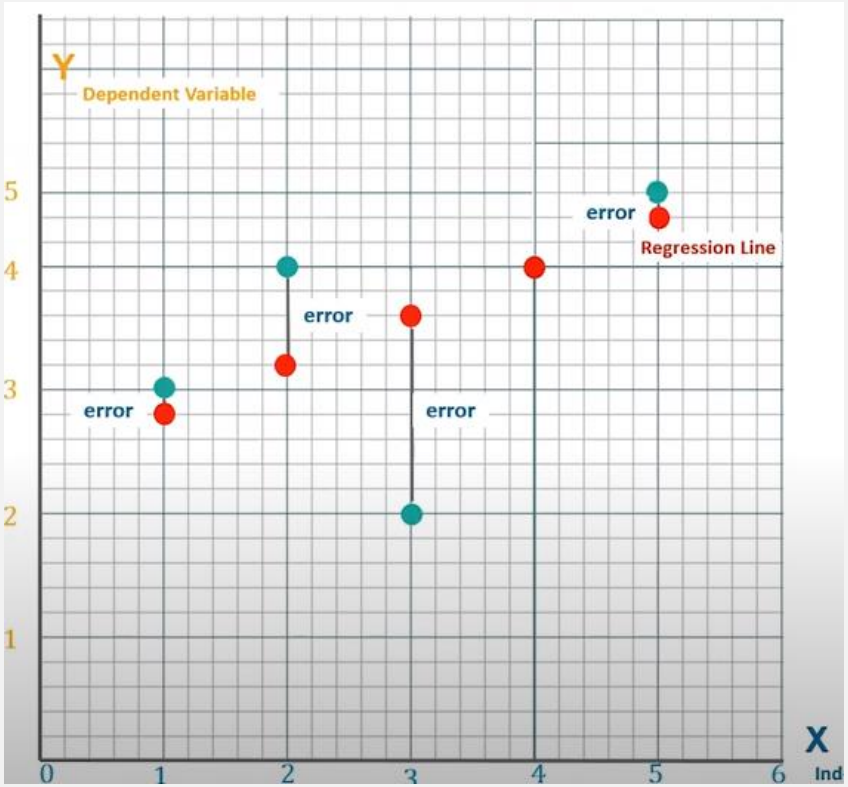


$$m = 0.4$$
$$c = 2.4$$
$$y = 0.4x + 2.4$$

For given $m = 0.4$ & $c = 2.4$, lets predict values for y for $x = \{1,2,3,4,5\}$

$$y = 0.4 \times 1 + 2.4 = 2.8$$
$$y = 0.4 \times 2 + 2.4 = 3.2$$
$$y = 0.4 \times 3 + 2.4 = 3.6$$
$$y = 0.4 \times 4 + 2.4 = 4.0$$
$$y = 0.4 \times 5 + 2.4 = 4.4$$

Regression Equation



Distance between actual & predicted value

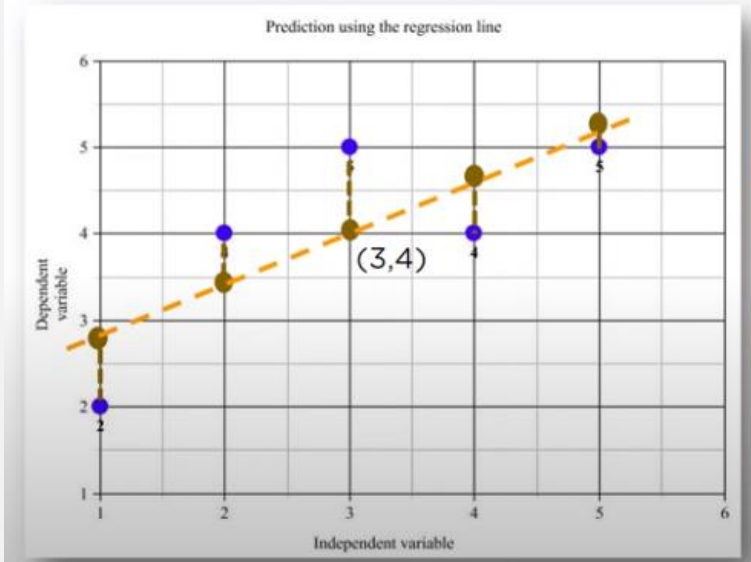
Regression Equation – Mean Square Error

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Regression Equation

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*.

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Model Performance

R Squared Concept and Formula

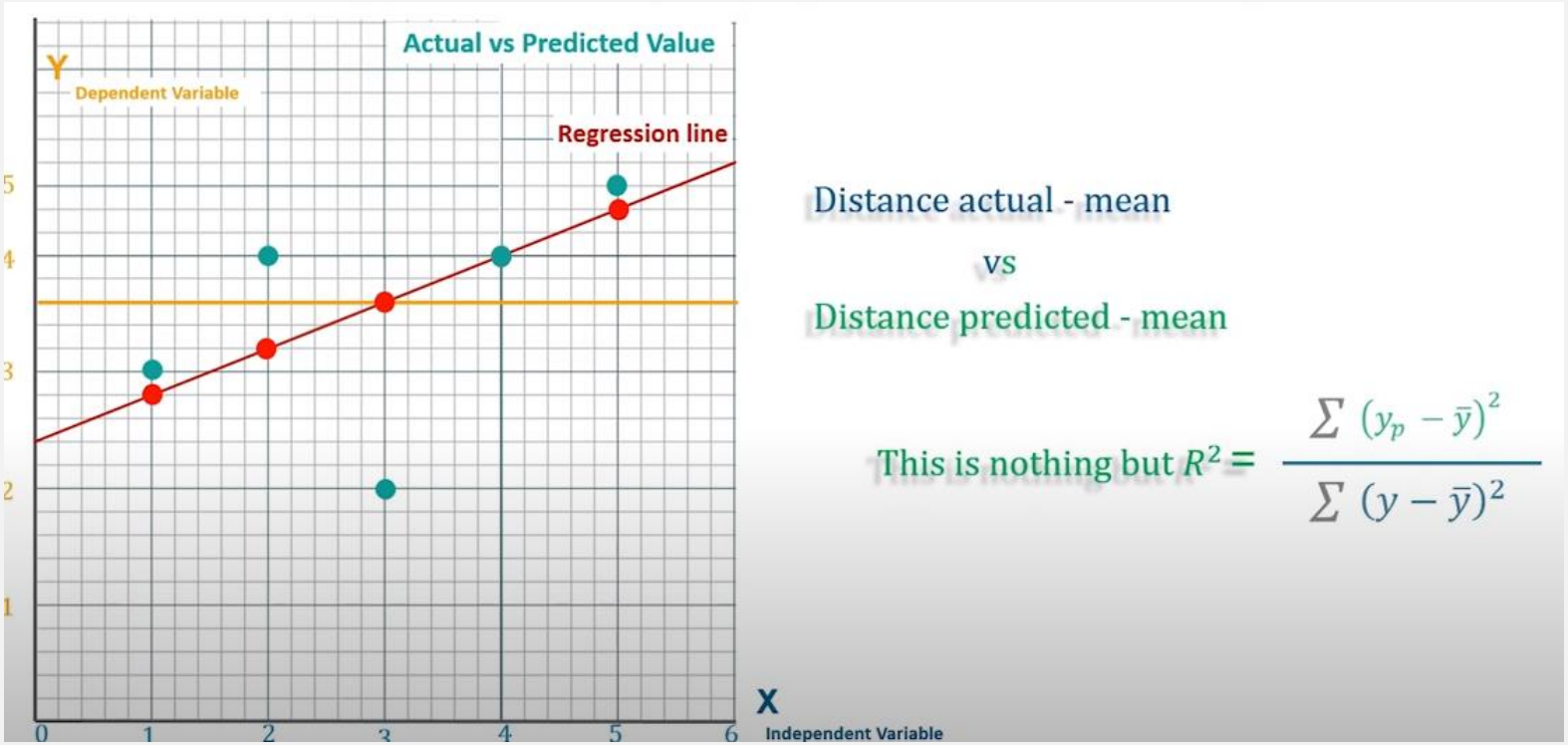
R-Squared is also known as the **Coefficient of Determination**.

The value of R-Squared ranges from **0 to 1**. The higher the R-Squared value of a model, the better is the model fitting on the data.

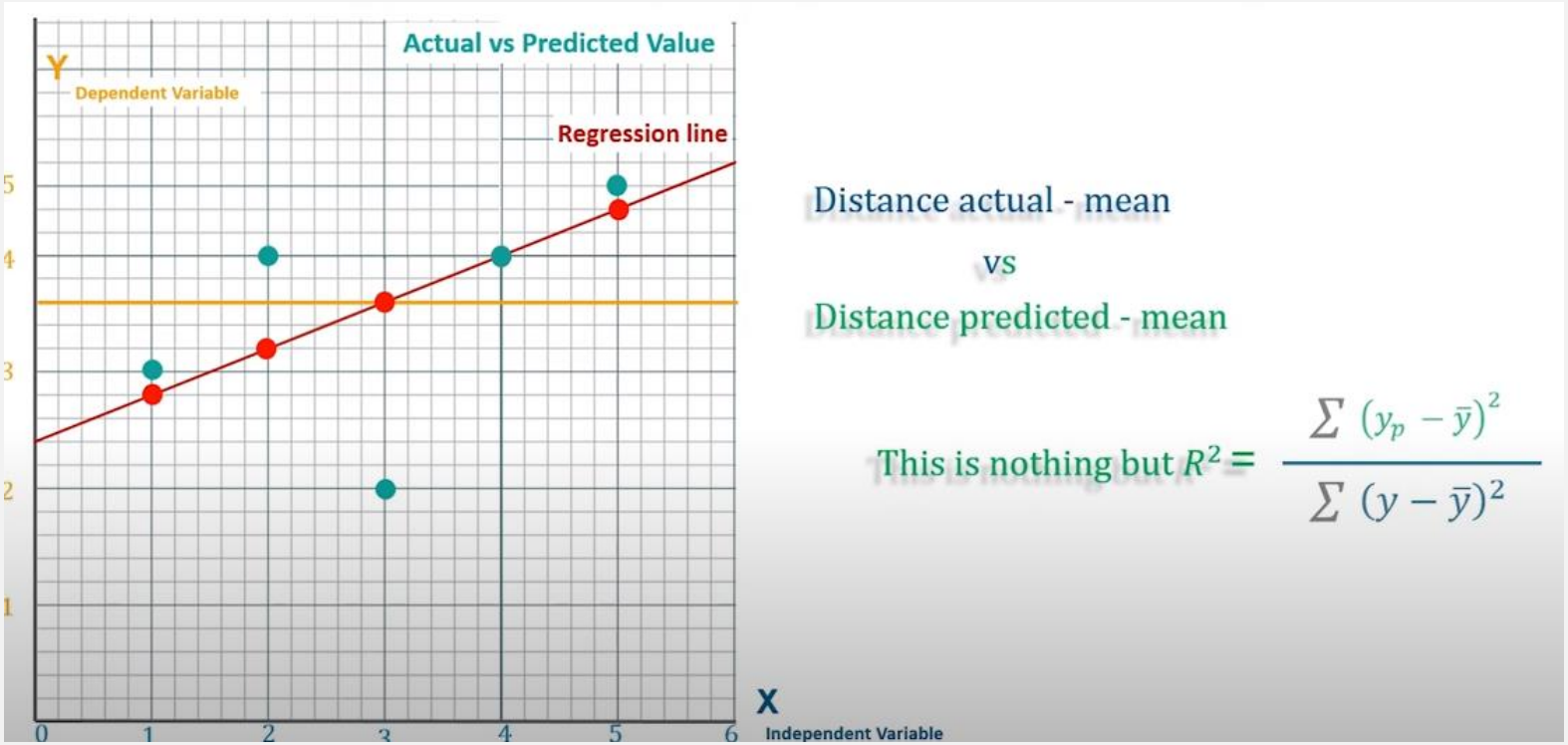
However, if the R-Squared value is very close to 1, then there is a possibility of model overfitting, which should be avoided.

A good model should have an R-Squared above 0.8.

Calculation



Calculation

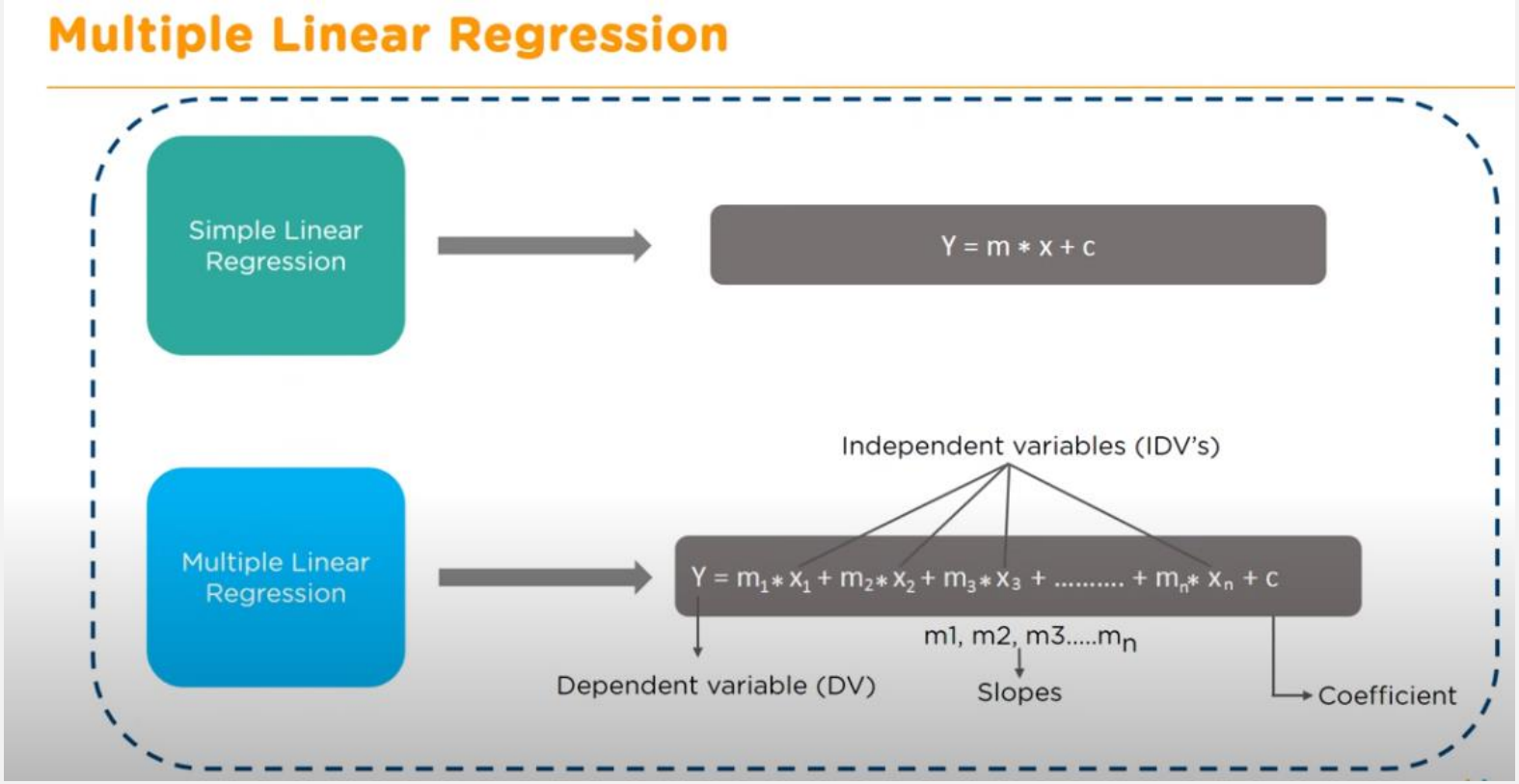


Calculation

x	y	$y - \dot{y}$	$(y - \dot{y})^2$	y_p	$(y_p - \dot{y})$	$(y_p - \dot{y})^2$
1	3	– 0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	–1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64
mean y	3.6		$\sum$ 5.2			$\sum$ 1.6

$$R^2 = \frac{1.6}{5.2} = \frac{\sum \left((y_p - \bar{y}\right)^2}{\sum (y - \dot{y})^2}$$

Multiple Linear Regression



THANK YOU !!!

Amol Patil - 9822291613