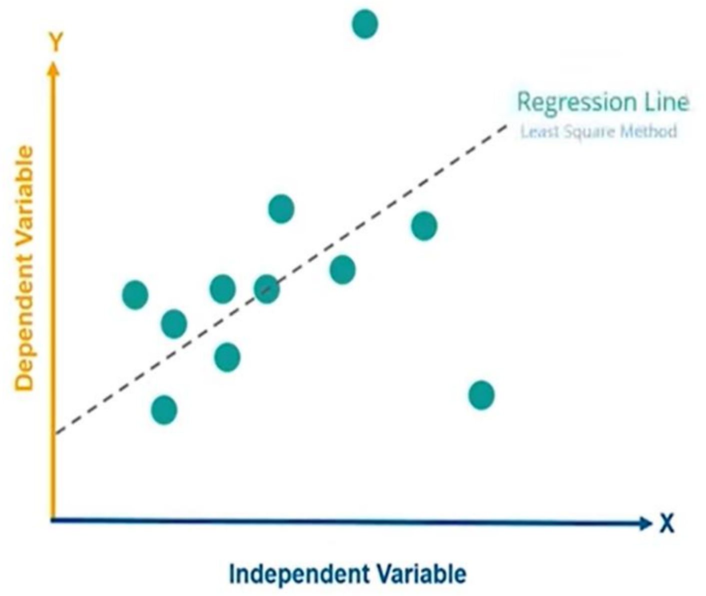
UNIT – II

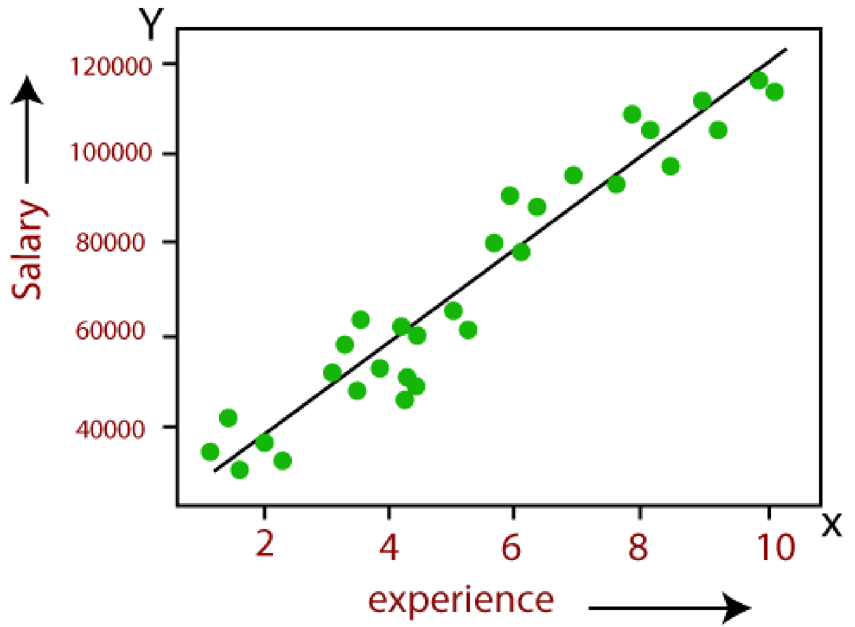
Supervised Learning: Regression

Supervised Learning: Regression

Regression

- Regression is a method for understanding the **relationship** between **independent variables/ features** and a **dependent variable/outcome.**
- Regression is a supervised machine learning technique which is used to **predict continuous values.**
- A regression problem is when the **output variable** is **a real or continuous values** such as **temperature, age, salary, price**, etc.



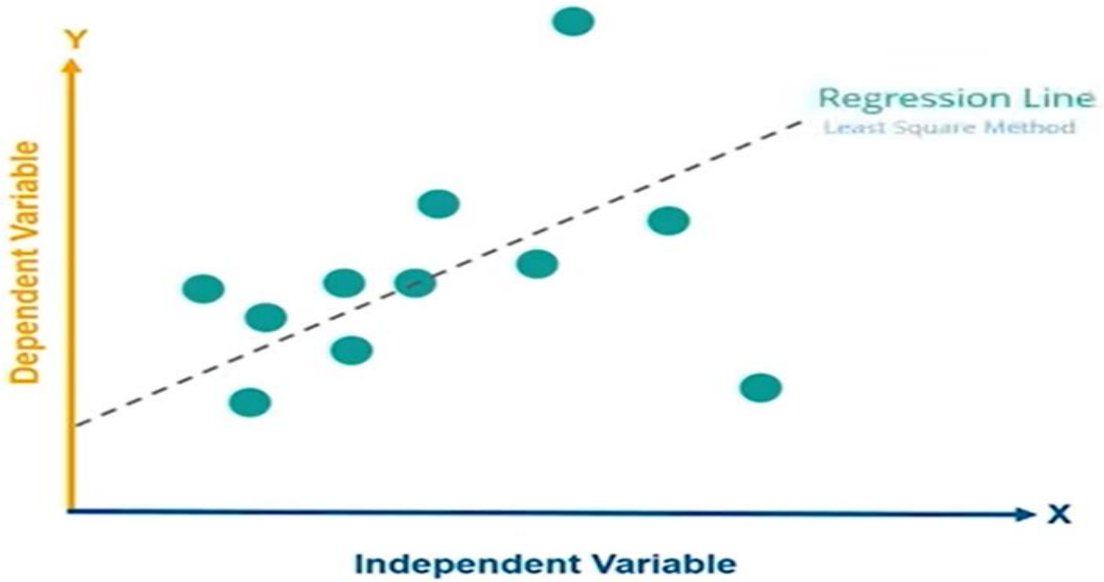


The most common regression algorithms are:

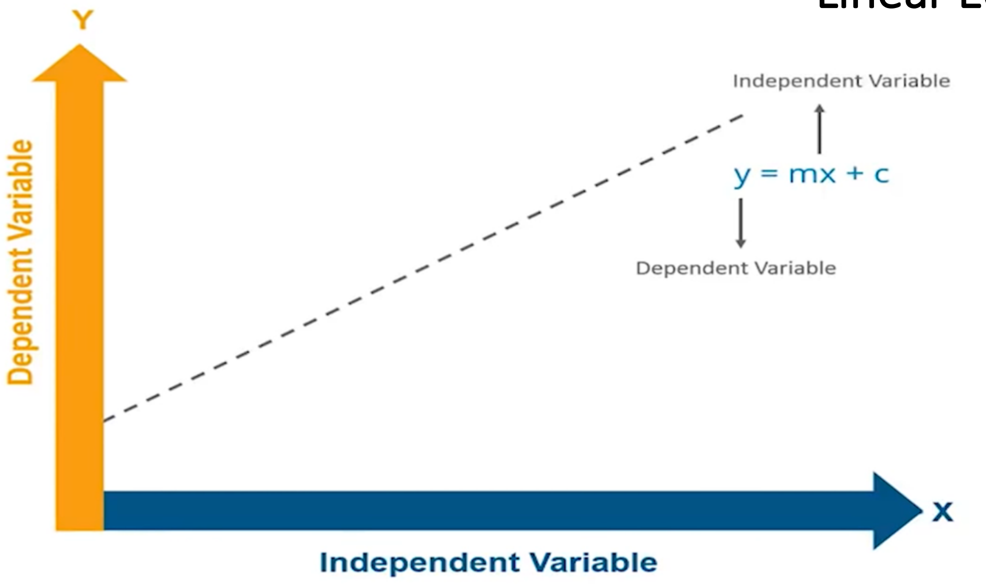
- Simple linear regression
- Multiple linear regression
- Polynomial regression
- Logistic regression
- Maximum likelihood estimation (least squares)

Simple Linear Regression

- **Simple Linear regression** is used to predict the **relationship** between **two variables** by applying **a linear equation to observed data.** There are two types of variable, one variable is called an **independent variable**, and the other is **a dependent variable**. Linear regression is commonly used for predictive analysis.
- In Linear regression, the objective is to **predict numerical features** like real estate or stock price, temperature, marks in an examination, sales revenue, etc.
- The underlying **predictor variable** and the **target variable** are **continuous** in nature.
- The measure of the **relationship** between two variables is shown by the **correlation coefficient**. The **range of the coefficient** lies between **-1 to +1.** This coefficient shows the **strength of the association** of the observed data **between two variables.**



Linear Equation

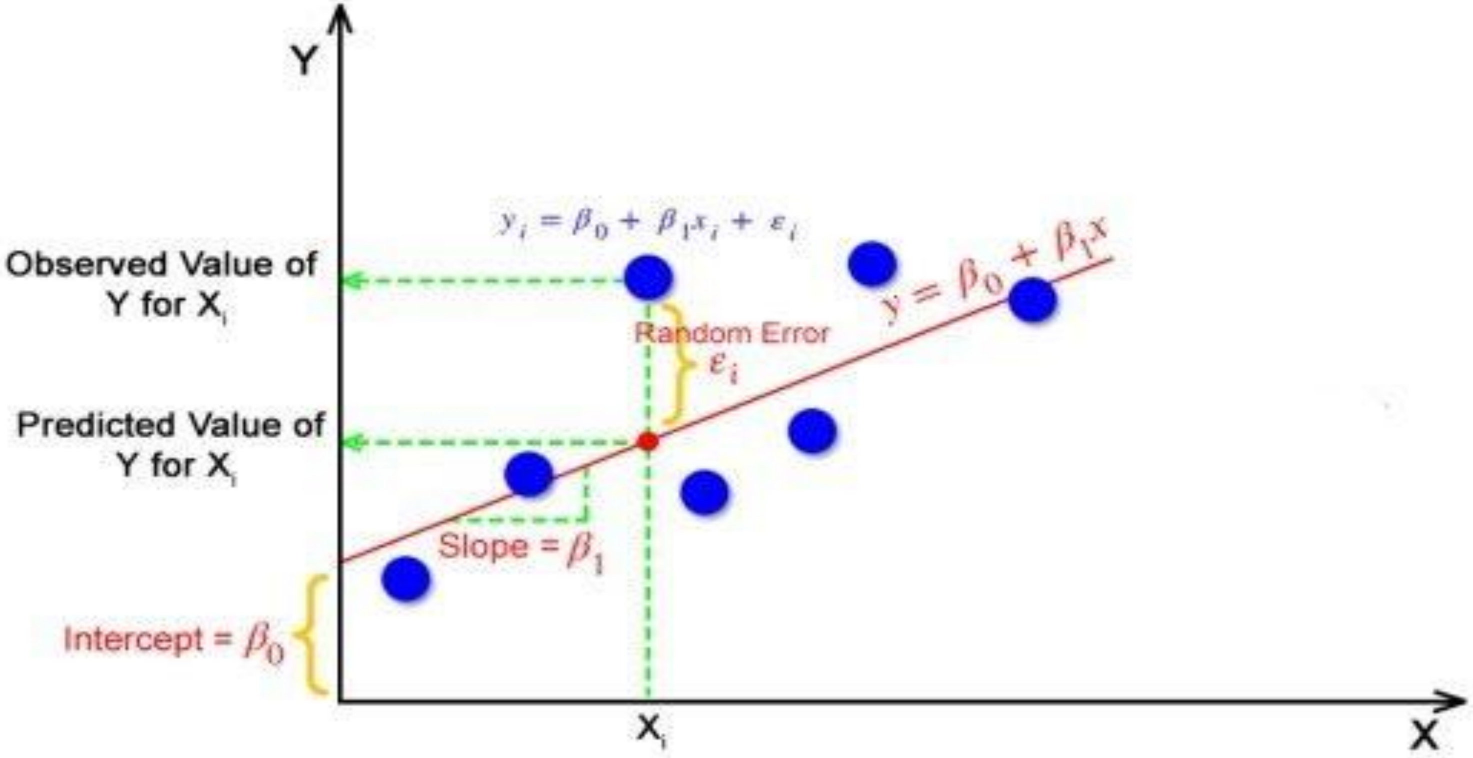


- Simple linear regression allows us to study the correlation between only two variables:
- One variable (X) is called independent variable or predictor.
- The other variable (Y), is known as dependent variable or outcome.
- and the simple linear regression equation is:

$$Y = B0 + B1.X + \varepsilon$$

or

$$Y=mx+c$$



Where:

- X – the value of the **independent variable,**
- Y – the value of the **dependent variable.**
- B0 – is a **constant / Intercept** point (shows the value of Y when the value of X=0)
- B1 – **slope / the regression coefficient** (shows how much Y changes for each unit change in X– regression coefficient estimates of some unknown parameters to describe the relationship between a predictor variable and the corresponding response.)
- ε – Random Error--is required as the best fit line also doesn't include the data points perfectly.

Linear Equation : y = mx+c & calculation of slope

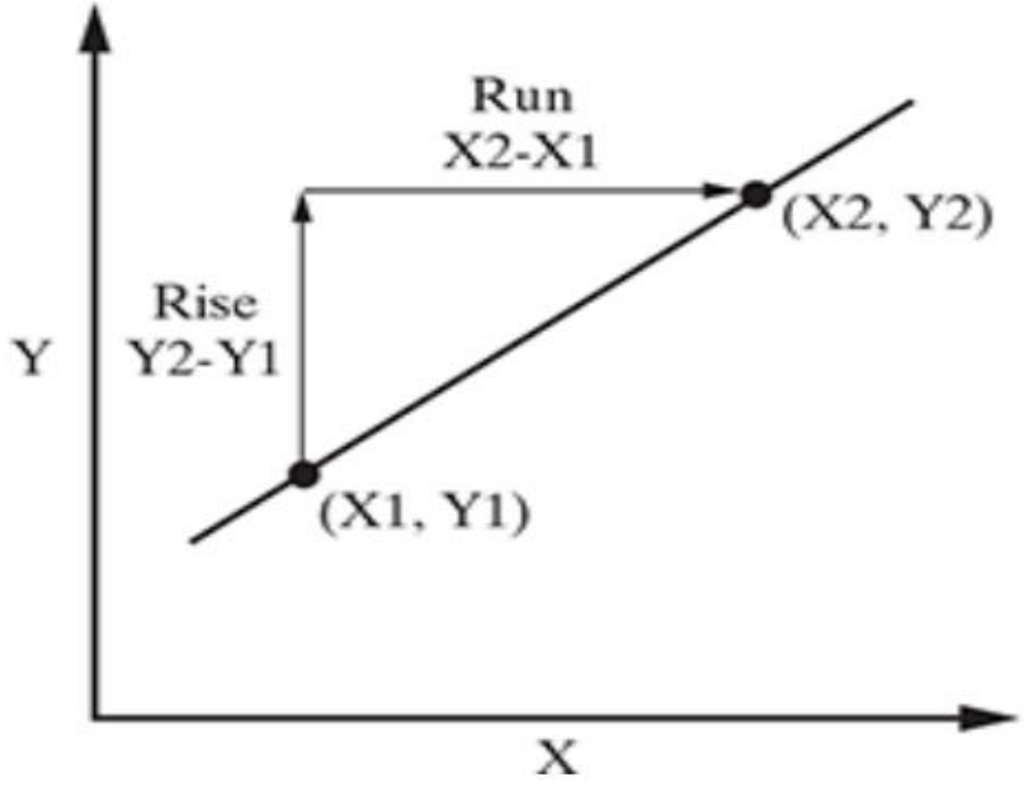


FIG. 8.2 Rise and run representation

- **Slope = Change in Y/Change in X**
- **Slope = Rise/Run**

- **Rise** is the change in Y-axis (Y2 – Y1) and
- **Run** is the change in X-axis (X2 – X1)
- For example:

$(X_1, Y_1) = (-3, -2)$ and $(X_2, Y_2) = (2, 2)$

Rise $= (Y_2 - Y_1) = (2 - (-2)) = 2 + 2 = 4$

Run $= (X_2 - X_1) = (2 - (-3)) = 2 + 3 = 5$

Slope = Rise/Run = 4/5 = 0.8

Positive slopes vs Negative slopes

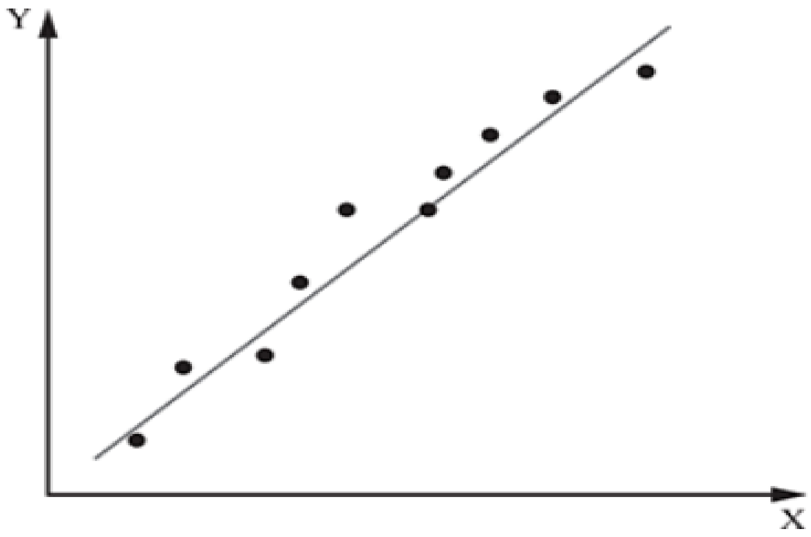


FIG. 8.3 Linear positive slope

Slope = Rise/Run = $(Y_2 - Y_1) / (X_2 - X_1)$ = Delta (Y) / Delta(X)

- Scenario 1 for positive slope: Delta (Y) is positive and Delta (X) is positive
- Scenario 2 for positive slope: Delta (Y) is negative and Delta (X) is negative

Curve linear positive slope

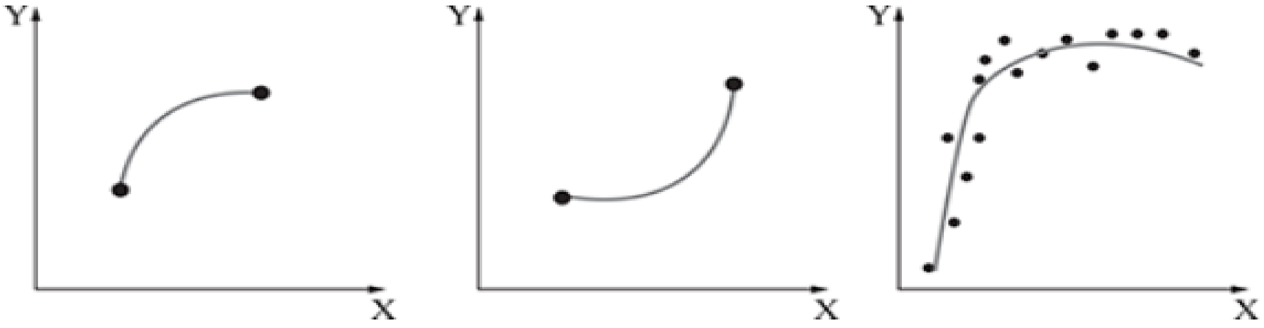


FIG. 8.4 Curve linear positive slope

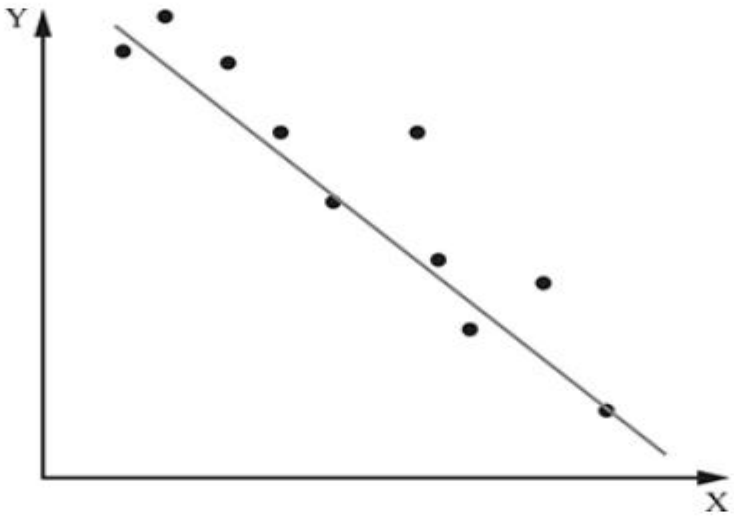


FIG. 8.5 Linear negative slope

Slope = Rise/Run = $(Y_2 - Y_1) / (X_2 - X_1)$ = Delta (Y) / Delta(X)

- Scenario 1 for negative slope: Delta (Y) is positive and Delta (X) is negative
- Scenario 2 for negative slope: Delta (Y) is negative and Delta (X) is positive

Curve linear negative slope

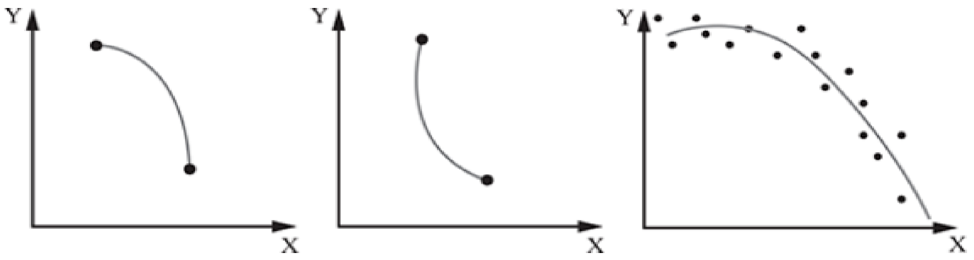
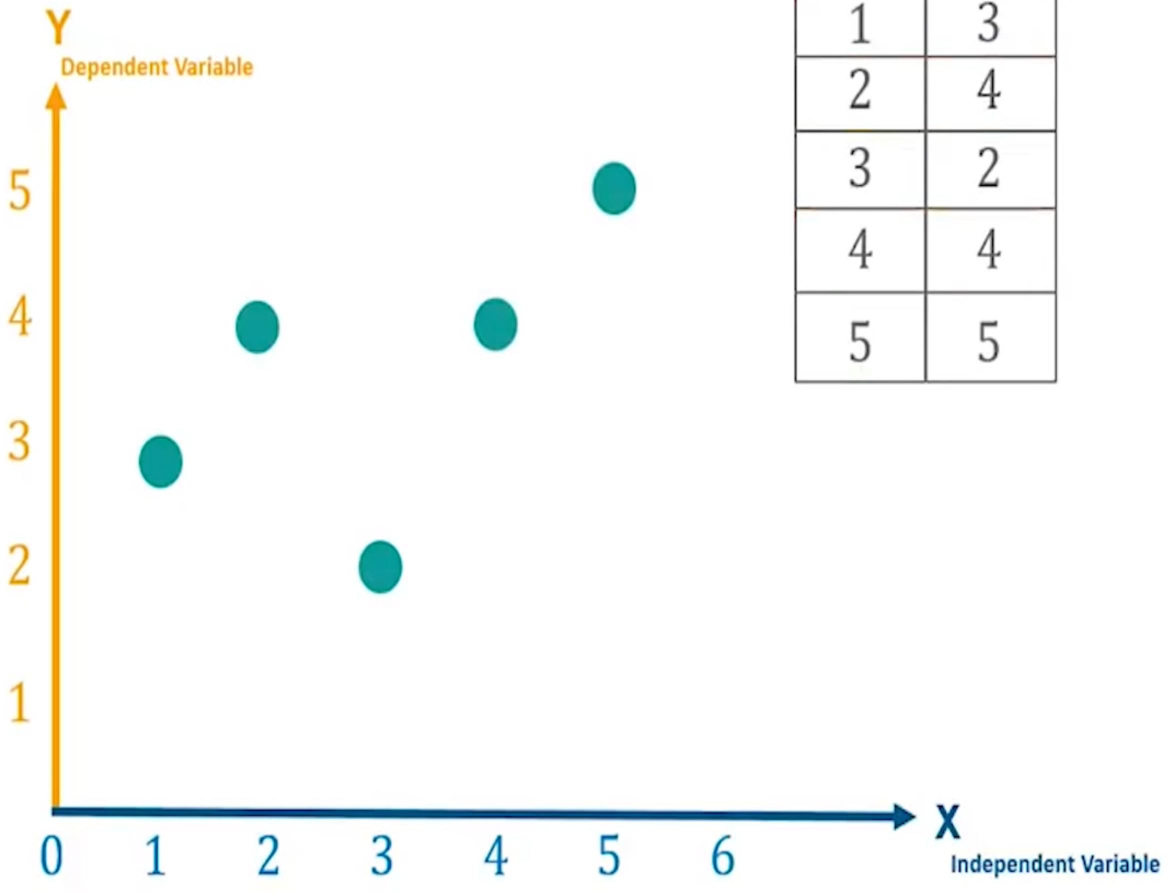


FIG. 8.6 Curve linear negative slope

Example-1

Understanding Linear Regression Algorithm

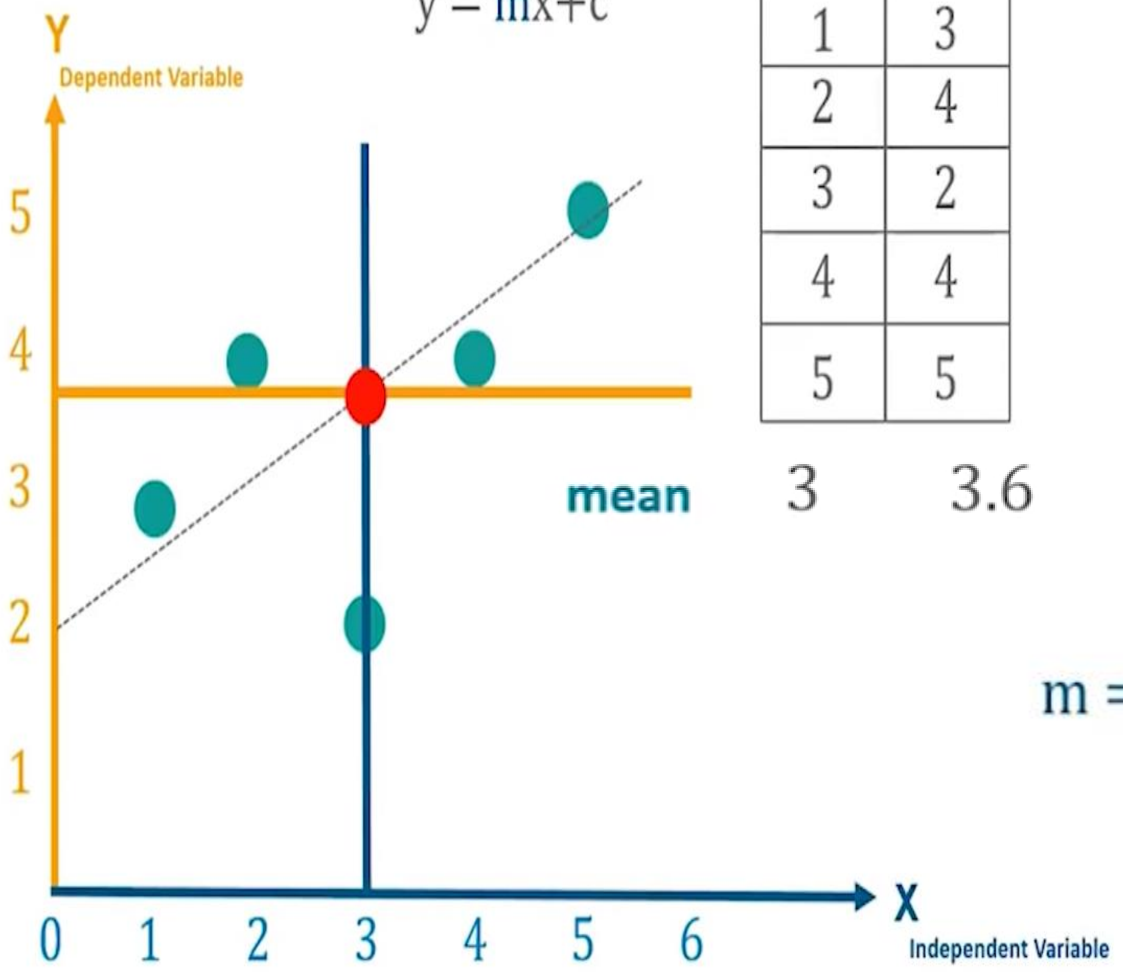


x	y
1	3
2	4
3	2
4	4
5	5

Understanding Linear Regression Algorithm

x	y
1	3
2	4
3	2
4	4
5	5
mean 3	3.6

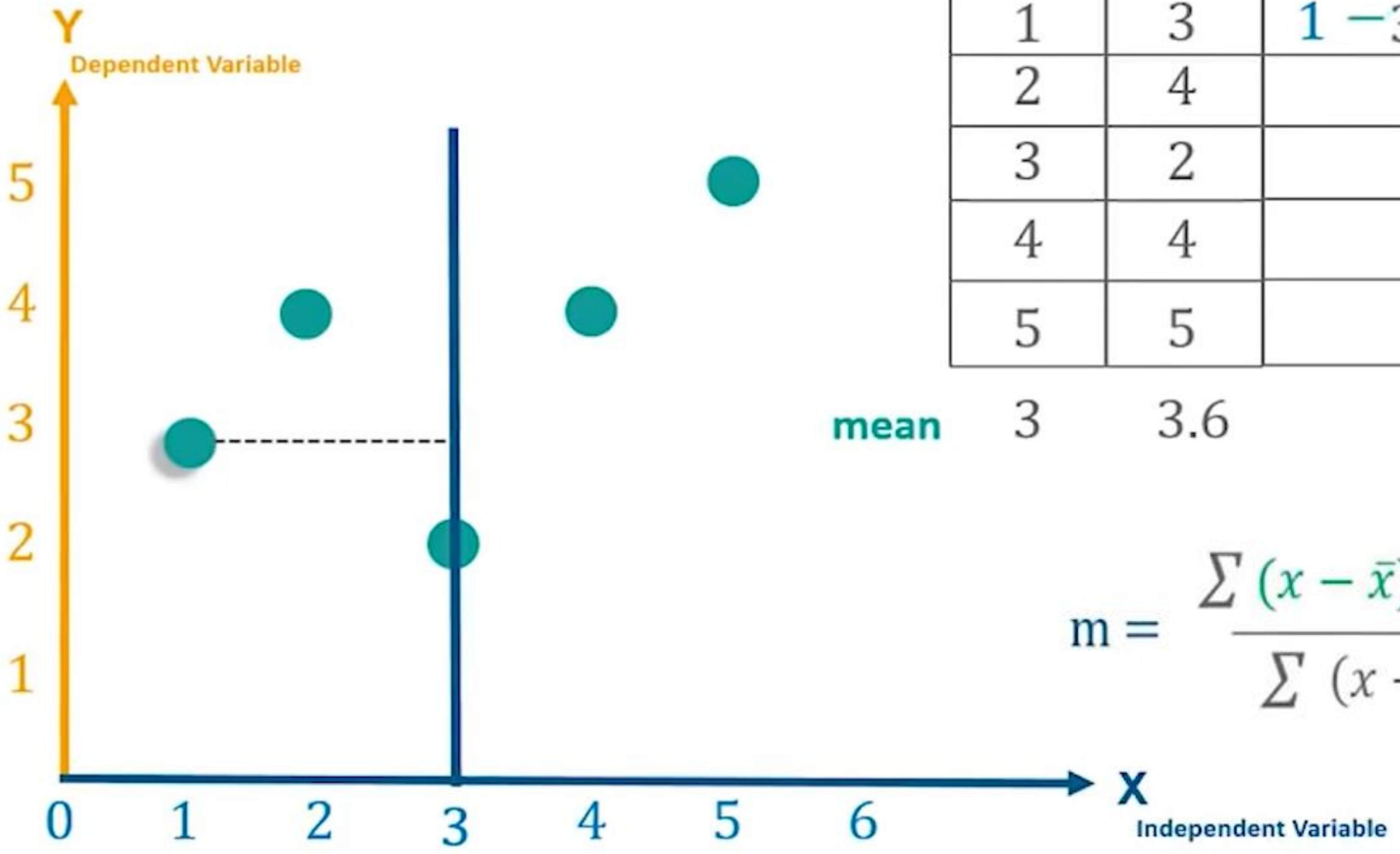
Understanding Linear Regression Algorithm



x	y
1	3
2	4
3	2
4	4
5	5
mean 3	3.6

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

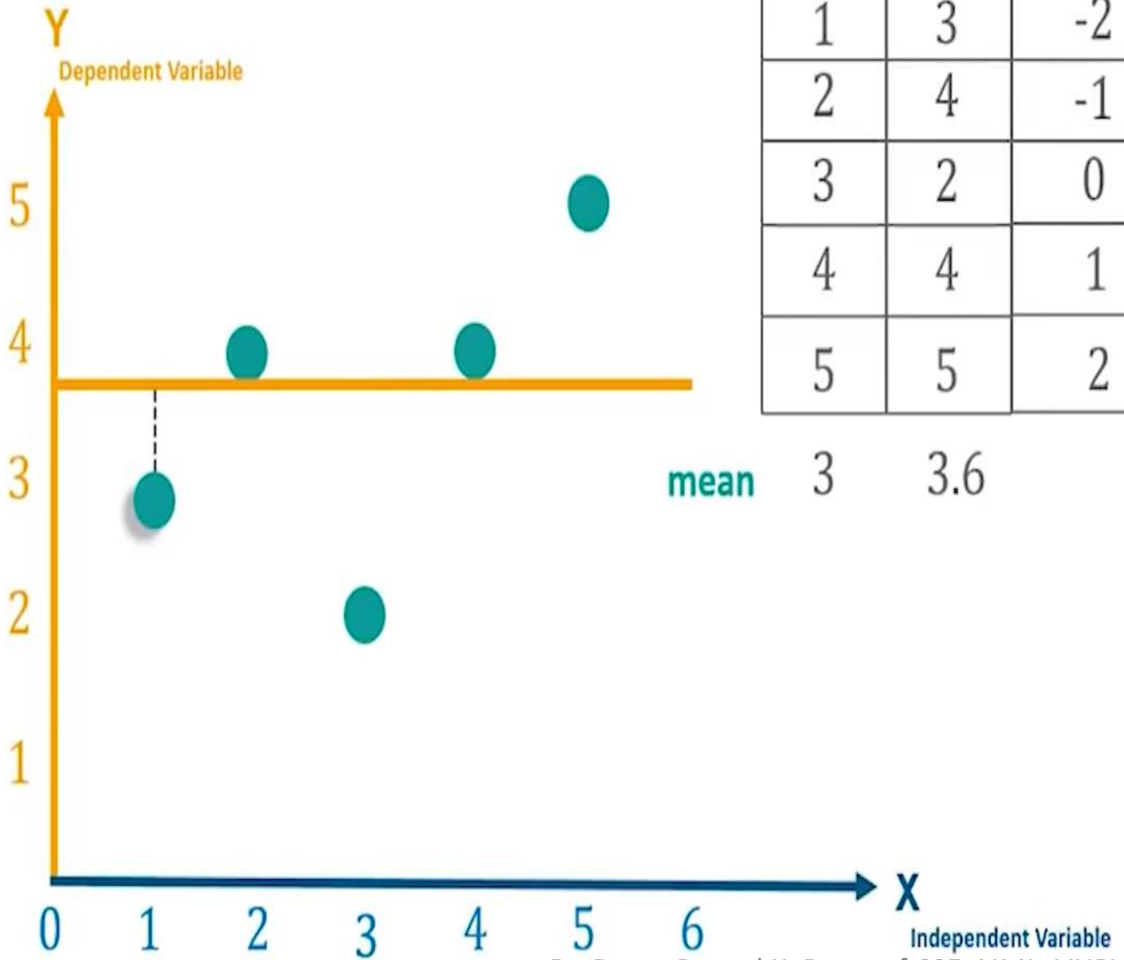
Understanding Linear Regression Algo



x	y	$x - \bar{x}$
1	3	1 − 3
2	4	
3	2	
4	4	
5	5	
mean 3	3.6	

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

Understanding Linear Regression Algorithm



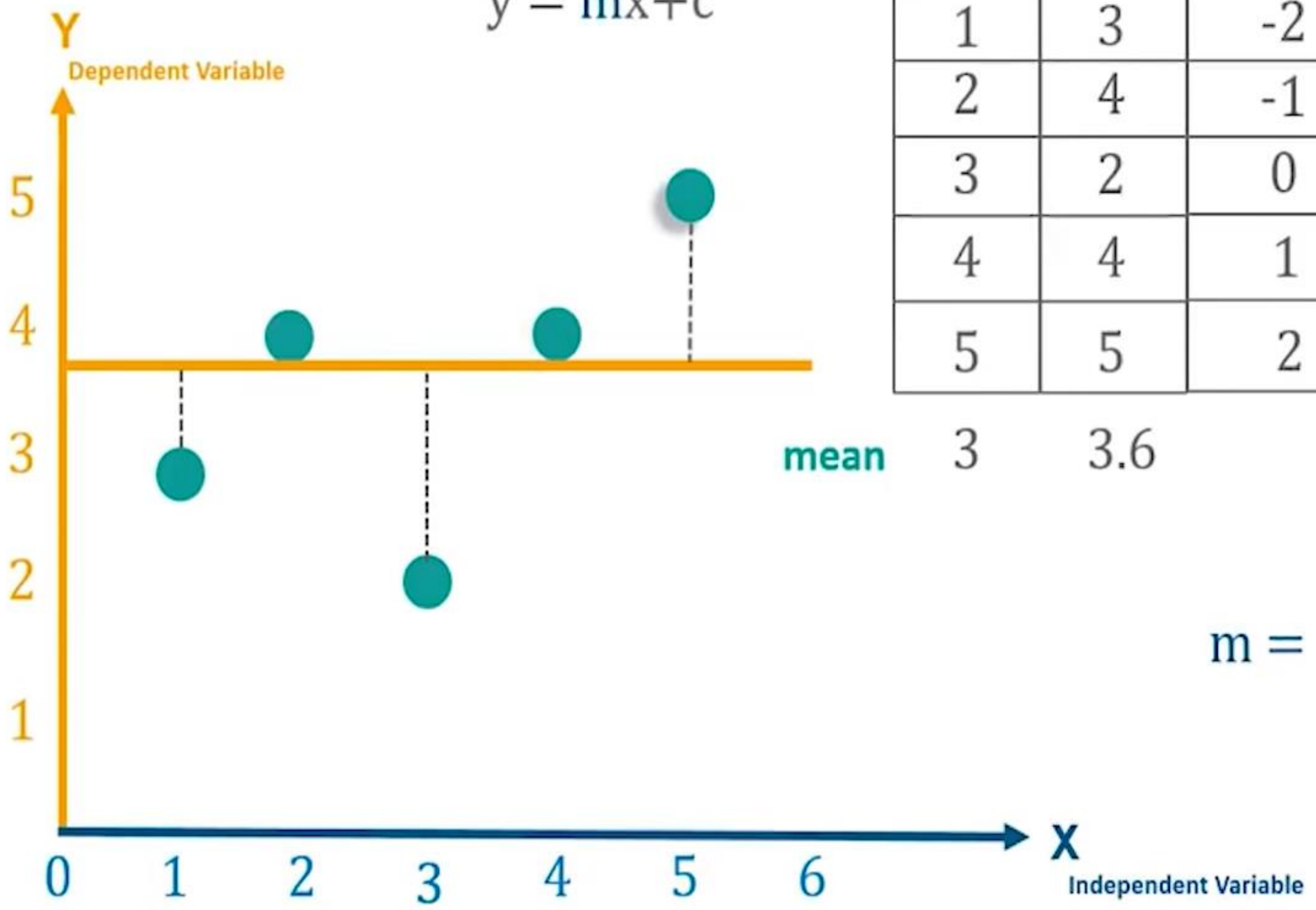
x	y	$x - \bar{x}$	$y - \bar{y}$
1	3	-2	3 − 3.6
2	4	-1	
3	2	0	
4	4	1	
5	5	2	

mean 3 3.6

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

Understanding Linear Regression Algorithm

$y = mx + c$

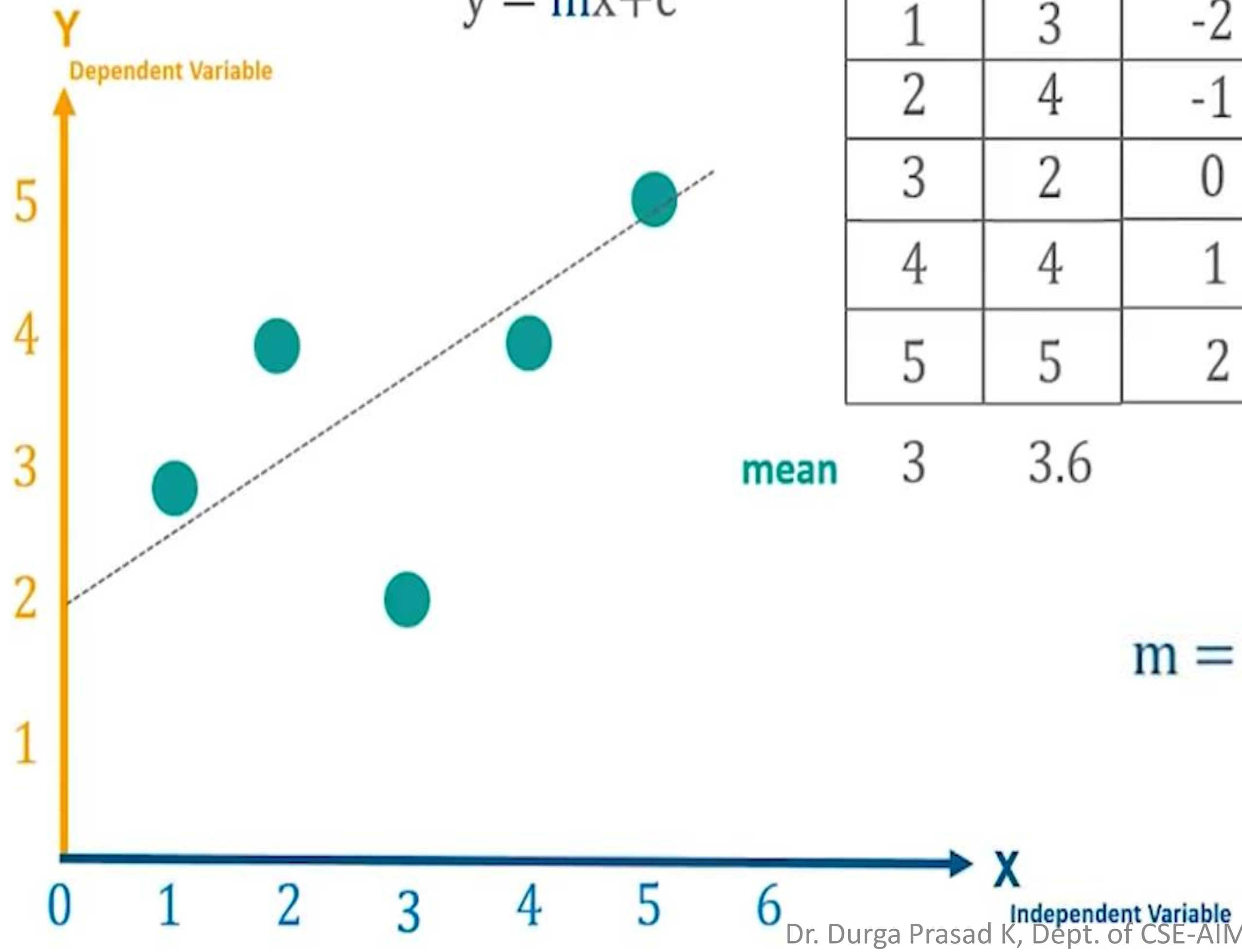


x	y	$x - \bar{x}$	$y - \bar{y}$
1	3	-2	-0.6
2	4	-1	0.4
3	2	0	-1.6
4	4	1	0.4
5	5	2	1.4
mean 3	3.6		

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

Understanding Linear Regression Algorithm

$y = mx + c$



x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$
1	3	-2	-0.6	4
2	4	-1	0.4	1
3	2	0	-1.6	0
4	4	1	0.4	1
5	5	2	1.4	4

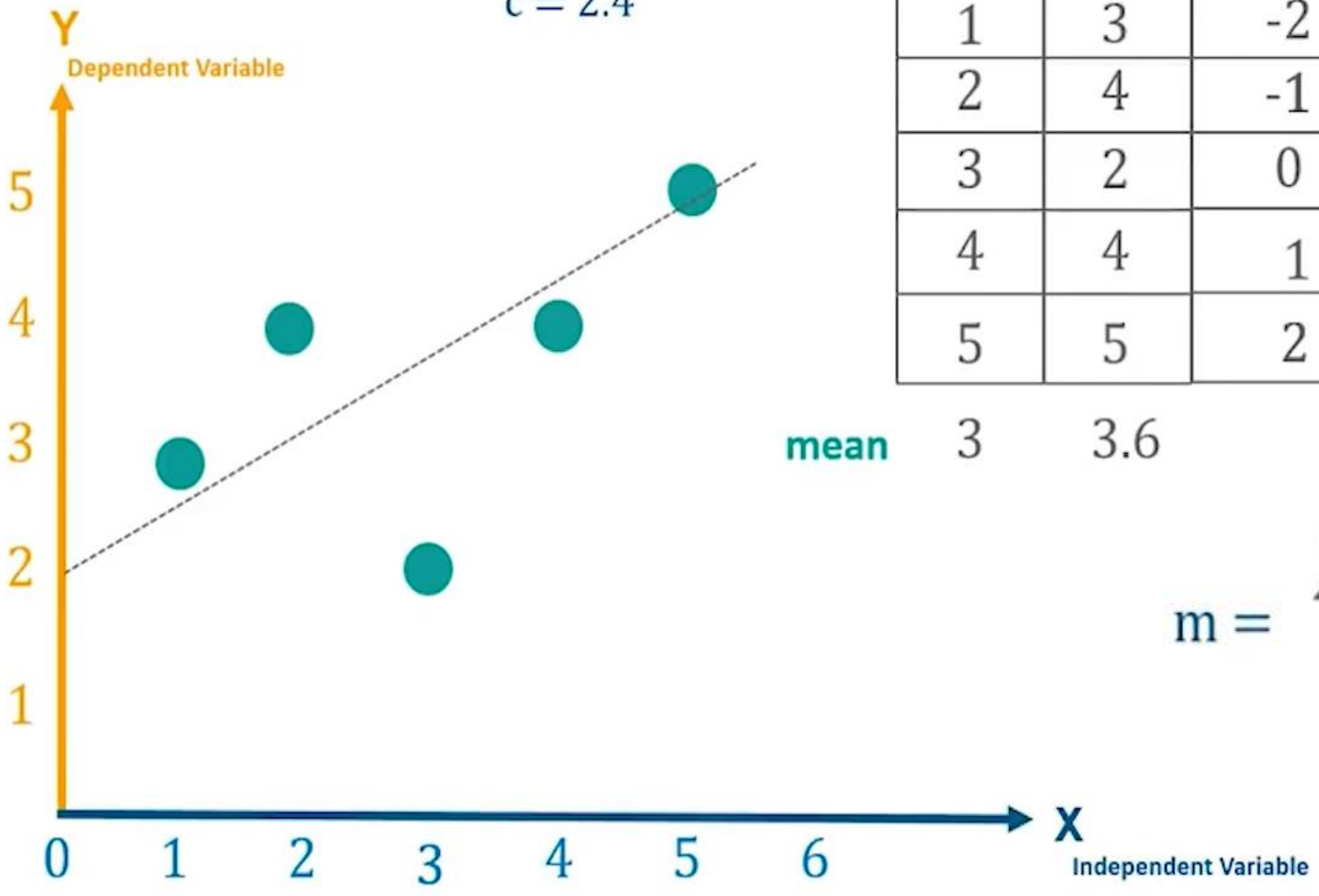
mean 3 3.6

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

Understanding Linear Regression Algorithm

$y = mx + c$

$c = 2.4$



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

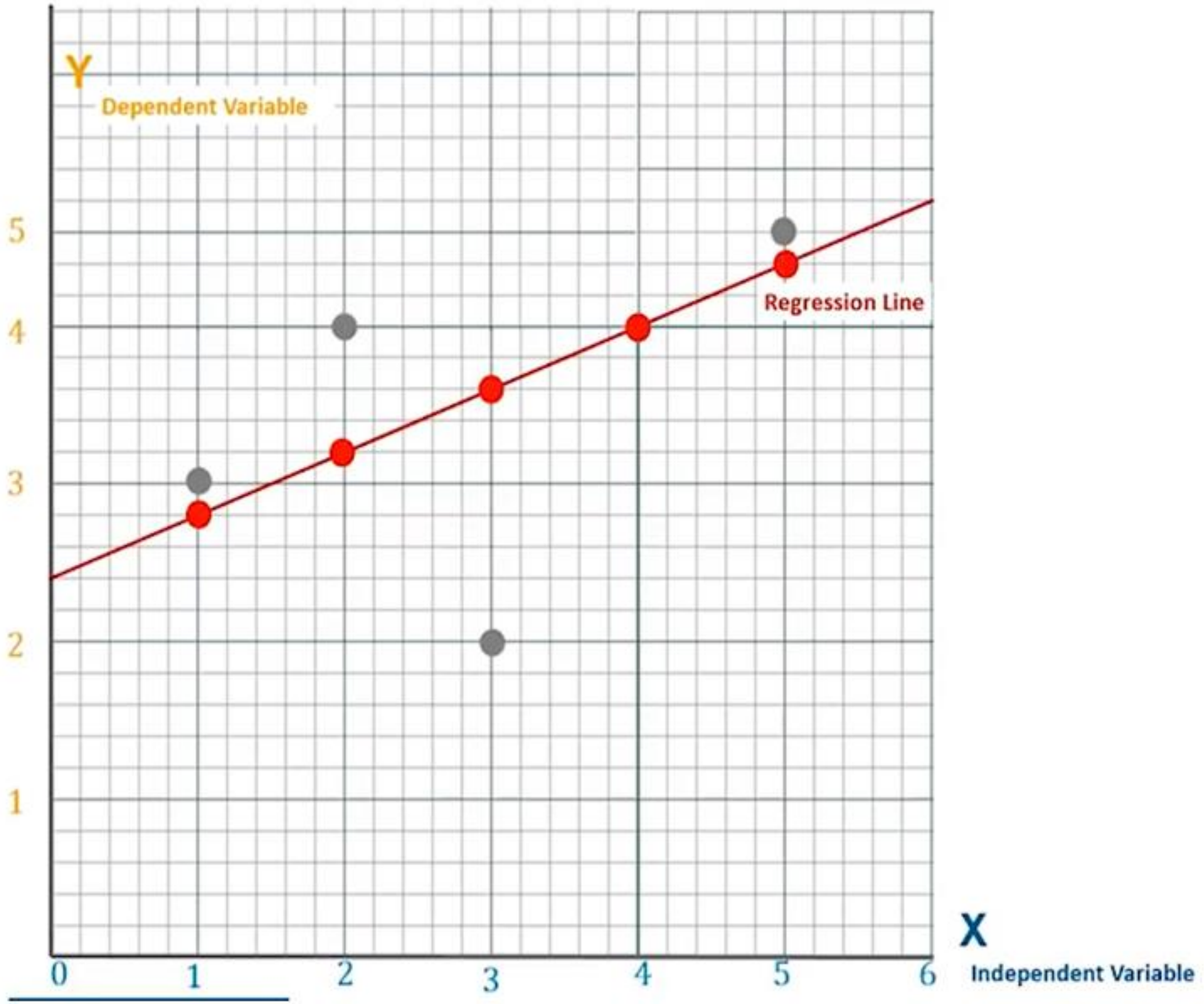
mean 3 3.6 $\Sigma = 10$ $\Sigma = 4$

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

$m = 0.4$

$c = 2.4$

$y = 0.4x + 2.4$



$m = 0.4$

$c = 2.4$

$y = 0.4x + 2.4$

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

$y = 0.4 \times 1 + 2.4 = 2.8$

$y = 0.4 \times 2 + 2.4 = 3.2$

$y = 0.4 \times 3 + 2.4 = 3.6$

$y = 0.4 \times 4 + 2.4 = 4.0$

$y = 0.4 \times 5 + 2.4 = 4.4$

Performance metrics for Regression

The metrics used for regression are different from the **classification metrics**. It means we cannot use the **Accuracy** metric to evaluate a regression model; instead, the performance of a Regression model is reported **as errors in the prediction**.

1. **Mean Absolute Error**--MAE measures the absolute difference between actual and predicted values, where **absolute means** taking **a number as Positive.**

   Y is the Actual outcome, Y' is the predicted outcome, and N is the total number of data points.

   mean absolute error = it's the mean of the sum of the absolute values of residuals/errors.

$$MAE = 1/N \sum |Y - Y'|$$

2. **Mean Squared Error**--It measures the average of the **Squared difference** between predicted values and the actual value given by the model.

   mean square error = it's the mean of the sum of the squares of residuals/errors.

$$MSE = 1/N \sum (Y - Y')^2$$

3. **R2 Score**--**R squared error** is also known as **Coefficient of Determination**, which is another popular metric used for **Regression model evaluation**. The R-squared metric enables us to compare our model with **a constant baseline** to determine the performance of the model. To select the **constant baseline**, we need to take **the mean of the data** and **draw the line** at the mean.

$$R^2 = 1 - \frac{MSE(Model)}{MSE(Baseline)}$$

or

$$R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$$

Performance metrics

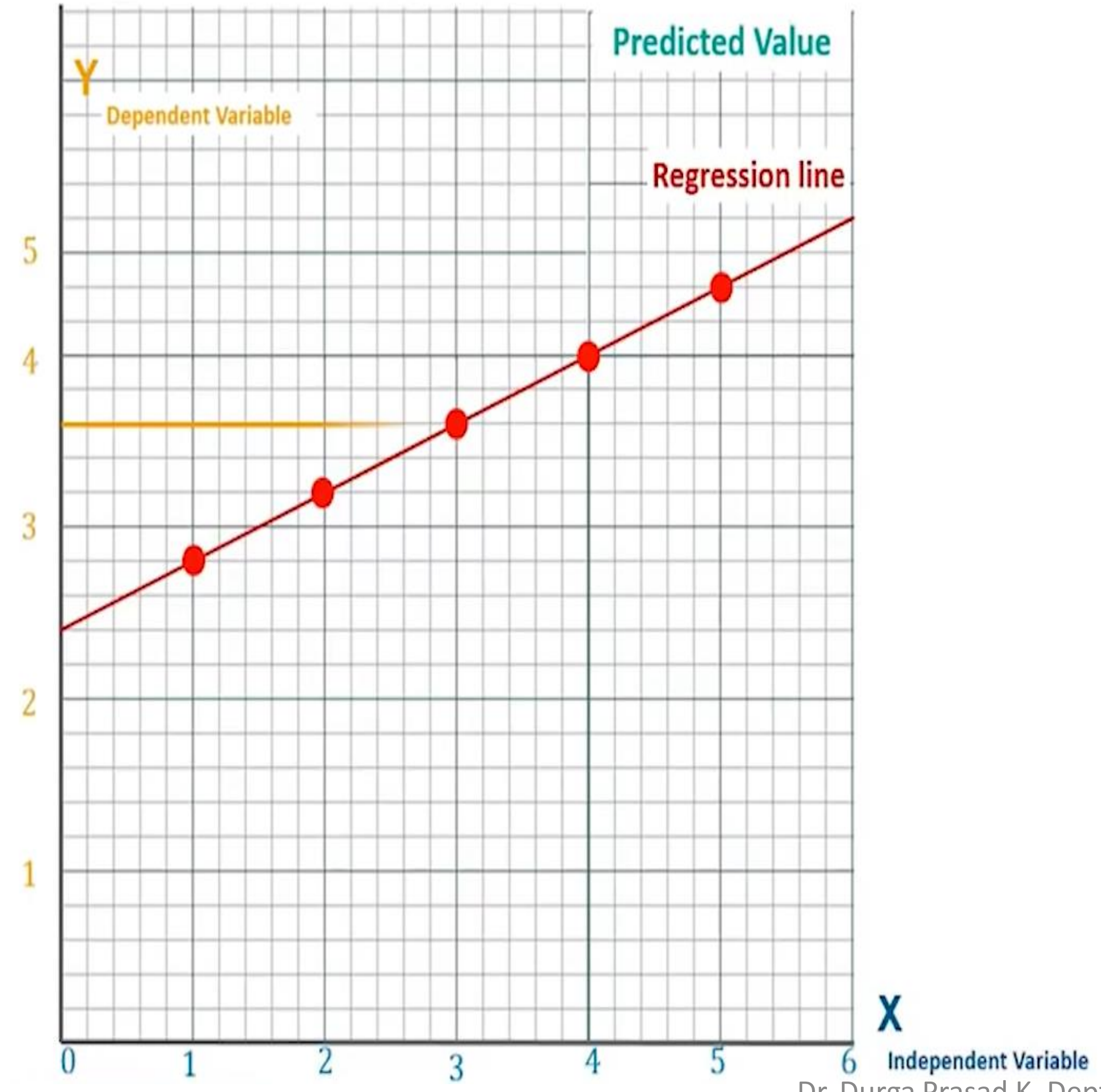
4. Adjusted R2

- Adjusted R squared, as the name suggests, is the **improved version of R squared error**. R square has a limitation of improvement of a score on increasing the terms, even though the **model is not improving**, and it may mislead the data scientists.
- To overcome the issue of R square, adjusted R squared is used, which will always **show a lower value than R^2**. It is because it adjusts the values of increasing predictors and only shows improvement if there is a real improvement.
- We can calculate the adjusted R squared as follows:

$$R_a^2 = 1-\left[\left(\frac{n-1}{n-k-1}\right)\times\left(1-R^2\right)\right]$$

- n is the number of observations
- k denotes the number of independent variables
- and R_a^2 denotes the adjusted R^2

Calculation of R^2



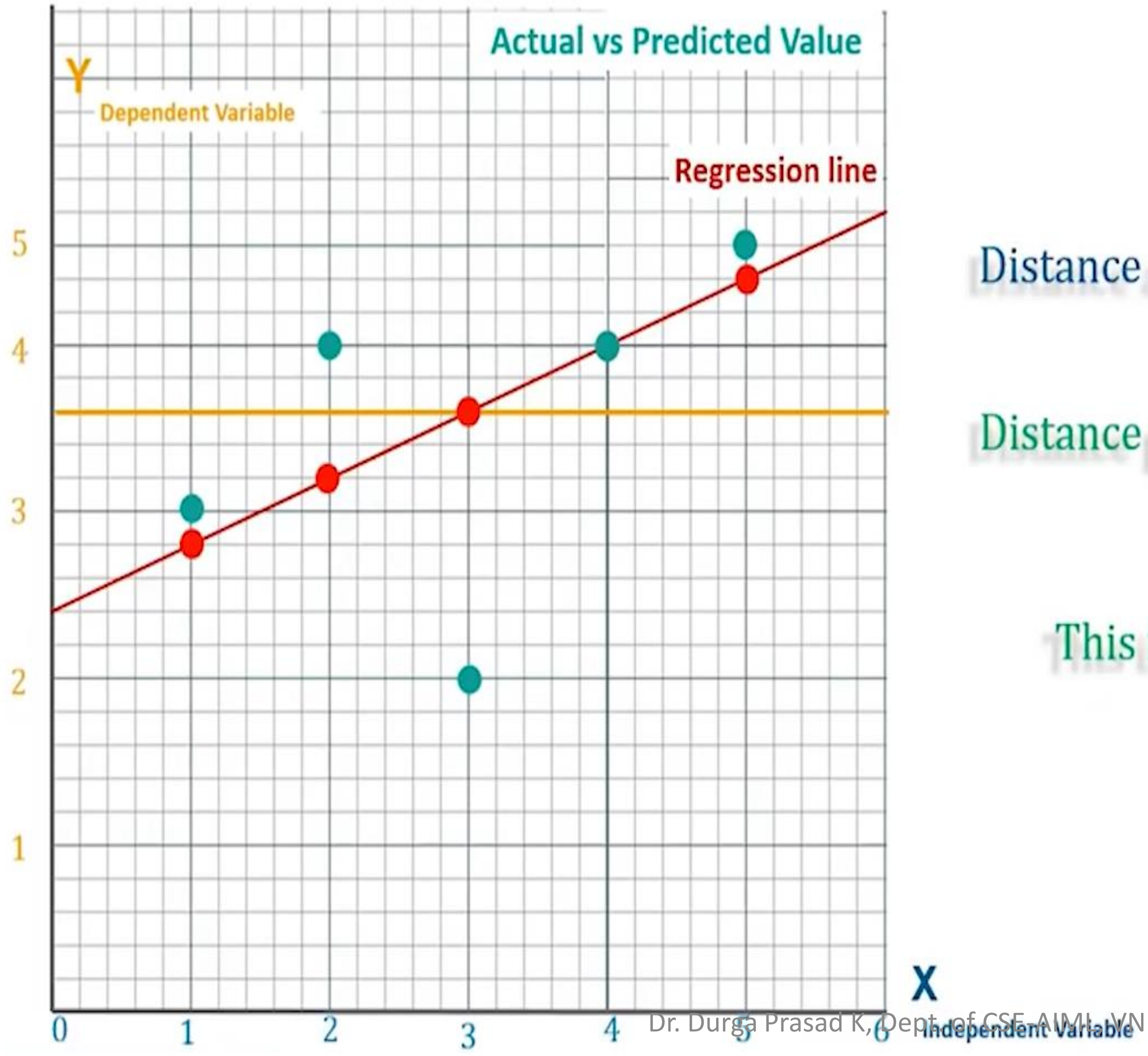
x	y_p
1	2.8
2	3.2
3	3.6
4	4.0
5	4.4

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

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

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

Calculation of R^2



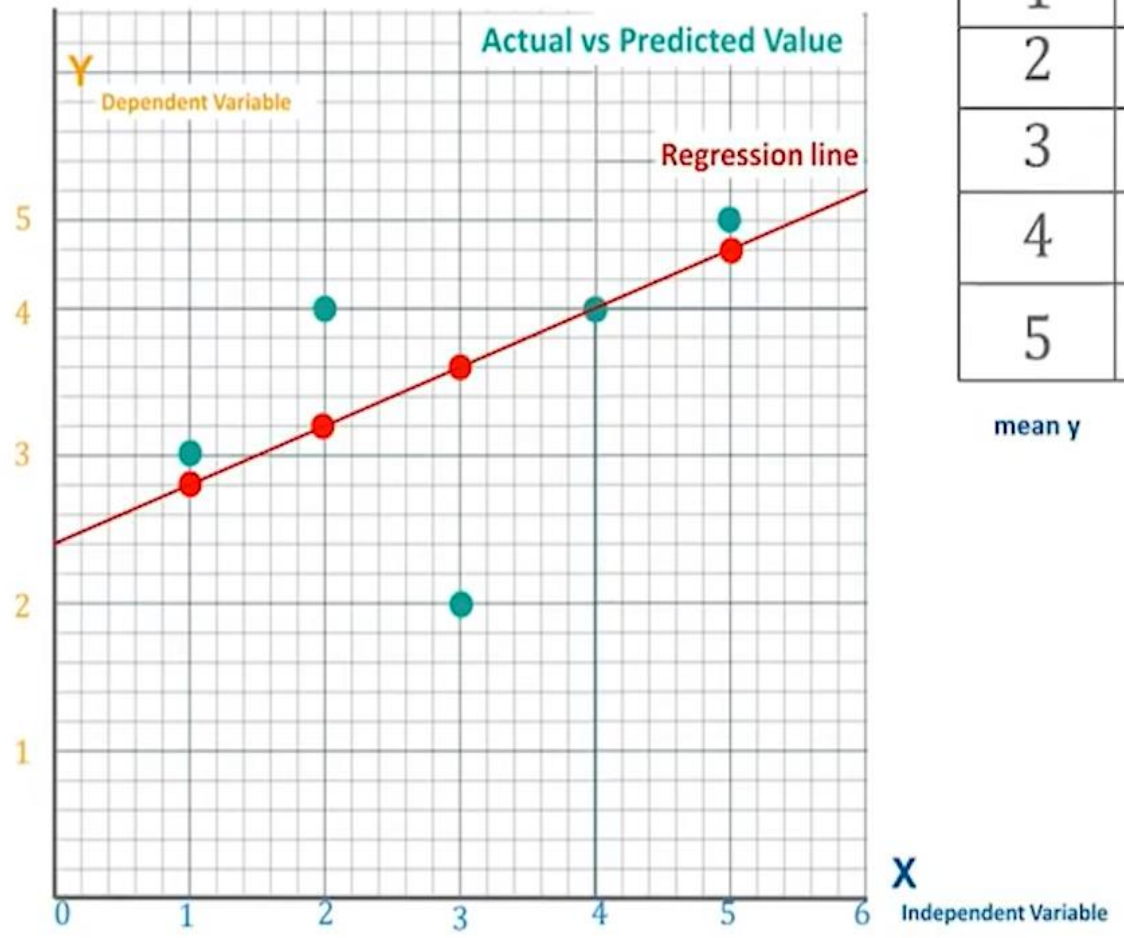
Distance actual - mean

vs

Distance predicted - mean

This is nothing but $R^2 = \dfrac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$

Calculation of R^2

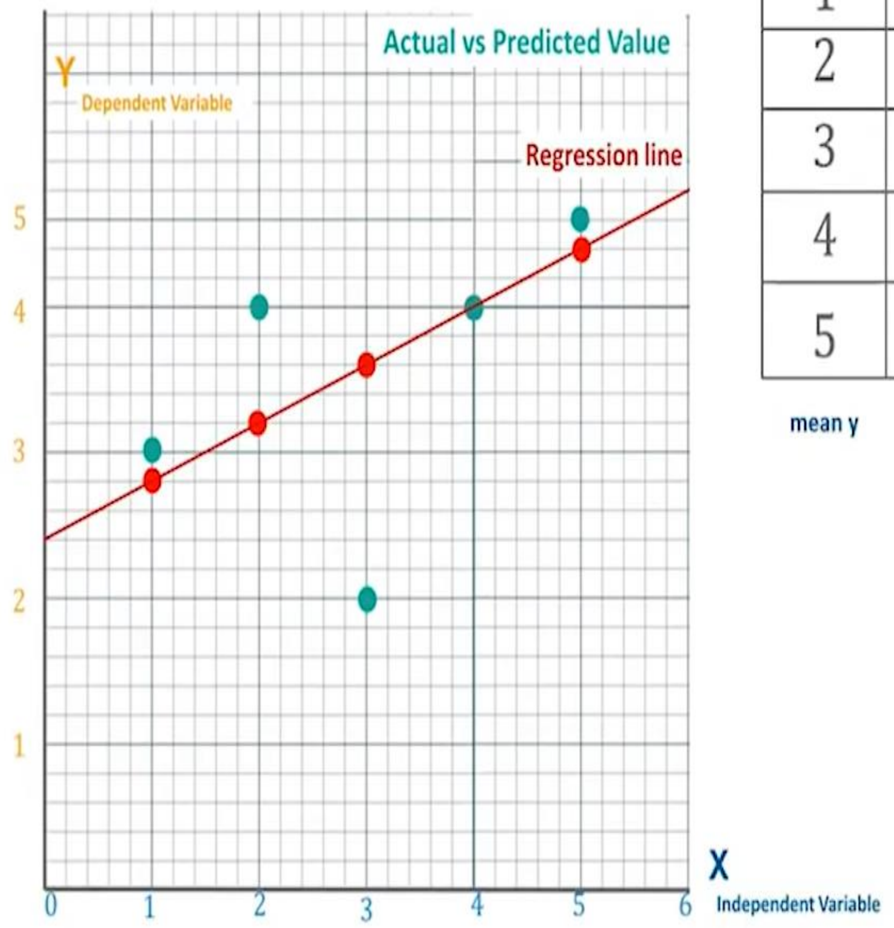


x	y	$y - \bar{y}$	$(y - \bar{y})^2$	y_p	$(y_p - \bar{y})$
1	3	– 0.6	0.36	2.8	-0.8
2	4	0.4	0.16	3.2	-0.4
3	2	–1.6	2.56	3.6	0
4	4	0.4	0.16	4.0	0.4
5	5	1.4	1.96	4.4	0.8

mean y 3.6

$$R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$$

Calculation of R^2

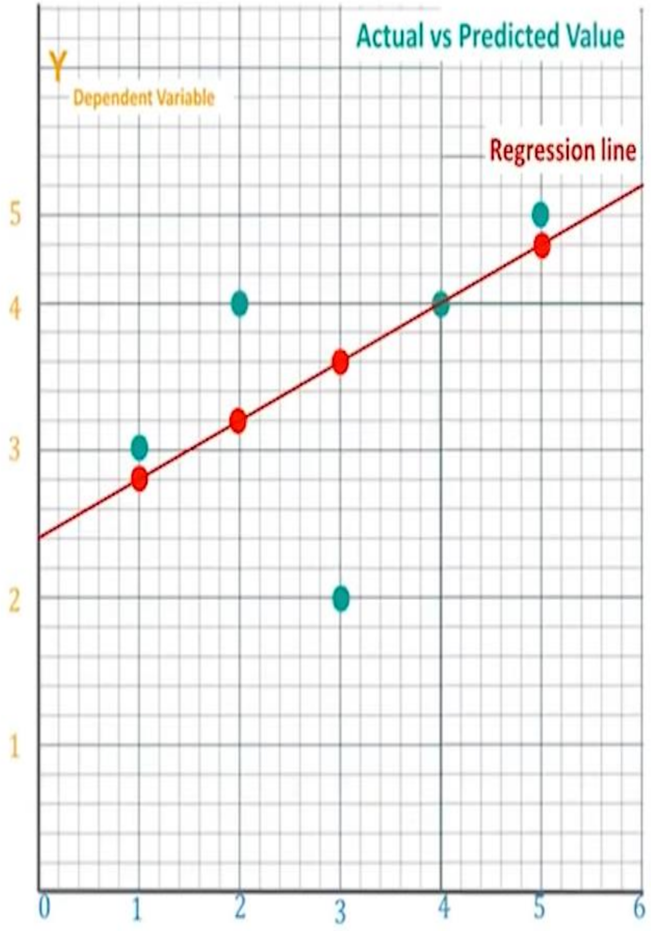


x	y	$y-\bar{y}$	$(y-\bar{y})^2$	y_p	$(y_p - \bar{y})$
1	3	– 0.6	0.36	2.8	-0.8
2	4	0.4	0.16	3.2	-0.4
3	2	–1.6	2.56	3.6	0
4	4	0.4	0.16	4.0	0.4
5	5	1.4	1.96	4.4	0.8

mean y 3.6

$$R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$$

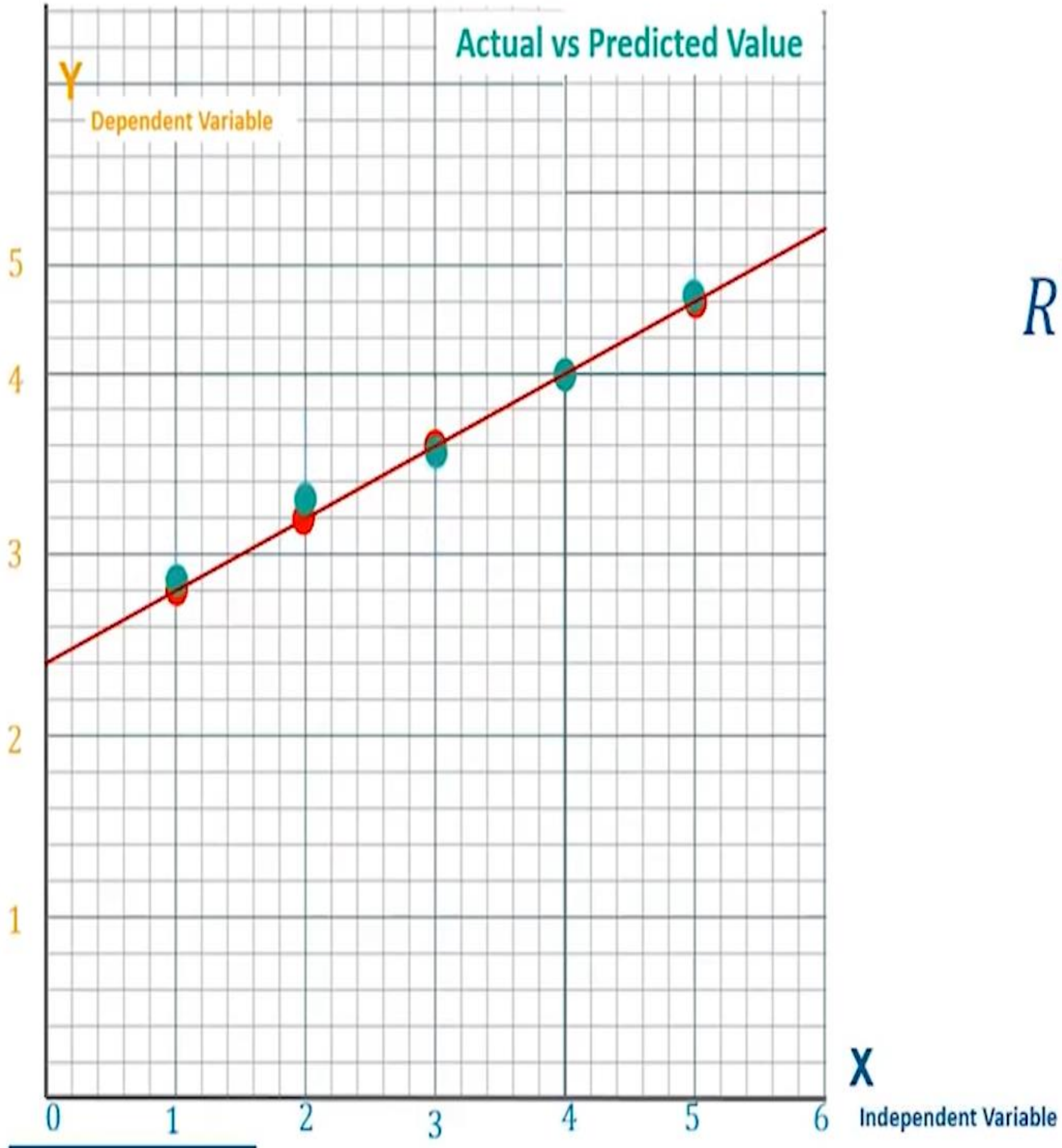
Calculation of R^2



x	y	$y - \bar{y}$	$(y - \bar{y})^2$	y_p	$(y_p - \bar{y})$	$(y_p - \bar{y})^2$
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

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

Calculation of R^2



$$R^2 \approx 0.9$$

Simple Linear Regression- Example-2

***When coefficient not given or need to calculate coefficient also from the given problem statement**

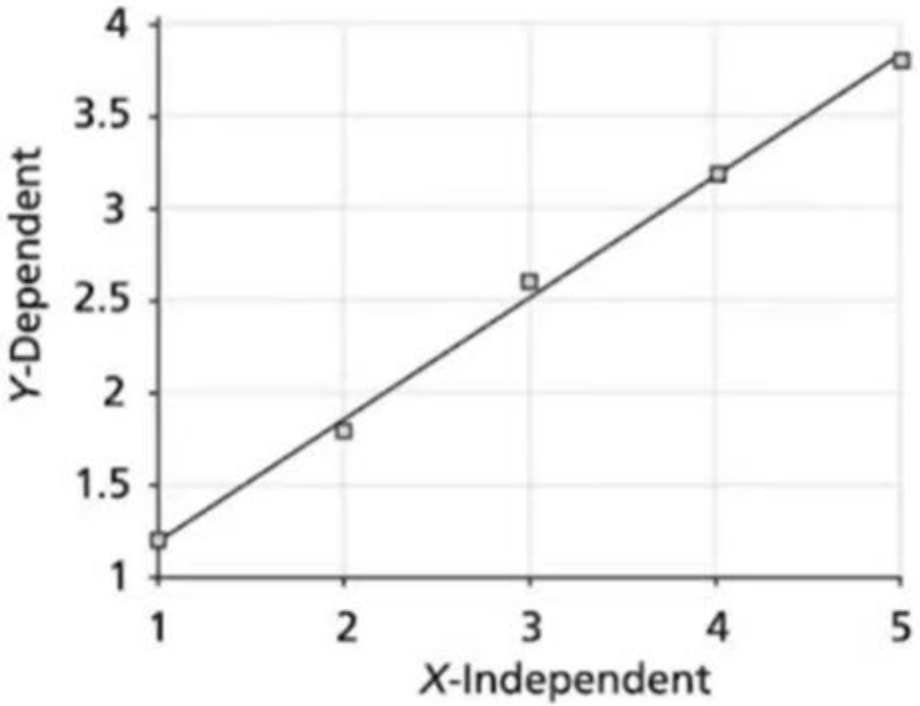
- Let us consider an example where the five weeks' sales data (in Thousands) is given as shown in Table.
- Apply linear regression technique to predict the 7th and 12th week sales.

x_i (Week)	y_j (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

What are the sales of 7th week and 12th week ?

Simple Linear Regression- Example-2

***When coefficient not given or need to calculate coefficient also from the given problem statement**



x_i (Week)	y_j (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

What are the sales of 7th week and 12th week ?

Simple Linear Regression- Example-2

- Linear regression equation is given by
- $y = a_0 + a_1 * x + e$
- *where*
- $a_1 = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2}$
- $a_0 = \bar{y} - a_1 * \bar{x}$

x_i $(Week)$	y_j $(Sales\ in\ Thousands)$
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

Simple Linear Regression- Example-2

- Here, there are 5 items, i.e., i = 1, 2, 3, 4, 5.

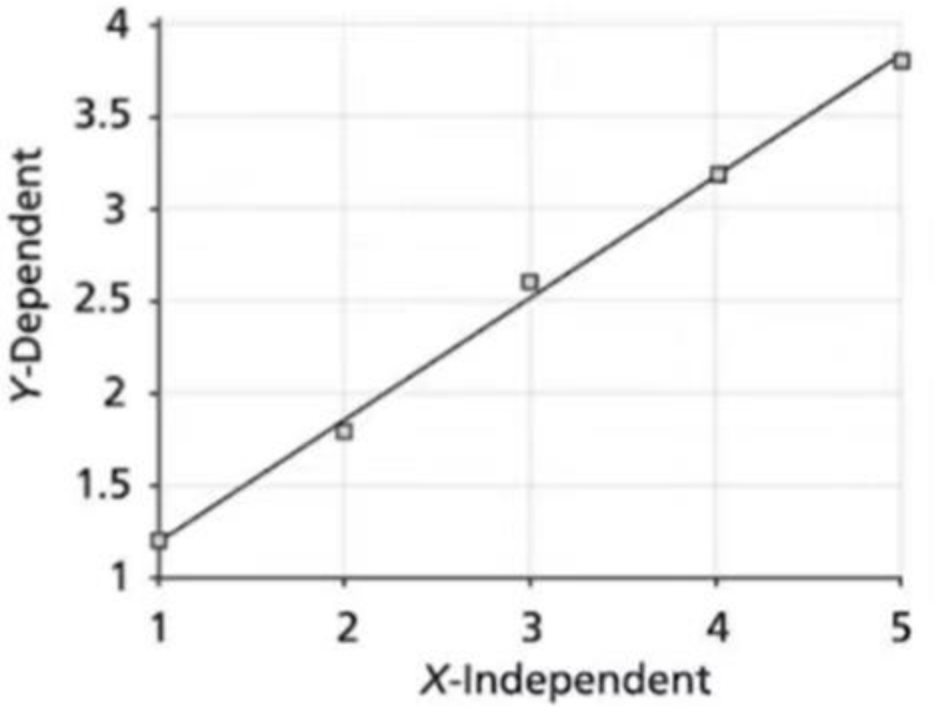
	x_i (Week)	y_i (Sales in Thousands)	x_i^2	$x_i * y_i$
	1	1.2	1	1.2
	2	1.8	4	3.6
	3	2.6	9	7.8
	4	3.2	16	12.8
	5	3.8	25	19
Sum	15	12.6	55	44.4
Average	$\bar{x} = 3$	$\bar{y} = 2.52$	$\overline{x^2} = 11$	$\overline{xy} = 8.88$

- $y = a_0 + a_1 * x + e$
- where
- $a_1 = \dfrac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2}$
- $a_0 = \bar{y} - a_1 * \bar{x}$

Simple Linear Regression- Example-2

- $\bar{x} = 3$ $\bar{y} = 2.52$ $\overline{x^2} = 11$ $\overline{xy} = 8.88$

- $a_1 = \frac{(\overline{xy})-(\bar{x})(\bar{y})}{\overline{x^2}-\bar{x}^2} = \frac{8.88-3*2.52}{11-3^2} = 0.66$
- $a_0 = \bar{y} - a_1 * \bar{x} = 2.52 - 0.66 * 3 = 0.54$
- **Regression equation is**
- $\boldsymbol{y = a_0 + a_1 * x}$
- $\boldsymbol{y = 0.54 + 0.66 * x}$



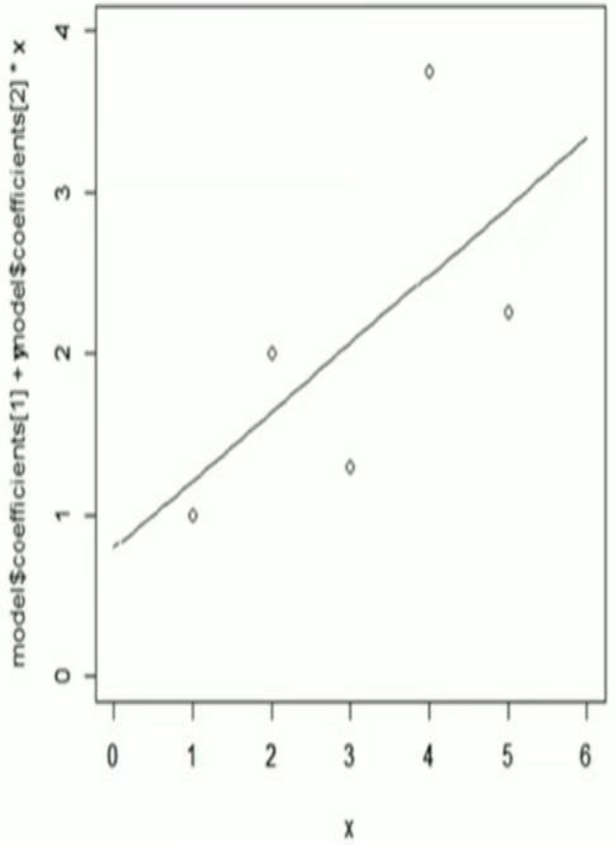
Simple Linear Regression- Example-2

- **Regression equation is**
- $y = a_0 + a_1 * x$
- $y = 0.54 + 0.66 * x$
- The predicted 7th week sale (when x = 7) is,
- **y = 0.54 + 0.66 x 7 = 5.16**
- the predicted 12th week sale (when x = 12) is,
- **y = 0.54 + 0.66 x 12 = 8.46**

Simple Linear Regression

Class room Assignment ?

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25



- $y = a_0 + a_1 * x + e$
- $where$
- $a_1 = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2}$
- $a_0 = \bar{y} - a_1 * \bar{x}$

Multiple Linear Regression

Multiple Linear Regression

- One of the most common types of **predictive analysis** is **multiple linear regression**. This type of analysis allows you to understand the **relationship** between **a continuous dependent variable** and **two** or **more independent variables/predictors**.

- In a multiple regression model, **two** or **more** **independent variables**, i.e. **predictors** are involved in the model.

$$\text{Price}_{\text{Property}} = f\,(\text{Area}_{\text{Property,}}\ \text{location, floor, Ageing, Amenities})$$

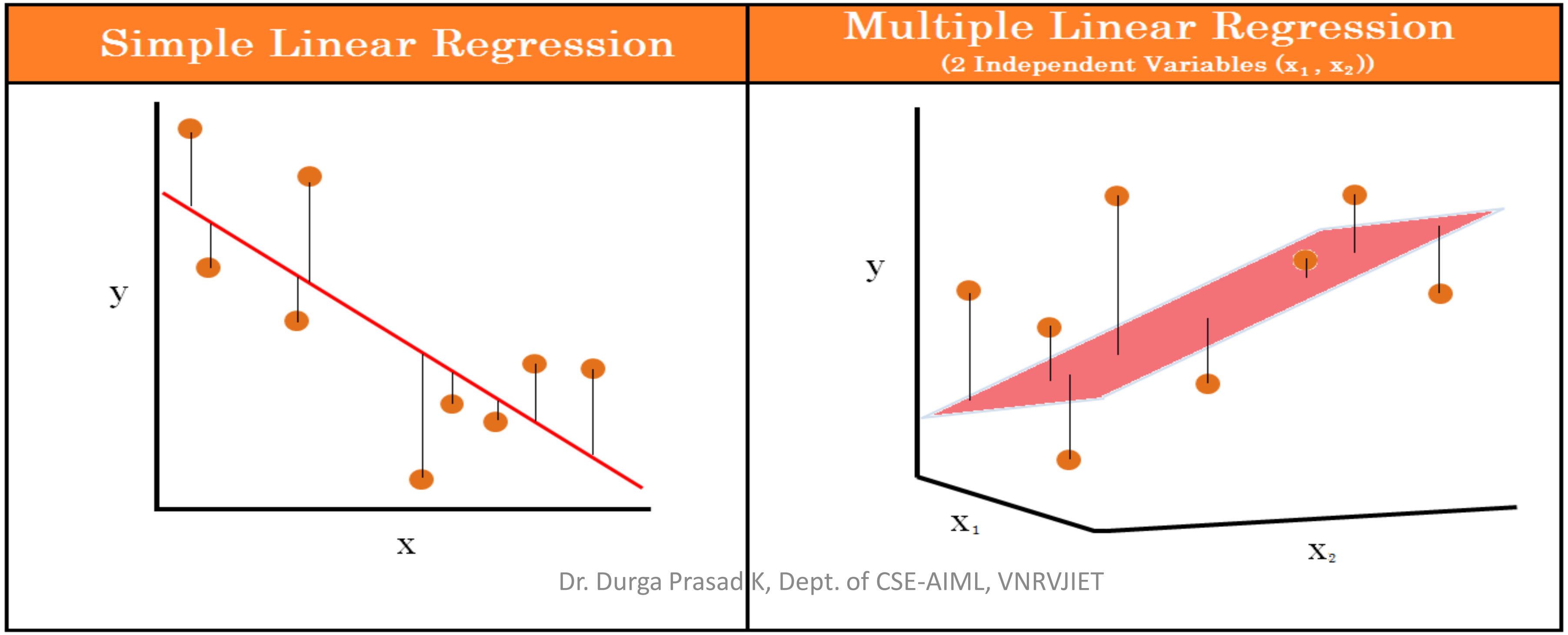
- **Equation for 2 independent variables** $\hat{Y} = a + b_1X_1 + b_2X_2$ or $\boldsymbol{y = a_0 + a_1x_1 + a_2x_2}$

- **X1 and x2**… are independent variables and **Y** is dependent variable

Parameter 'a' is the intercept of this plane. Parameters 'b_1' and 'b_2' are referred to as **partial regression coefficients**. Parameter b_1 represents the change in the mean response corresponding to a unit change in X_1 when X_2 is held constant. Parameter b_2 represents the change in the mean response corresponding to a unit change in X_2 when X_1 is held constant.

Simple Vs Multiple Linear Regression

- In linear regression model we have one dependent and one independent variable.
- Multiple regression model involves multiple predictors or independent variables and one dependent variable.
- This is an extension of the linear regression problem.



Multiple Linear Regression

- The multiple regression of two variables x1 and x2 is given as follows:

$$y = f(x_1, x_2)$$

$$y = a_0 + a_1 x_1 + a_2 x_2$$

- In general, this is given for 'n' independent variables as:

$$y = f(x_1, x_2, \ldots . x_n)$$

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ \ldots \ldots + a_n x_n + \varepsilon$$

- Here, $x_1, x_2, \ldots . x_n$ are predictor variables, y is the dependent variable, $(a_0, a_1, a_2, \ldots . a_n)$ are the coefficients of the regression equation and ε is the error term.

Multiple Linear Regression – Solved Example

- Apply multiple regression for the values given in Table where weekly sales along with sales for products x1 and x2 are provided.
- Use matrix approach for finding multiple regression.

x1 Product 1 Sales	x2 Product 2 Sales	Y Weekly Sales
1	4	1
2	5	6
3	8	8
4	2	12

$$y = a_0 + a_1x_1 + a_2x_2$$

Multiple Linear Regression

- Here, the matrices for Y and X are given as follows:

$$X = \begin{pmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 2 \end{pmatrix} \text{ and } Y = \begin{pmatrix} 1 \\ 6 \\ 8 \\ 12 \end{pmatrix}$$

- The coefficient of the multiple regression equation is given as

$$a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}.$$

x1 Product 1 Sales	x2 Product 2 Sales	Y Weekly Sales
1	4	1
2	5	6
3	8	8
4	2	12

$$y = a_0 + a_1 x_1 + a_2 x_2$$

Multiple Linear Regression

- The regression coefficient for multiple regression is calculated the same way as linear regression:

$$\hat{a} = ((X^T X)^{-1} X^T)Y$$

x1 Product 1 Sales	x2 Product 2 Sales	Y Weekly Sales
1	4	1
2	5	6
3	8	8
4	2	12

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 4 & 5 & 8 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 10 & 19 \\ 10 & 30 & 46 \\ 19 & 46 & 109 \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 2 \end{pmatrix}$$

Multiple Linear Regression

- The regression coefficient for multiple regression is calculated the same way as linear regression:

$$\hat{a} = ((X^T X)^{-1} X^T)Y$$

$$(X^T X)^{-1} = \begin{pmatrix} 4 & 10 & 19 \\ 10 & 30 & 46 \\ 19 & 46 & 109 \end{pmatrix}^{-1} = \begin{pmatrix} 3.15 & -0.59 & -0.30 \\ -0.59 & 0.20 & 0.016 \\ -0.30 & 0.016 & 0.054 \end{pmatrix}$$

x1 Product 1 Sales	x2 Product 2 Sales	Y Weekly Sales
1	4	1
2	5	6
3	8	8
4	2	12

Multiple Linear Regression

- The regression coefficient for multiple regression is calculated the same way as linear regression:

$$\hat{a} = ((X^T X)^{-1} X^T)Y$$

x1 Product 1 Sales	x2 Product 2 Sales	Y Weekly Sales
1	4	1
2	5	6
3	8	8
4	2	12

$$X^T X)^{-1} X^T = \begin{pmatrix} 3.15 & -0.59 & -0.30 \\ -0.59 & 0.20 & 0.016 \\ -0.30 & 0.016 & 0.054 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 4 & 5 & 8 & 2 \end{pmatrix} = \begin{pmatrix} 0.05 & 0.47 & -1.02 & 0.19 \\ -0.32 & -0.098 & 0.155 & 0.26 \\ -0.065 & 0.005 & 0.185 & -0.125 \end{pmatrix}$$

Multiple Linear Regression

- The regression coefficient for multiple regression is calculated the same way as linear regression:

$$\hat{a} = ((X^T X)^{-1} X^T)Y$$

x1 Product 1 Sales	x2 Product 2 Sales	Y Weekly Sales
1	4	1
2	5	6
3	8	8
4	2	12

$$\hat{a} = ((X^T X)^{-1} X^T)Y = \begin{pmatrix} 0.05 & 0.47 & -1.02 & 0.19 \\ -0.32 & -0.098 & 0.155 & 0.26 \\ -0.065 & 0.005 & 0.185 & -0.125 \end{pmatrix} \times \begin{pmatrix} 1 \\ 6 \\ 8 \\ 12 \end{pmatrix} = \begin{pmatrix} -1.69 \\ 3.48 \\ -0.05 \end{pmatrix}$$

Multiple Linear Regression

$$a_0 = -1.69$$
$$a_1 = 3.48$$
$$a_2 = -0.05$$

x1 Product 1 Sales	x2 Product 2 Sales	Y Weekly Sales
1	4	1
2	5	6
3	8	8
4	2	12

- $y = a_0 + a_1x_1 + a_2x_2$
- Hence, the constructed model is:
- $y = -1.69 + 3.48x_1 - 0.05x_2$

*We will get the **y** (dependent variable) value **once substitute** the values of **x1** and **x2** from the **respective table**

Polynomial Regression Model

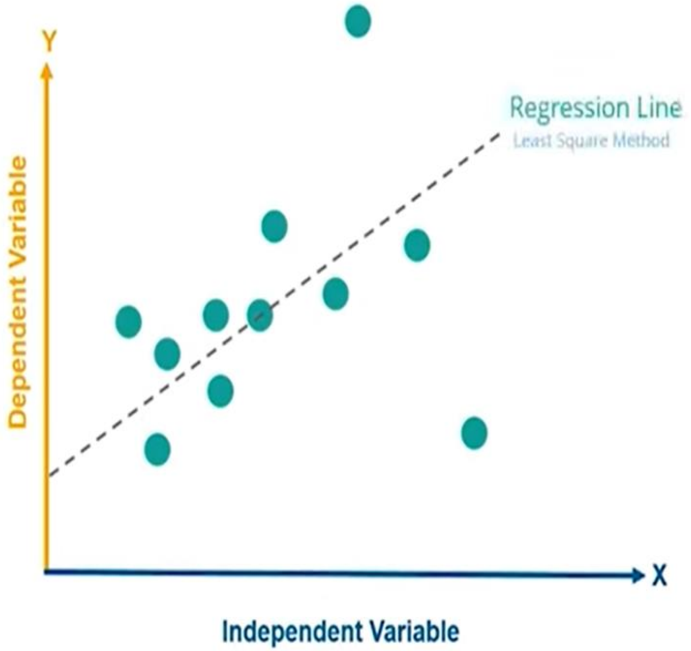
Polynomial Regression Model

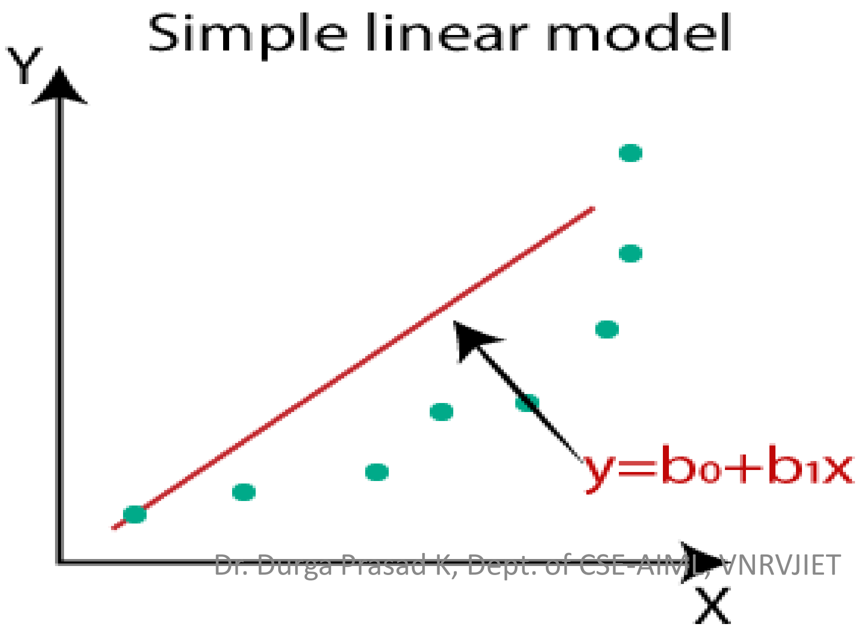
- Polynomial regression is a type of **regression analysis** used in **statistics and machine learning** when the relationship between the **independent variable (input)** and the **dependent variable (output)** **is not linear.**

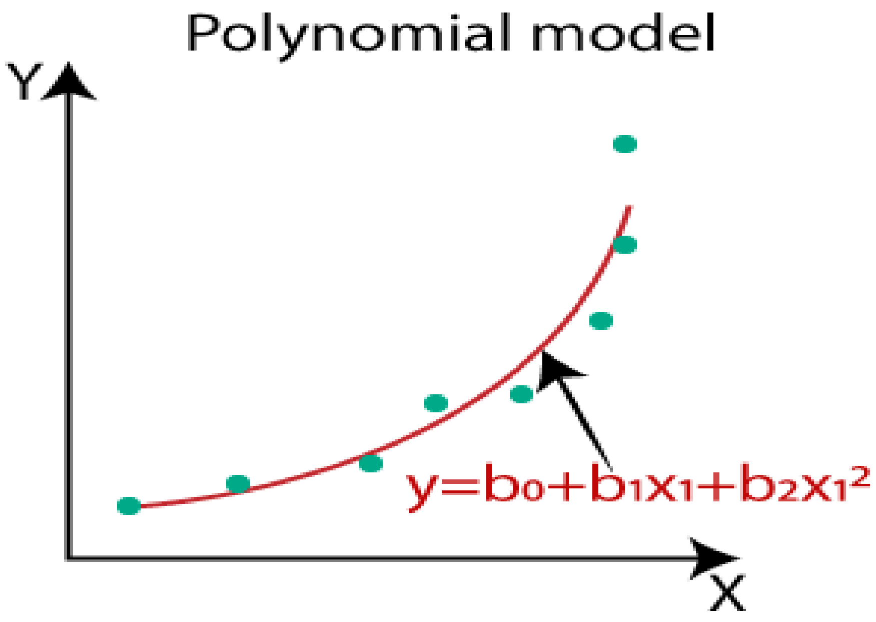
$$y= b_0+b_1x_1+ b_2x_1^2+ b_2x_1^3+\ldots\ldots B_nx_1^n \quad \text{or} \quad f(x) = c_0 + c_1.X^1 + c_2.X^2 + c_3.X^3$$

- Polynomial regression is a **form of Multiple Linear regression** where only due to the **Non-linear relationship** between **dependent** and **independent variables**, we add **some polynomial terms** to **linear regression** to convert it into **Polynomial regression.**

- When the **relationship** between the **variables** is **better** represented by **a curve** rather than **a straight line**, **polynomial regression** can **capture** the **non-linear patterns** in the **data.**

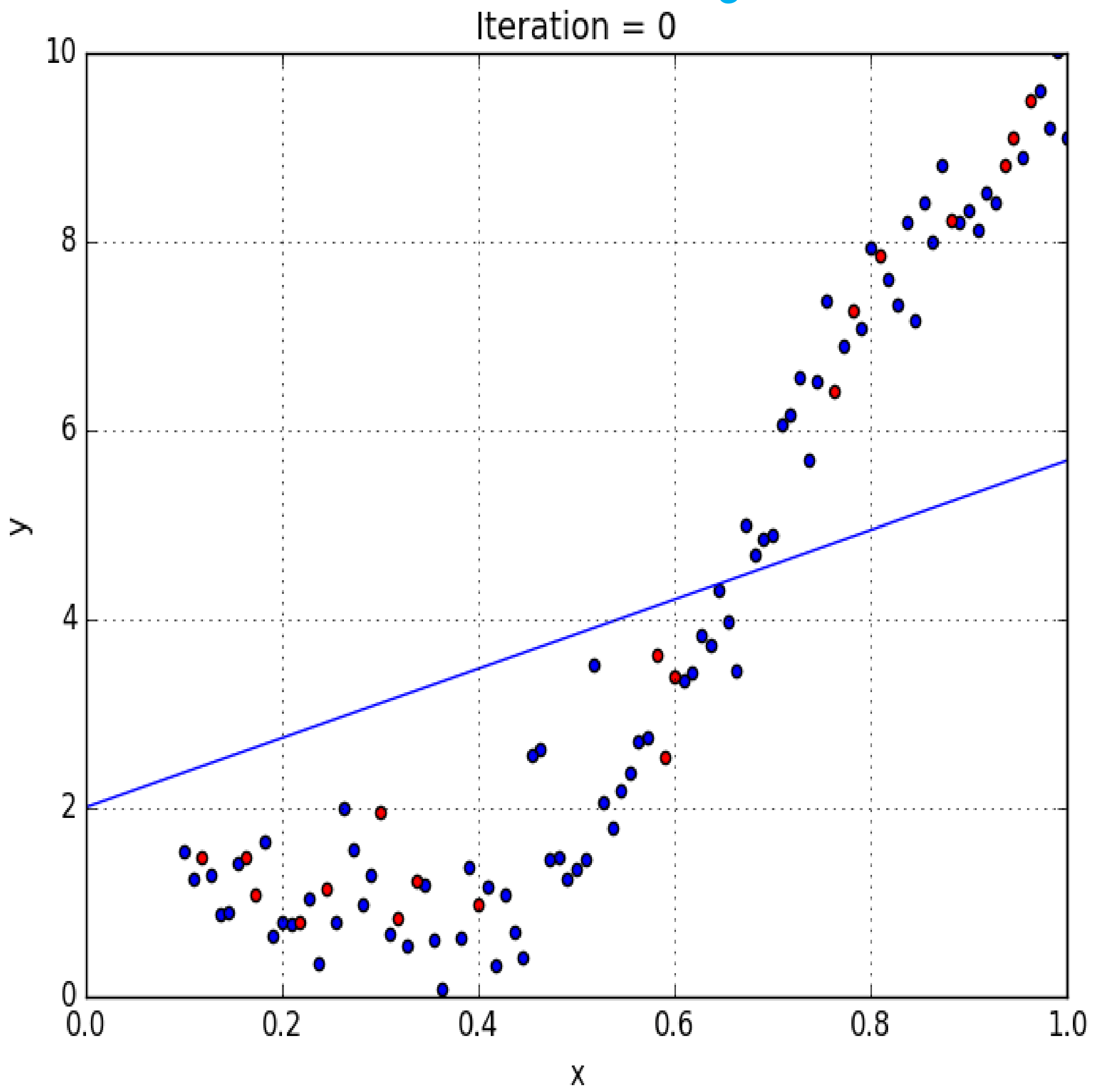


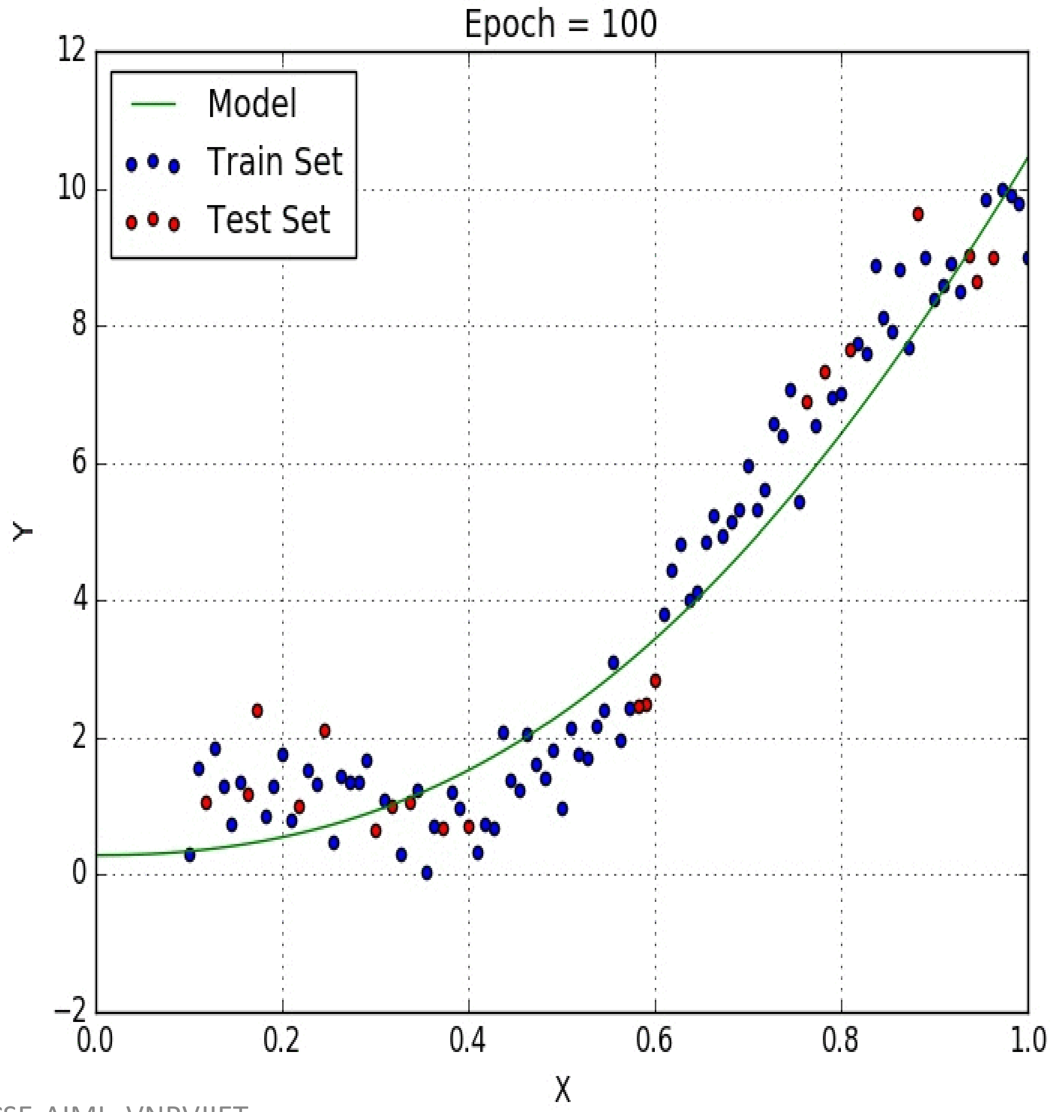




Linear/Multiple linear Regression Vs Polynomial Regression Model

Analysis on Non-Linear Dataset





Polynomial Regression Model

- In polynomial regression, the relationship between the **dependent variable** and the **independent variable is** modeled as **an nth-degree polynomial function**. When the **polynomial** is of **degree 2**, it is called **a quadratic model;** when the degree of a **polynomial** is **3,** it is called a **cubic model**, and so on.

- Consider an example of **input value is 35**, and the **degree of a polynomial is 2**, so I will find **35 power 0**, **35 power 1**, and **35 power 2 ...**this **helps** to **interpret** the **non-linear relationship** in data.

Equation of the Polynomial Regression Model:

Simple Linear Regression	$y = b_0 + b_1 x_1$
Multiple Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$
Polynomial Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$

Types of Polynomial Regression

- A **quadratic equation** is a general term for a **second-degree polynomial equation**. This degree, on the other hand, can go **up to nth values**. Here is the categorization of Polynomial Regression:
- **Linear –** if degree as 1
- **Quadratic –** if degree as 2
- **Cubic –** if degree as 3 and goes on, on the basis of degree.

Polynomials	Form	Degree	Examples
Linear Polynomial	$p(x): ax+b,\ a \neq 0$	Polynomial with Degree 1	$x + 8$
Quadratic Polynomial	$p(x): ax^2+b+c,\ a \neq 0$	Polynomial with Degree 2	$3x^2-4x+7$
Cubic Polynomial	$p(x): ax^3+bx^2+cx,\ a \neq 0$	Polynomial with Degree 3	$2x^3+3x^2+4x+6$

Polynomial Regression Model

- If your **data points** clearly will **not fit a linear regression** (a straight line through all data points), it might be **ideal** for **polynomial regression.**
- **While** simple linear regression models the relationship **as a straight line**, **polynomial regression** allows for **more flexibility** by **fitting** a **polynomial equation to the data.**
- It is a **linear model with some modification** in order to **increase the accuracy.**
- The dataset used in **Polynomial regression** for **training is of non-linear nature**.
- It makes use of a **linear regression model** to **fit** the **complicated** and **non-linear functions** and datasets.
- Hence, "In Polynomial regression, the **original features** are converted into **Polynomial features** of **required degree (2,3,..,n)** and then **modeled using a linear model**."

How to solve **the problem** of **linear/Multiple Linear Regression** for **Non-Linear data**?

- The problem of non-linear regression can be solved by two methods:

1. Transformation of non-linear data to linear data, so that the linear regression can handle the data
2. Using polynomial regression

How to solve **the problem** of **linear/Multiple Linear Regression** for **Non-Linear data**?

*First solution: Applying Transformations (Non-linear data to Linear & vice-versa) and use Linear/Multiple Linear Regression

Transformations

- The trick is to convert non-linear data to linear data that can be handled using the linear regression method.
- Let us consider an exponential function $y = ae^{bx}$.
- The transformation can be done by applying log function to both sides to get:

$$ln(y) = ln(a) + ln(e^{bx})$$

$$ln(y) = ln(a) + bx * ln(e)$$

$$ln(y) = ln(a) + bx$$

How to solve **the problem** of **linear/Multiple Linear Regression** for **Non-Linear data**?

***Second solution: Apply Directly Polynomial Regression**

Polynomial Regression

Example (on Polynomial of 2nd degree)

- Consider the polynomial of 2nd degree.
- The polynomial equation in given by $\boldsymbol{y} = \boldsymbol{a_0} + \boldsymbol{a_1 x} + \boldsymbol{a_2 x^2}$
- The coefficients $\boldsymbol{a_0}, \boldsymbol{a_1}$ $\boldsymbol{and}$ $\boldsymbol{a_2}$ are calculated using the formula,

$$\boldsymbol{a = X^{-1}B}$$

- Where,

$$X = \begin{bmatrix} n & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{bmatrix} \quad B = \begin{bmatrix} \Sigma y_i \\ \Sigma(x_i, y_i) \\ \Sigma(x_i^2, y_i) \end{bmatrix}$$

Polynomial Regression

Example (on Polynomial of 2nd degree)

x	y
1	1
2	4
3	9
4	15

- Consider the polynomial of 2nd degree.
- The polynomial equation in given by $\boldsymbol{y = a_0 + a_1 x + a_2 x^2}$
- The coefficients $\boldsymbol{a_0, a_1\ and\ a_2}$ are calculated using the formula,

$$\boldsymbol{a = X^{-1}B}$$

- Where,

$$X = \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \quad B = \begin{bmatrix} \sum y_i \\ \sum (x_i, y_i) \\ \sum (x_i^2, y_i) \end{bmatrix}$$

Polynomial Regression

Example (on Polynomial of 2nd degree)

x_i	y_i	$x_i y_i$	x_i^2	$x_i^2 y$	x_i^3	x_i^4
1	1	1	1	1	1	1
2	4	8	4	16	8	16
3	9	27	9	81	27	81
4	15	60	16	240	64	256
$\Sigma x_i = 10$	$\Sigma y_i = 29$	$\Sigma x_i y_i = 96$	$\Sigma x_i^2 = 30$	$\Sigma x_i^2 y_i = 338$	$\Sigma x_i^3 = 100$	$\Sigma x_i^4 = 354$

$$X = \begin{bmatrix} n & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{bmatrix} \quad B = \begin{bmatrix} \Sigma y_i \\ \Sigma(x_i, y_i) \\ \Sigma(x_i^2, y_i) \end{bmatrix}$$

Polynomial Regression

x_i	y_i	$x_i y_i$	x_i^2	$x_i^2 y$	x_i^3	x_i^4
1	1	1	1	1	1	1
2	4	8	4	16	8	16
3	9	27	9	81	27	81
4	15	60	16	240	64	256
$\Sigma x_i = 10$	$\Sigma y_i = 29$	$\Sigma x_i y_i = 96$	$\Sigma x_i^2 = 30$	$\Sigma x_i^2 y_i = 338$	$\Sigma x_i^3 = 100$	$\Sigma x_i^4 = 354$

$$a = X^{-1}B \quad X = \begin{bmatrix} n & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{bmatrix} \quad B = \begin{bmatrix} \Sigma y_i \\ \Sigma(x_i, y_i) \\ \Sigma(x_i^2, y_i) \end{bmatrix} \quad \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix}^{-1} \times \begin{bmatrix} 29 \\ 96 \\ 338 \end{bmatrix} = \begin{pmatrix} -0.75 \\ 0.95 \\ 0.75 \end{pmatrix}$$

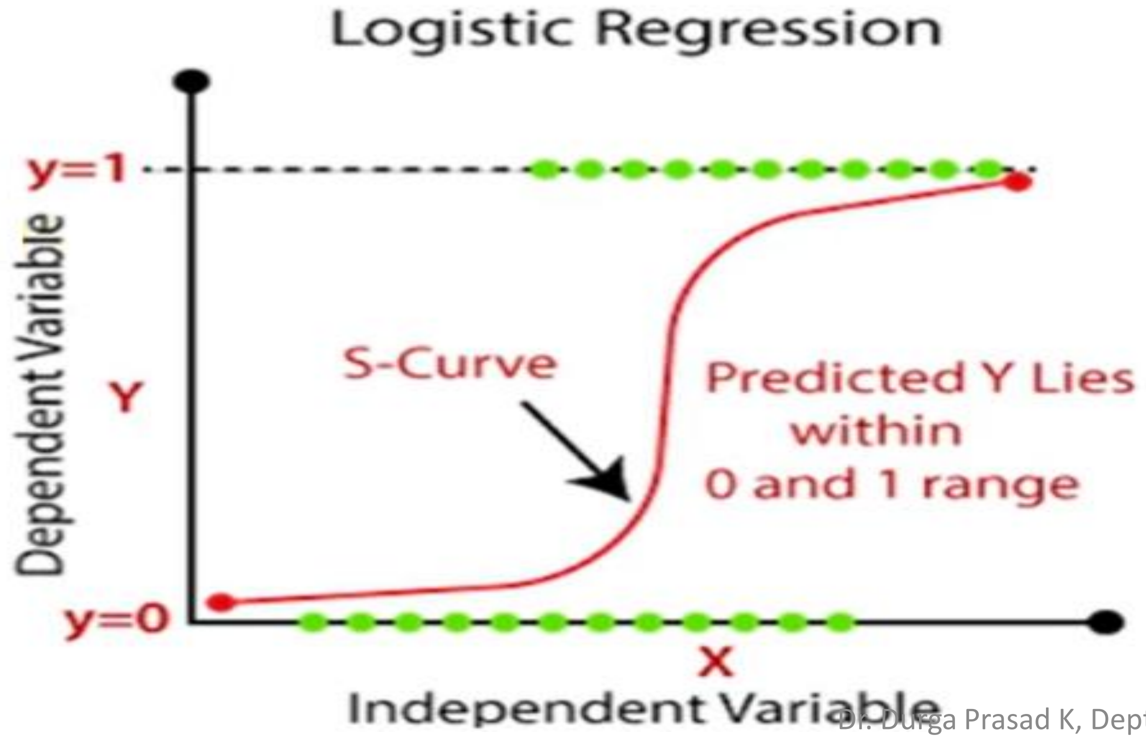
Polynomial Regression

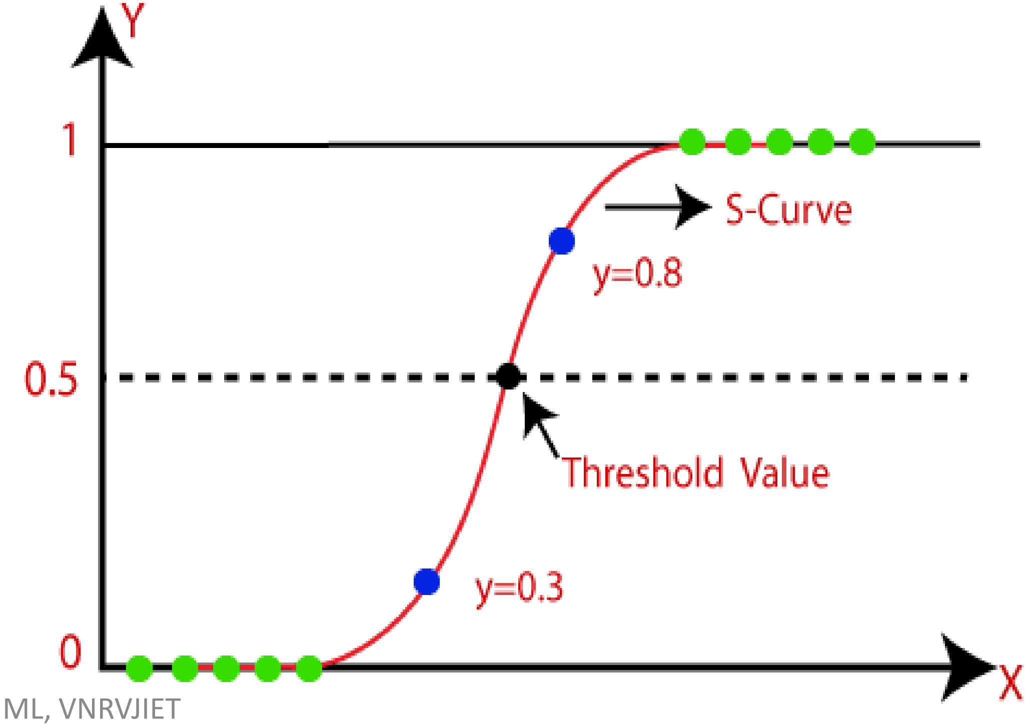
x_i	y_i	$x_i y_i$	x_i^2	$x_i^2 y$	x_i^3	x_i^4
1	1	1	1	1	1	1
2	4	8	4	16	8	16
3	9	27	9	81	27	81
4	15	60	16	240	64	256
$\Sigma x_i = 10$	$\Sigma y_i = 29$	$\Sigma x_i y_i = 96$	$\Sigma x_i^2 = 30$	$\Sigma x_i^2 y_i = 338$	$\Sigma x_i^3 = 100$	$\Sigma x_i^4 = 354$

$$a = X^{-1}B \qquad \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix}^{-1} \times \begin{bmatrix} 29 \\ 96 \\ 338 \end{bmatrix} = \begin{pmatrix} -0.75 \\ 0.95 \\ 0.75 \end{pmatrix} \qquad y = -0.75 + 0.95\,x + 0.75\,x^2$$

Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the **Supervised Learning technique**. It is used for predicting the **categorical dependent variable** using a given set of independent variables (continues).
- Therefore the outcome must be a **categorical or discrete value**. It can be either Yes or No, 0 or 1, true or False, etc.
- But instead of giving the exact value as 0 and 1, **it** gives the **probabilistic values** which **lie between 0 and 1.**





Logistic Regression

- Logistic regression is a process of modeling the probability of a discrete outcome given an input variable (continues). The most common logistic regression models a binary outcome; something that can take two values such as **true/false, yes/no,** and so on.
- Logistic Regression is much similar to the Linear Regression except that **how they are used.** Linear Regression is used for **solving Regression problems**, whereas **Logistic regression is used for solving the classification problems.**
- In Logistic regression, instead of **fitting a regression line**, we fit an **"S" shaped logistic function**.
- Logistic regression is used for **binary classification** where we use **sigmoid function**, that takes input as **independent variables** and **produces a probability value between 0 and 1.**
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a **significant machine learning algorithm** because it has the ability to provide probabilities and classify new data **using continuous** and **discrete datasets.**

Where to apply Linear and Logistic Regression

- Linear regression predicts the numerical response but is not suitable for predicting the categorical variables.
- When categorical variables are involved, it is called classification problem.
- Logistic regression is suitable for binary classification problem.

Where to apply Linear Regression

How Does the Logistic Regression Algorithm Work?

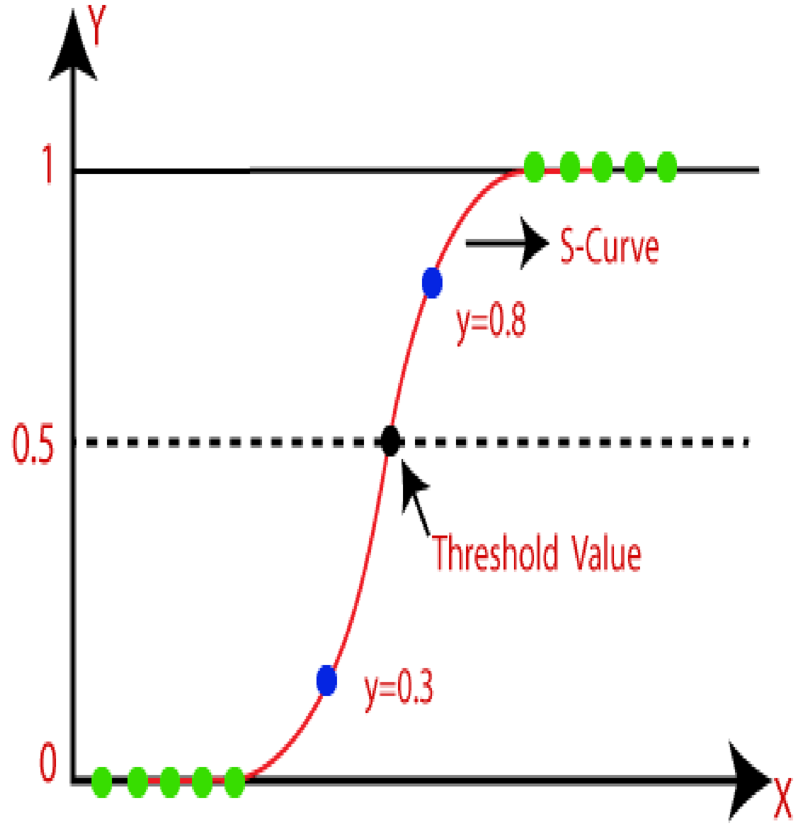
- Consider the following example:
- An organization wants to determine an employee's salary increase based on their performance.
- For this purpose, a linear regression algorithm will help them decide.
- Plotting a regression line by considering the employee's performance as the independent variable, and the salary increase as the dependent variable will make their task easier.

Where to apply Logistic Regression

- Now, what if the organization wants to know whether an employee would get a promotion or not based on their performance?
- The above linear graph won't be suitable in this case.
- As such, we clip the line at zero and one, and convert it into a sigmoid curve (S curve).
- Based on the threshold values, the organization can decide whether an employee will get a salary increase or not.

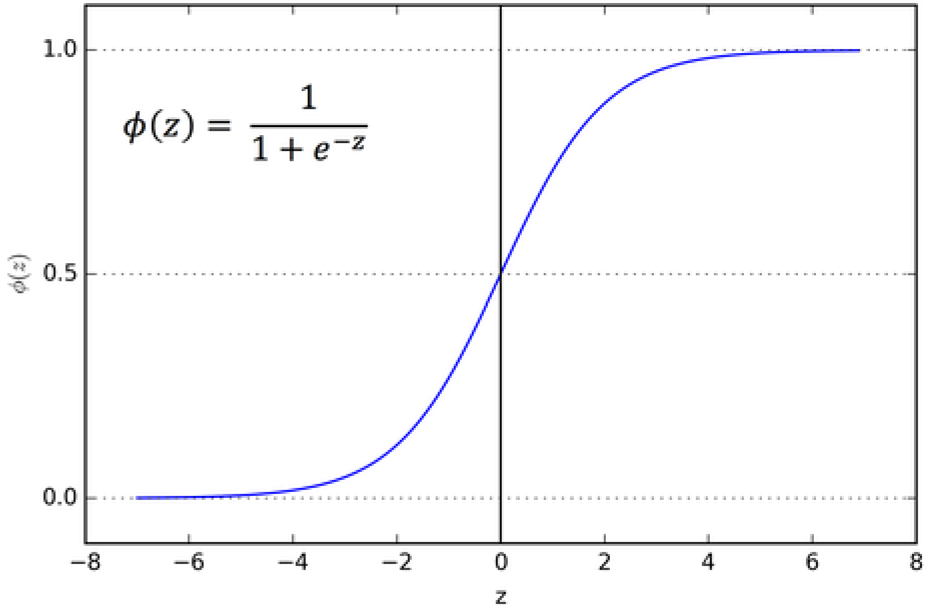
Logistic Function – Sigmoid Function

- The **sigmoid function** is a mathematical function used to map the **predicted values** to **probabilities.**
- It maps any **real value** into **another value** within **a range of 0 and 1**. The value of the **logistic regression** must be between 0 and 1, which cannot go beyond this limit, so **it forms a curve like the "S" form.** The **S-form curve** is called the **Sigmoid function** or the **logistic function.**
- In logistic regression, we use the concept of the **threshold value,** which defines the probability of either 0 or 1.
- Such as values **above** the **threshold value** tends to 1, and a value **below** the threshold values tends to 0.
- For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is **greater than 0.5** (threshold value) then it belongs to Class 1, otherwise, it belongs to Class 0.



Logistic Regression-Sigmoid function

- To understand logistic regression, let’s go over the odds of success.
- $Odds\ (\theta) = \frac{Probability\ of\ an\ event\ happening}{Probability\ of\ an\ event\ not\ happening}$
- $Odds\ (\theta) = \frac{p}{1-p}$
- The values of odds range from **zero to ∞** and the value probability lies between **zero and one**.
- Consider the equation of a straight line:
- $y\ =\ \beta 0\ +\ \beta 1 * \ x$



***Equation for Sigmoid function** $$\sigma(z) = \frac{1}{1-e^{-z}}$$

where the **input** will be **Z** and we find the probability between 0 and 1. i.e. **predicted y.**

***Equation for Probability calculation using Sigmoid function** $$p(x) = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 x)}}$$ **or** $$p = \frac{1}{1+e^{-z}}$$

Logistic Function – Sigmoid Function

- **Odds:** It is the **ratio** of **something occurring** to **something not occurring**. Or The odds of success are defined as the ratio of the **probability of success** over **the probability of failure**.
- it is different from **probability** as the probability is the ratio of something occurring to everything that could possibly occur.
- **Log-odds:** The log-odds, also known as the **logit function**, is the **natural logarithm** of the **odds.**
- In logistic regression, the **log odds** of the dependent variable are modeled as a **linear combination of** the **independent variables** and **the intercept.**

Role of Logistic function:

The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0.

Logistic Regression Function

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y = \beta_0 + \beta_1 X$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = \beta_0 + \beta_1 X$$

Logistic Regression-Example-1

How to predict a student selected or not selected using Logistic Regression from the student dataset ?

- The student dataset has entrance mark based on the historic data of those who are selected or not selected.
- Based on the logistic regression, the values of the learnt parameters are $\beta_0 = 1$ *and* $\beta_1 = 8$.
- Assuming marks of x = 60, compute the resultant class.

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

$$\beta_0 + \beta_1 x = 481$$

$$p(x) = \frac{1}{1 + e^{-481}} = 0.44$$

- If we assume the threshold value as 0.5, then it is observed that 0.44 < 0.5, therefore, the candidate with marks 60 is not selected.

Logistic Regression-Example-2

- The dataset of pass or fail in an exam of 5 students is given in the table.
- Use logistic regression as classifier to answer the following questions.

1. Calculate the probability of pass for the student who studied 33 hours.
2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%.

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

Assume the model suggested by the optimizer for odds of passing the course is,

$$\log(odds) = -64 + 2 * hours$$

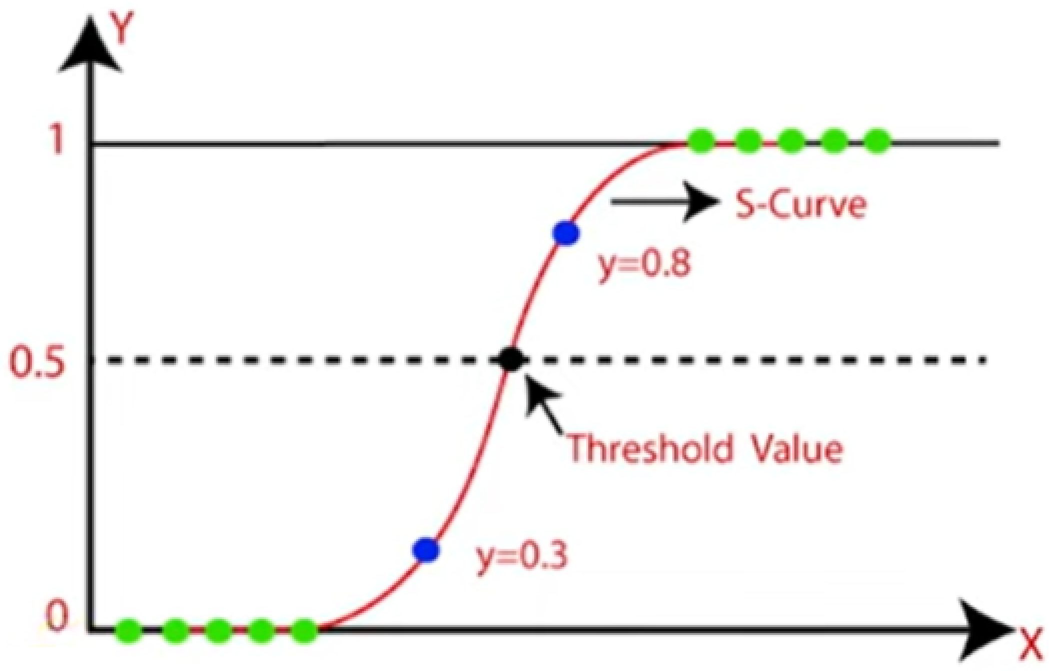
***Odds:** It is the ratio of **something occurring** to **something not occurring.**

***Log-odds:** The log-odds, also known as the **logit function**, is the <u>**natural logarithm**</u> of the **odds.**

Logistic Regression-Example

- We use Sigmoid Function in logistic regression

- $s(x) = \frac{1}{1+e^{-x}}$



Logistic Regression-Example

1. Calculate the probability of pass for the student who studied 33 hours.

$$s(x) = \frac{1}{1+e^{-x}}$$

- $p = \frac{1}{1+e^{-z}}$
- $z = -64 + 2 * 33 = -64 + 66 = 2$
- $p = \frac{1}{1+e^{-2}} = 0.88$
- That is, if student studies 33 hours, then there is **88% chance** that the student will pass the exam

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

$$\log(odds) = z = -64 + 2 * hours$$

***Odds:** It is the ratio of **something occurring** to **something not occurring.**

***Log-odds:** The log-odds, also known as the **logit function**, is the **natural logarithm** of the **odds.**

Logistic Regression-Example

2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%.

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

- $\boldsymbol{p = \frac{1}{1+e^{-z}} = 0.95}$
- $0.95 * (1 + e^{-z}) = 1$
- $0.95 * e^{-z} = 1 - 0.95$
- $e^{-z} = \frac{0.05}{0.95} = 0.0526$
- $\ln(e^{-z}) = \ln(0.0526)$

$\ln(e^{x}) = x$

$-z = ln(0.0526) = -2.94$

$\boldsymbol{z = 2.94}$

Logistic Regression-Example

- $z = 2.94$
- $\log(odds) = z = -64 + 2 * hours$
- $2.94 = -64 + 2 * hours$
- $2 * hours = 2.94 + 64$
- $2 * hours = 66.94$
- $hours = \frac{66.94}{2}$
- $\boldsymbol{hours = 33.47\ Hours}$

Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

- The student should study **at least 33.47 hours**, so that he will pass the exam with more than 95% probability

Maximum Likelihood Estimation (MLE)

- The Maximum Likelihood Estimation (MLE) is a method of **estimating** the **parameters** of a **logistic regression model.** This estimation method is one of the most widely used.
- The **coefficients** in a logistic regression are estimated using a process called Maximum Likelihood Estimation (MLE).
- **MLE** is about **predicting the value** for the **parameters** that **maximizes** the **likelihood function**
- The method of maximum likelihood selects **the set of values of the model parameters** that **maximize** the likelihood function.
- It involves **maximizing** a **likelihood function** in order to find the **probability distribution** and **parameters** that **best explain** the **observed data.**
- It provides a **framework for predictive modeling** in **machine learning** where **finding model parameters** can **be framed** as an **optimization problem.**
- The **likelihood function** is the **probability** that the **observed values of the dependent variable** may be **predicted** from the **observed values of the independent variables**. The likelihood varies from **0 to 1.**

Maximum Likelihood Estimation (MLE)

- Maximum Likelihood Estimation (MLE) used to estimate the coefficient in logistic regression
- A coin flips problem, outcome equally heads and tails of the same number of times.
- If we toss the coin 10 times, it is expected that we get five times Head and five times Tail.
- Let us now discuss about the probability of getting only Head as an outcome;
- it is 5/10 = 0.5 in the above case.
- If P is greater than 0.5, it is in favour of Head,
- P is lesser than 0.5, it is against the Head.

Maximum Likelihood Estimation (MLE)

- Let us represent 'n' flips of coin as X1 , X2 , X 3,..., Xn .
- Now Xi can take the value of 1 or 0.
- Xi = 1 if Head is the outcome
- Xi = 0 if Tail is the outcome
- When we use the Bernoulli distribution represents each flip of the coin:

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

Note: ' Θ ' refers to the **parameter space** i.e., the **range of values** the **unknown parameter 'θ' can take**.

For our case, since p indicates the probability that the coin lands as heads, p is bounded between 0 and 1.

Hence, Θ = [0, 1].

Maximum Likelihood Estimation (MLE)

- Each observation X is independent and identically distributed (iid), and the joint distribution simplifies to a product of distributions.

$$f(x_1, \ldots, x_n|\theta)\,{}_{i=1}^{n} f(x_i|\theta) = \theta^{x_1}(1-\theta)^{1-x_1} \ldots \theta^{x_n}(1-\theta)^{1-x_n} = \theta^{\#H}(1-\theta)^{n-\#H},$$

- where #H is the number of flips that resulted in the expected outcome (heads in this case).
- The likelihood equation is

$$L(\theta|x) = {}_{i=1}^{n} f(x_i|\theta)$$

Note: ‘ Θ ‘ refers to the **parameter space** i.e., the **range of values** the **unknown parameter θ can take**.

For our case, since p indicates the probability that the coin lands as heads, p is bounded between 0 and 1.

Hence, Θ = [0, 1].

Maximum Likelihood Estimation (MLE)

- But the likelihood function is not a probability.
- The likelihood for some coins may be 0.25 or 0 or 1.
- MLE is about predicting the value for the parameters that maximizes the likelihood function.

$$\log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Assumptions in Regression Analysis

1.The dependent variable (Y) can be calculated / predicated as a linear function of a specific set of independent variables (X’s) plus an error term (ε).

2. The number of observations (n) is greater than the number of parameters (k) to be estimated, i.e. n > k.

3. Relationships determined by regression are only relationships of association based on the data set and not necessarily of cause and effect of the defined class.

4. Regression line can be valid only over a limited range of data. If the line is extended (outside the range of extrapolation), it may only lead to wrong predictions.

Assumptions in Regression Analysis

5. The values of the error (ε) are independent and are not related to any values of X. This means that there are no relationships between a particular X, Y that are related to another specific value of X, Y.

6. The error term (ε) is normally distributed. This also means that the mean of the error (ε) has an expected value of **0.**

7. Variance is the same for all values of X (homoscedasticity/ homogeneity of variance).

8. If the business conditions change and the business assumptions underlying the regression model are no longer valid, then the past data set will no longer be able to predict future trends.

Improving Accuracy of the Linear Regression Model

General scenario for evaluation of **any ML model:**

Based on **Bias and Variance:**

- The concept of **bias** and **variance** is similar to accuracy and prediction.
- **Accuracy** refers to how close **the estimation** is **near the actual value**, where as **prediction** refers to continuous estimation of the value.

- **High bias** = **low accuracy** (not close to real value)
- **Low bias** = **high accuracy** (close to real value)
- **High variance** = **low prediction** (values are scattered)
- **Low variance** = **high prediction** (values are close to each other)

*[Reference from saikat Dutt ML text book]

Improving Accuracy of the Linear Regression Model

- A regression model is said to be good which is **highly accurate and highly predictive.** Therefore, the **overall error** of **our model will be low**, implying a **low bias (high accuracy)** and **low variance (high prediction)**. This is highly preferable.

- Similarly, we can say that if the **variance increases** (low prediction), the **spread of our data points increases**, which results in **less accurate prediction**. As the **bias increases** **(low accuracy),** **the error** between our predicted value and the observed values **increases.**

- Therefore, **balancing out** **bias** and **accuracy** is **essential** in a regression model.

Improving Accuracy of the Linear Regression Model

FOR IMPROVEMENT in Linear Regression:

- In the linear regression model, it is **assumed** that the number of observations (n) is **greater than** the number of parameters (k) to be estimated, i.e. n > k, and in that case, the least squares estimates tend to have **low variance** and hence will perform **well on test observations.**

- However, if observations (n) is **not much larger than** parameters (k), then there can be **high variability** in the least squares fit, resulting in **over-fitting** and leading to **poor predictions.**

- If k > n, then linear regression is **not usable.** This also indicates **infinite variance**, and so, the **method cannot be used at all.**

Improving Accuracy of the Linear Regression Model

- Accuracy of linear regression can be improved using the following three methods:

1. SHRINKAGE APPROACH

2. SUBSET SELECTION

3. DIMENSIONALITY (VARIABLE) REDUCTION

Improving Accuracy of the Linear Regression Model

1. SHRINKAGE (REGULARIZATION) APPROACH:

- **By limiting (shrinking)** the **estimated coefficients**, we can try to **reduce** the **variance** at the cost of a negligible increase in bias.
- This can in turn lead to **substantial improvements** in the **accuracy** of the model.
- **Few variables** used in the multiple regression model are in fact not associated with the **overall response** and are called as **irrelevant variables**; this may lead to **unnecessary complexity** in the **regression model.**
- This approach involves fitting a model involving all predictors. However, the **estimated coefficients** are **shrunken towards zero** relative to the least squares estimates. This shrinkage (also known as **regularization)** has the effect of **reducing the overall variance.**

Improving Accuracy of the Linear Regression Model

1. SHRINKAGE (REGULARIZATION) APPROACH:

- The **two** best-known techniques for shrinking the regression **coefficients towards zero are**

1. ridge regression
2. lasso (Least Absolute Shrinkage Selector Operator)

Improving Accuracy of the Linear Regression Model

1. **SHRINKAGE (REGULARIZATION) APPROACH:**
   1. ridge regression :

- Ridge regression performs **L2 regularization**, i.e. it adds penalty equivalent to square of the magnitude of coefficients.

Minimization objective of ridge = LS Obj + α × (sum of square of coefficients)

- ridge regression **works best** in situations where the **least squares** estimates have **high variance.**

- **One disadvantage** with **ridge regression** is that it will include **all k predictors in the final model**

Improving Accuracy of the Linear Regression Model

Shrinkage (Regularization) approach:

2. lasso (Least Absolute Shrinkage Selector Operator)

Lasso regression performs **L1 regularization,** i.e. it adds penalty equivalent to the absolute value of the magnitude of coefficients.

Minimization objective of ridge = LS Obj + α × (absolute value of the magnitude of coefficients)

- The lasso overcomes the disadvantage (from ridge regression) by forcing some of the coefficients to zero value.
- The lasso can be expected to **perform better** in a setting where a relatively **small number** of **predictors** have substantial coefficients

Improving Accuracy of the Linear Regression Model

2. SUBSET SELECTION:

Identify a subset of the predictors that is assumed to be related to the response and then fit a model

*There are two methods in which subset of the regression can be selected:

1. **Best subset selection (considers all the possible (2 power k))**

   In best subset selection, we fit a separate least squares regression for each possible subset of the k predictors.

2. **Stepwise subset selection**

   1. Forward stepwise selection (0 to k)
   2. Backward stepwise selection (k to 0)

Improving Accuracy of the Linear Regression Model

2. SUBSET SELECTION:

Stepwise subset selection:

1. Forward stepwise selection (0 to k)

- Forward stepwise selection is a computationally efficient alternative to best subset selection
- predictors are added one by one **to** the model, until all the k predictors are included in the model. In particular, at each step, the variable (X) that gives the highest additional improvement to the fit is added.

2. Backward stepwise selection (k to 0)

Backward stepwise selection begins with **the least squares model** which contains all k predictors and then iteratively **removes** the **least useful predictor one by one.**

Improving Accuracy of the Linear Regression Model

3. DIMENSIONALITY REDUCTION (VARIABLE REDUCTION)

- The earlier methods, namely subset selection and shrinkage, control variance either by using a subset of the original variables or by shrinking their coefficients towards zero.
- In dimensionality reduction, predictors (X) are transformed, and the model is set up using the transformed variables after dimensionality reduction.
- The number of variables is reduced using the dimensionality reduction method. **Principal component analysis** is one of the **most important dimensionality (variable) reduction** techniques.