Regression

What is Regression Analysis?

- ✓ Regression analysis is a form of ***predictive modelling technique*** which investigates the relationship between a dependent (target) and independent variable**(s)** (predictor).
- ✓ **Predicting:** This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.
- ✓ For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Why we need Regression Analysis?

Typically, a regression analysis is used for these purposes:

(1) Prediction of the target variable (forecasting).

(2) Modelling the relationships between the dependent variable and the explanatory variable.

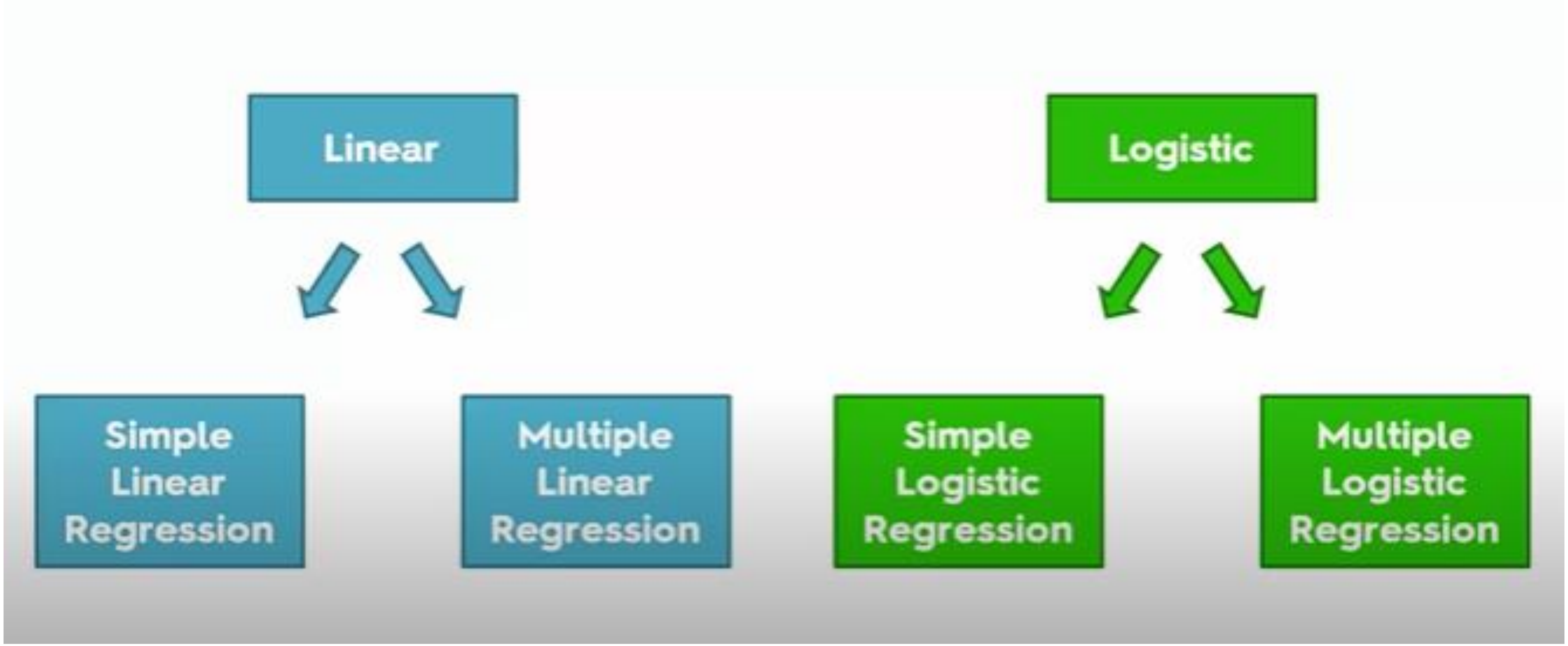
(3) Testing of hypotheses.

Benefits

1. It indicates the strength of impact of multiple independent variables on a dependent variable.
2. It indicates the significant relationships between dependent variable and independent variable.

These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

Types



Linear
Logistic
Simple Linear Regression
Multiple Linear Regression
Simple Logistic Regression
Multiple Logistic Regression

Regression

- Regression is a statistical measurement that attempts to determine the strength of the relationship between a dependent variable and a series of independent variables.

Linear regression

Linear regression always uses a linear equation, Y = a +bx, where x is the explanatory variable and Y is the dependent variable.

Multi Linear regression

- In multiple linear regression, multiple equations are added together but the parameters are still linear.

Non-linear regression

If the model equation does not follow the Y = a +bx form then the relationship between the dependent and independent variables will not be linear.

Regression

- **Simple linear regression** relates two variables (X and Y) with a **straight line ($y = mx + b$)**

- **Nonlinear regression** relates the two variables in a nonlinear ***(curved)*** relationship.

Types (x= studying, y=grade)

Simple Linear Regression

$$y = b_0 + b_1{}^*x_1$$

Multiple Linear Regression

$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 + \ldots + b_n{}^*x_n$$

Types of Regression Analysis

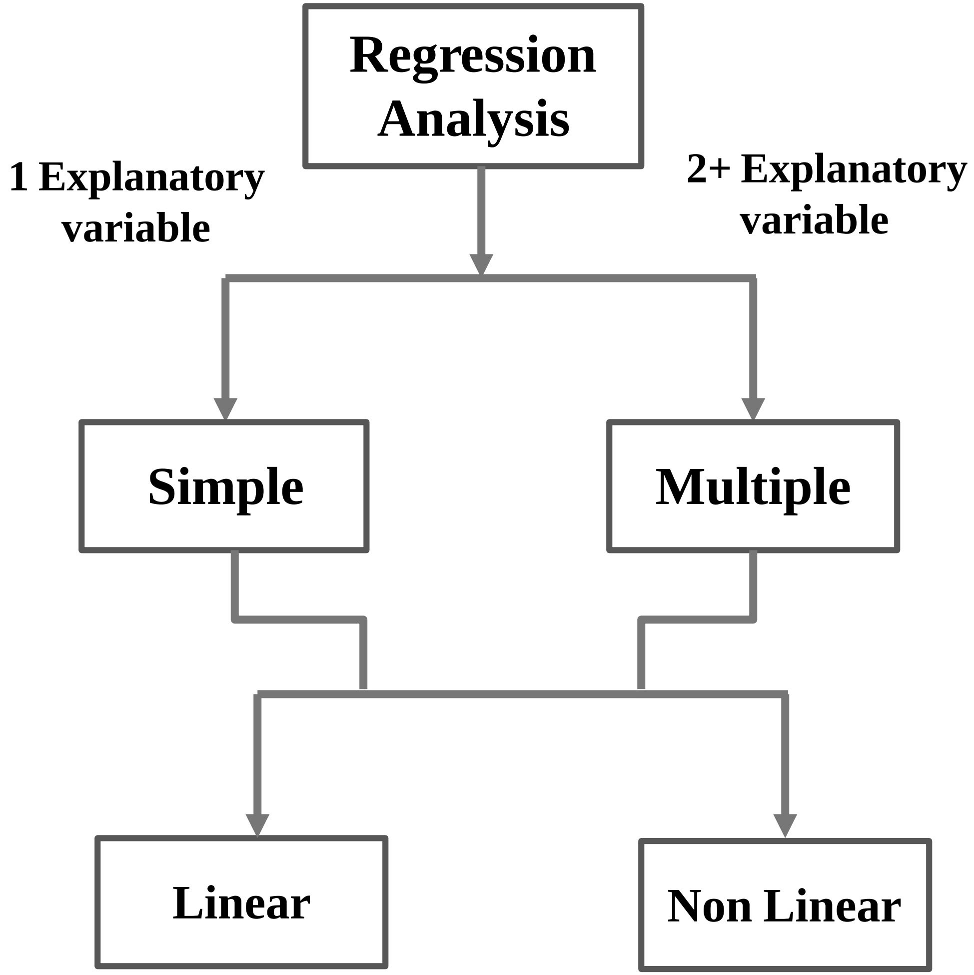
Types of regression analysis:

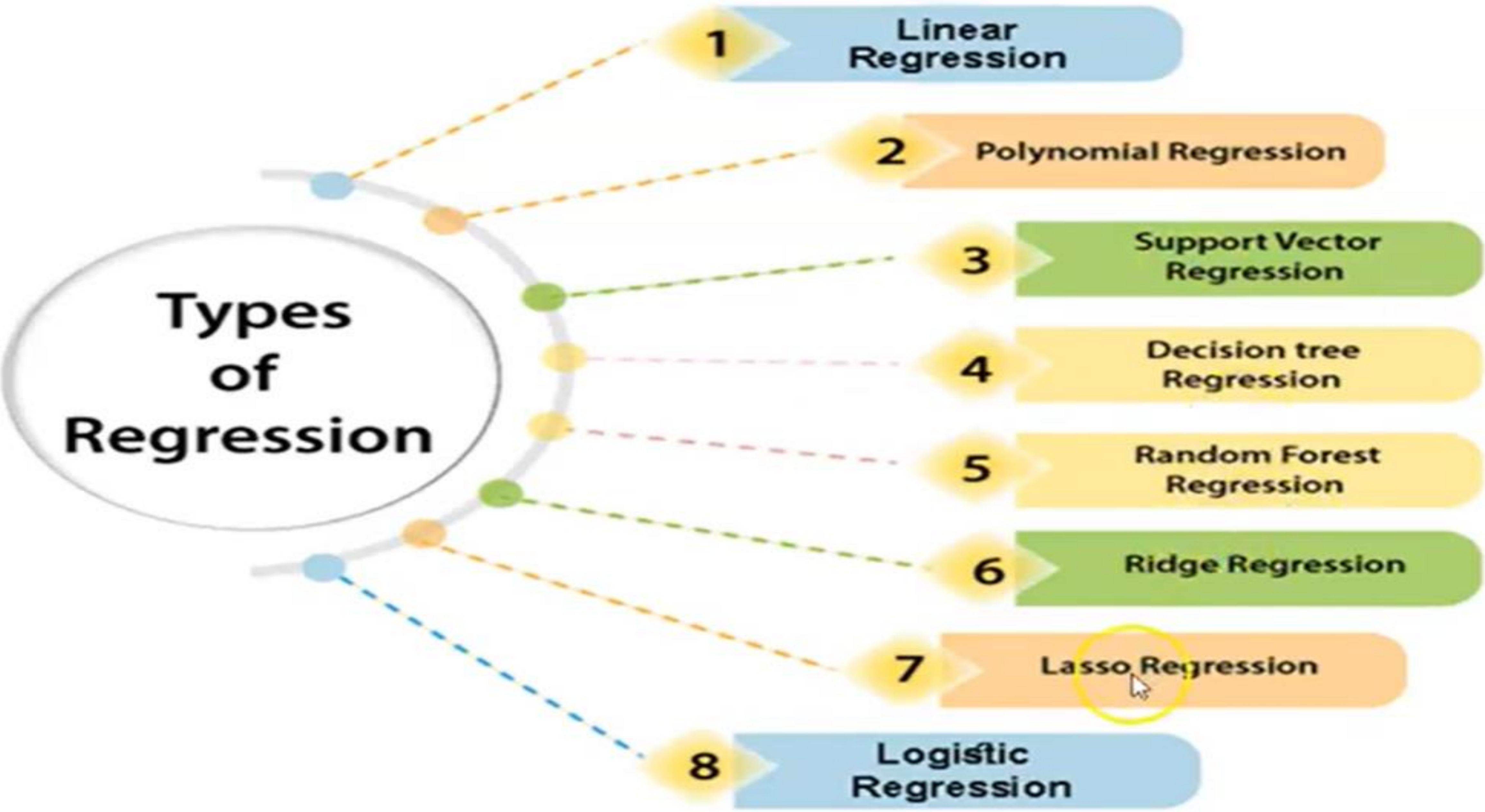
Regression analysis is generally classified into two kinds: simple and multiple.

Simple Regression:

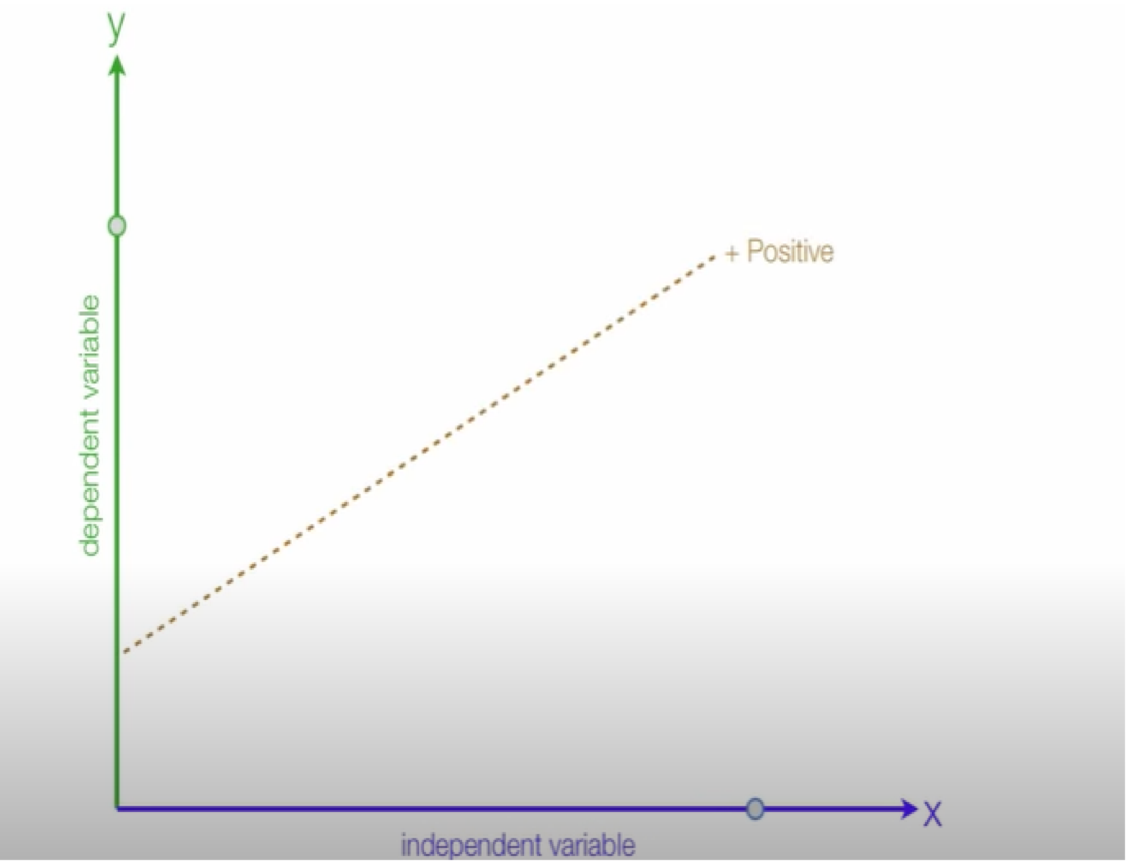
It involves only two variables: dependent variable , explanatory (independent) variable.

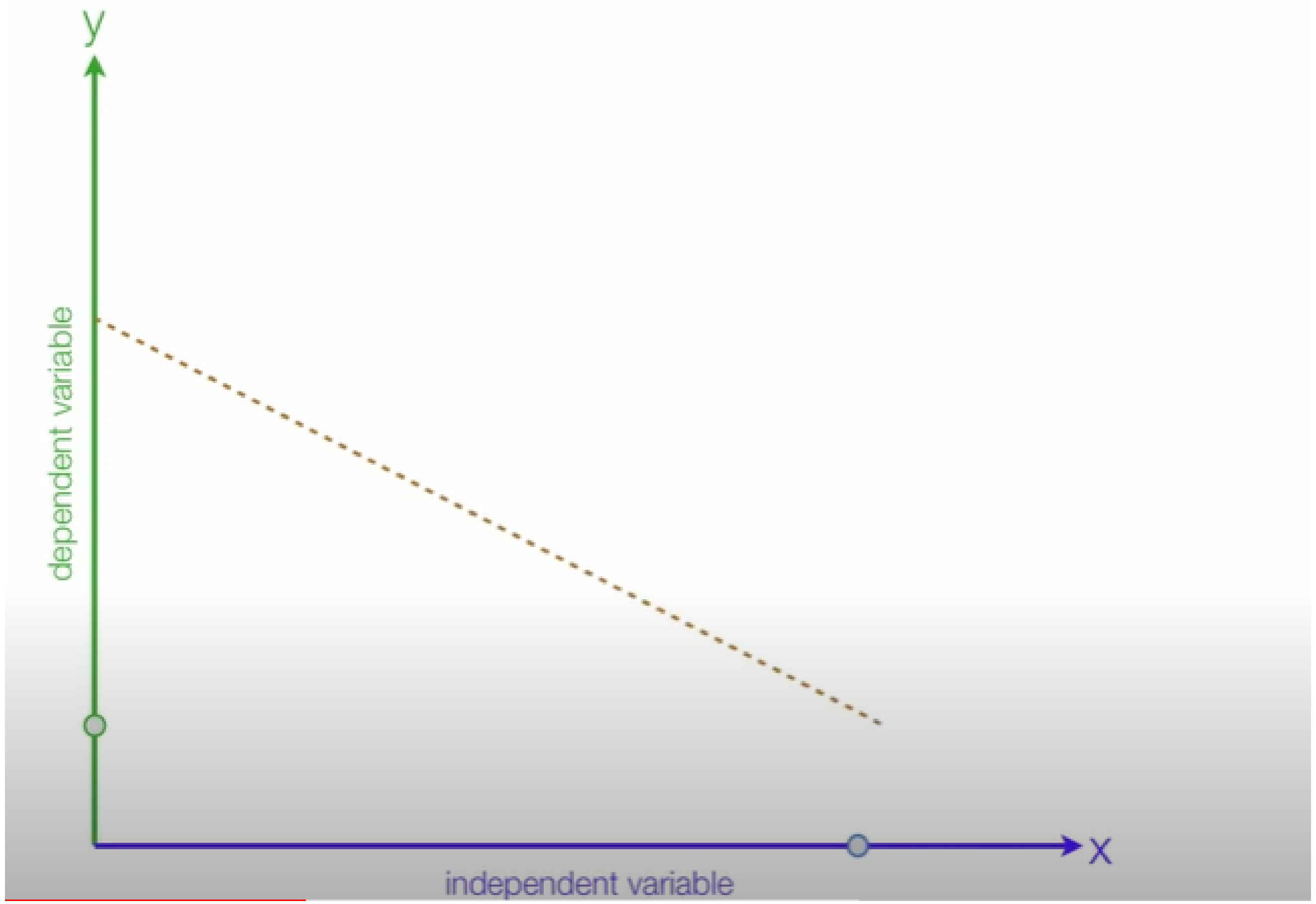
A regression analysis may involve a **linear** model or a **nonlinear** model.
The term linear can be interpreted in two different ways:
1. Linear in variable
2. Linearity in the parameter




Types of Regression
1
Linear Regression
2
Polynomial Regression
3
Support Vector Regression
4
Decision tree Regression
5
Random Forest Regression
6
Ridge Regression
7
Lasso Regression
8
Logistic Regression

Positive Relationship (x= Studying , y=Grades)
Negative Relationship (x= Instagram y= Grades)





Simple Linear Regression

- Least Square "Linear Regression" is a statistical method to regress the data with dependent variable having continuous values whereas independent variables can have either continuous or categorical values.
- **In other words "Linear Regression" is a method to predict dependent variable (Y) based on values of independent variables (X).**
- It can be used for the cases where we want to predict some continuous quantity.
- E.g., Predicting traffic in a retail store, predicting a user's dwell time or number of pages visited on a website, etc.,

Prerequisites

- To start with Linear Regression, you must be aware of a few basic concepts of statistics. i.e.,
- Correlation (r) – Explains the relationship between two variables, possible values -1 to +1
- Variance (σ^2)– Measure of spread in your data
- Standard Deviation (σ) – Measure of spread in your data (Square root of Variance)
- Normal distribution
- Residual (error term) – {Actual value – Predicted value}

Assumptions of Linear Regression

- Not a single size fits or all, the same is true for Linear Regression as well. In order to fit a linear regression line data should satisfy few basic but important assumptions. If your data doesn’t follow the assumptions, your results may be wrong as well as misleading.

- **Linearity & Additive**: There should be a linear relationship between dependent and independent variables and the impact of change in independent variable values should have additive impact on dependent variable.

Assumptions of Linear Regression

- **Normality of error distribution:** Distribution of differences between Actual & Predicted values (Residuals) should be normally distributed.
- **Homoscedasticity:** Variance of errors should be constant versus,
  - Time
  - The predictions
  - Independent variable values
- **Statistical independence of errors:** The error terms (residuals) should not have any correlation among themselves. E.g., In case of time series data there shouldn't be any correlation between consecutive error terms

Finding a Linear Regression Line

x	y	Predicted 'y'
1	2	B_0+B_1*1
2	1	B_0+B_1*2
3	3	B_0+B_1*3
4	6	B_0+B_1*4
5	9	B_0+B_1*5
6	11	B_0+B_1*6
7	13	B_0+B_1*7
8	15	B_0+B_1*8
9	17	B_0+B_1*9
10	20	B_0+B_1*10

Where,

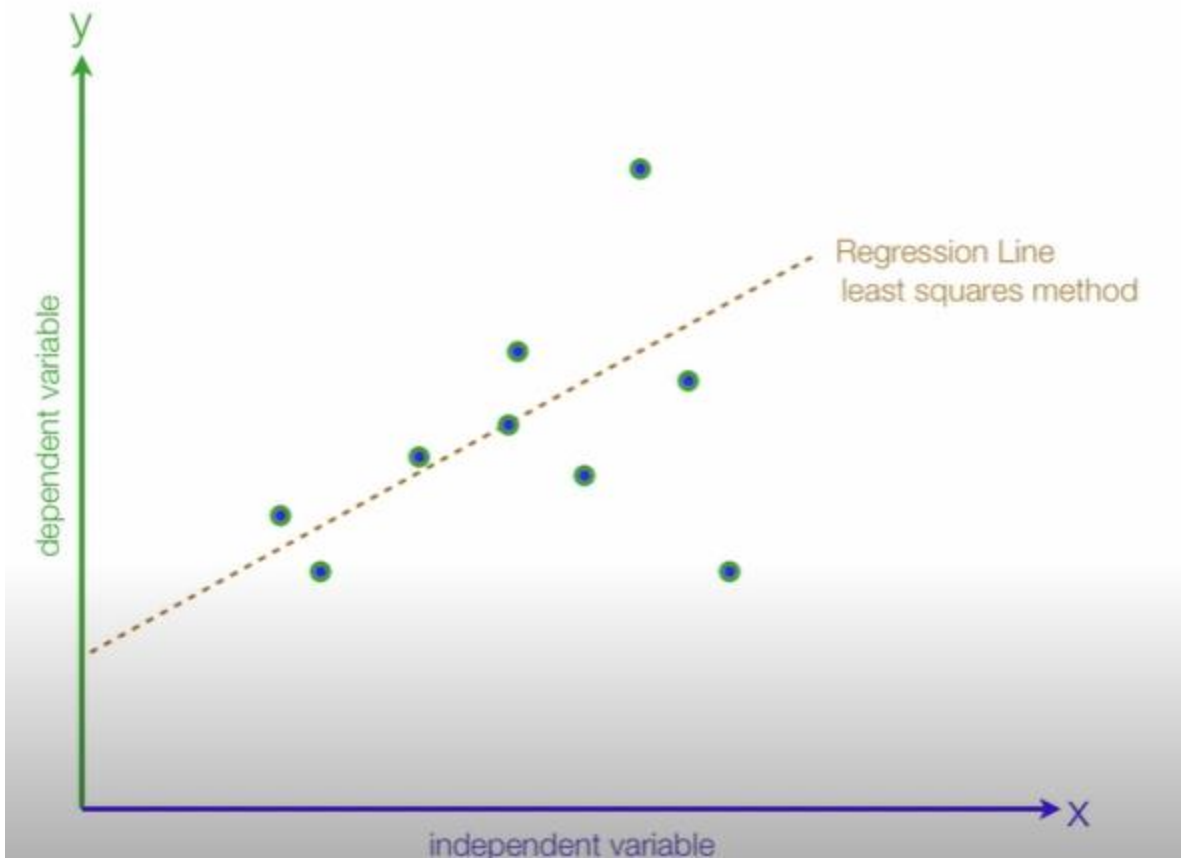
Table 1:

Std. Dev. of x	3.02765
Std. Dev. of y	6.617317
Mean of x	5.5
Mean of y	9.7
Correlation between x & y	.989938

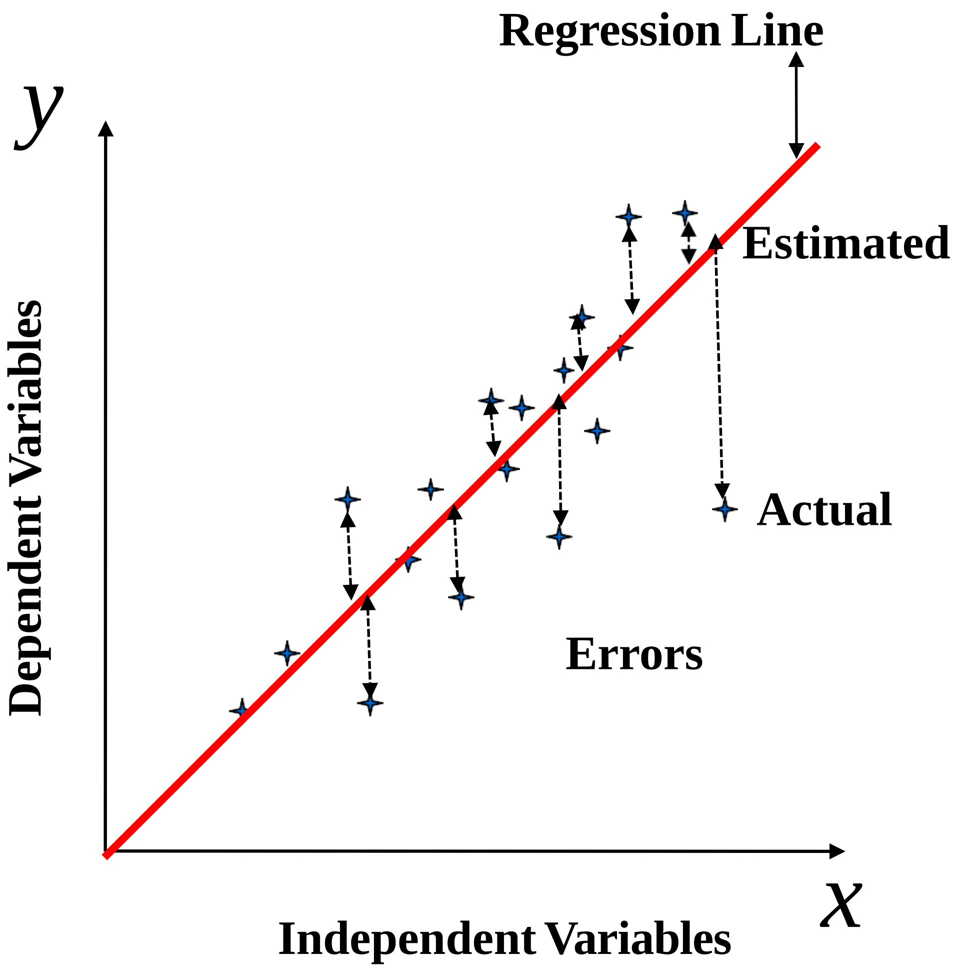
Correlation

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

Population Regression Line

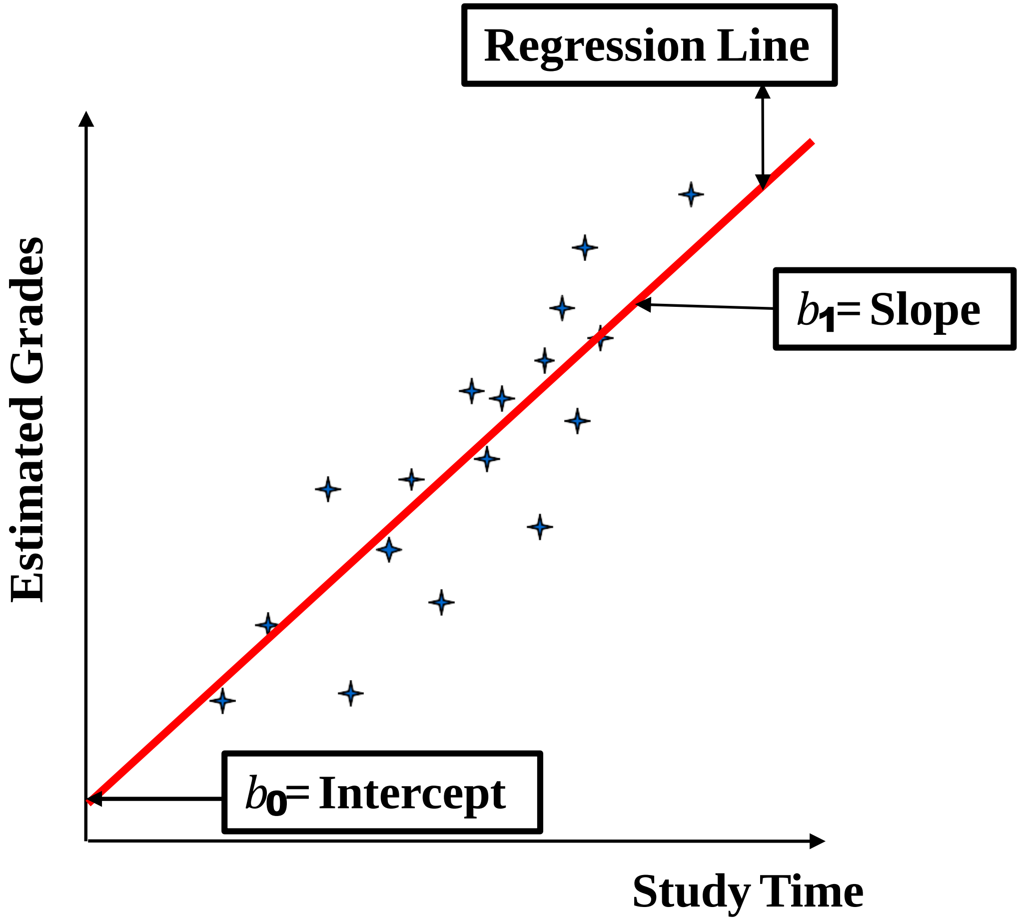


Population Regression Line



Population Regression Line

Example

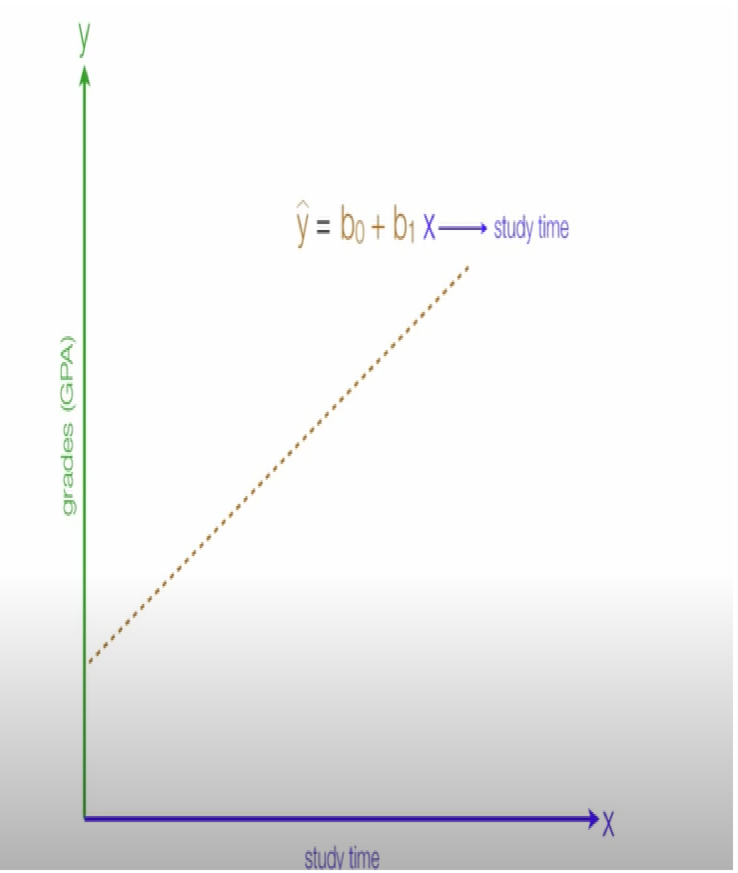


Population regression function =

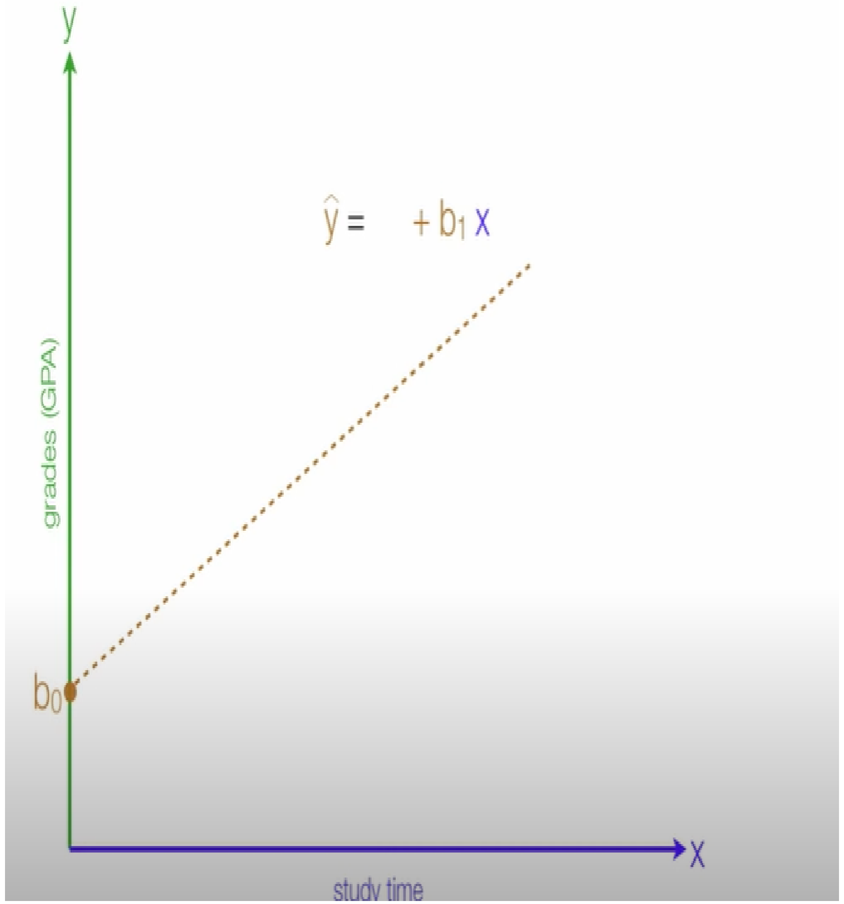
$$y = b_0 + b_1 x$$

y = **Estimated Grades**
x = **Study Time**
b_0= **Intercept**
b_1= **Slope**

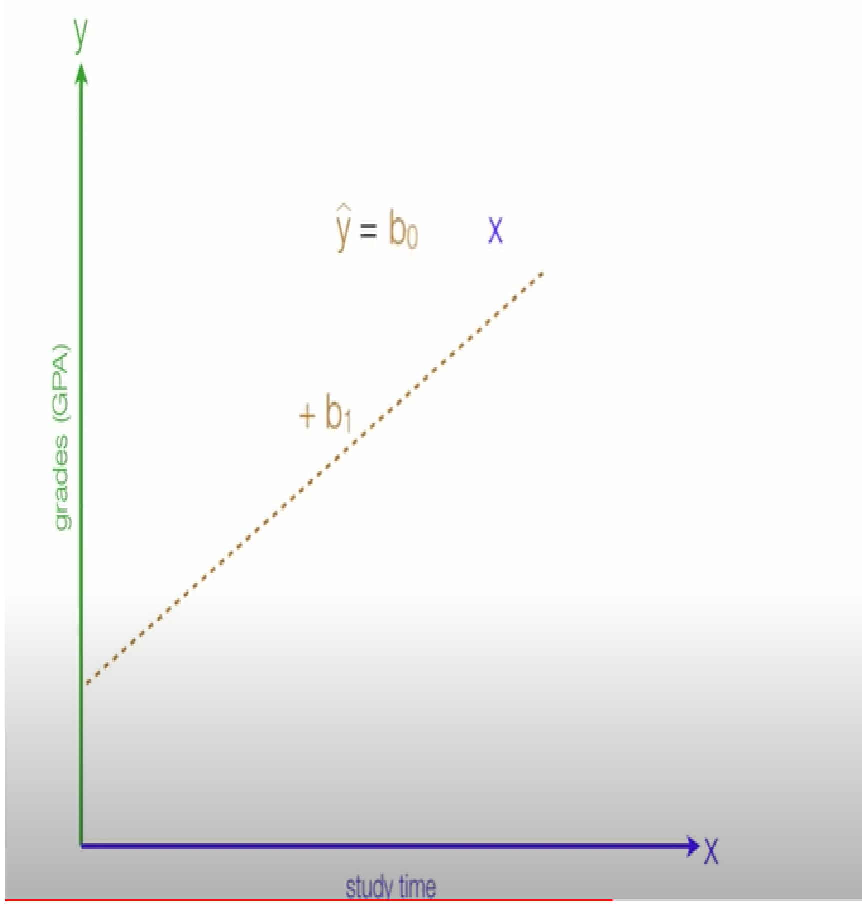
Regression



y
$\hat{y} = b_0 + b_1 X$ ⟶ study time
grades (GPA)
X
study time

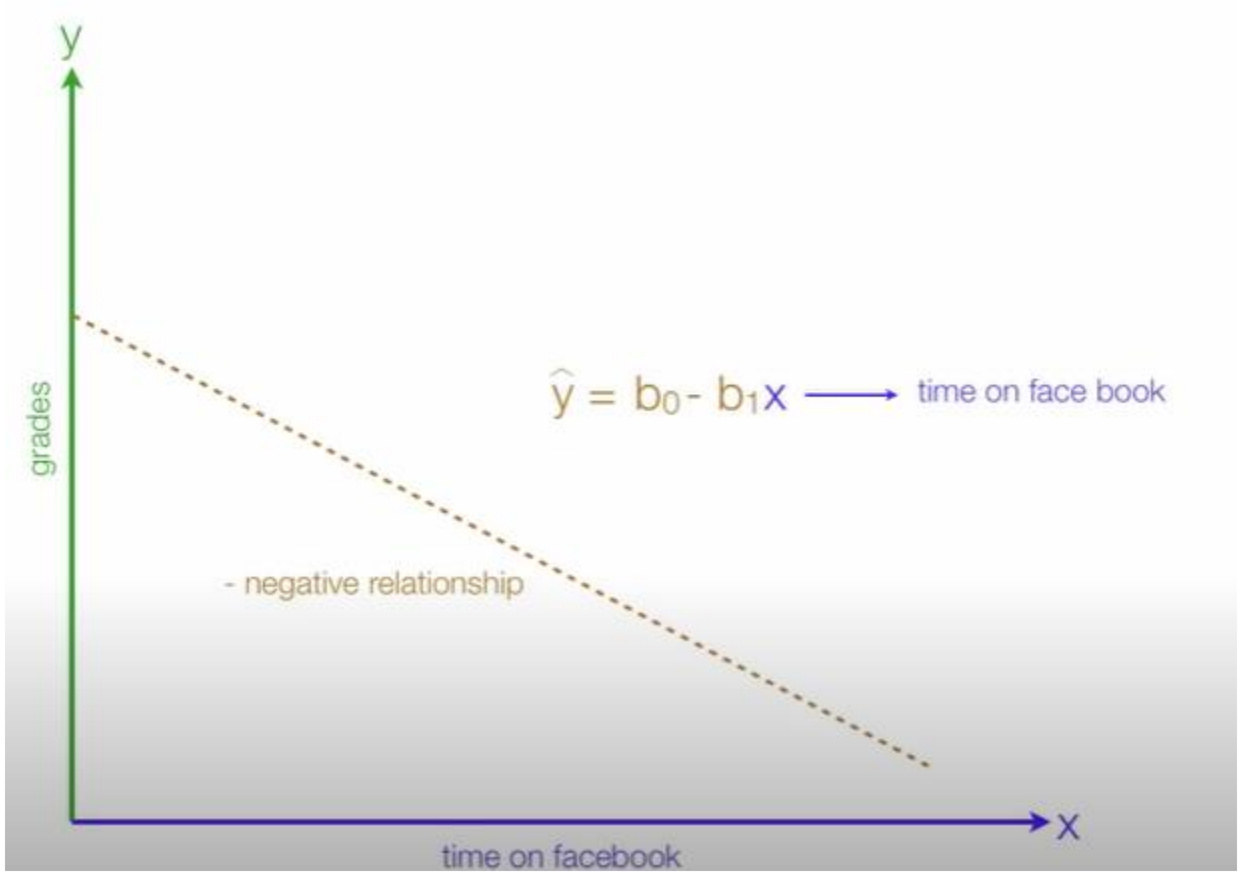


y
$\hat{y} = \quad + b_1 X$
grades (GPA)
b_0
X
study time



y
$\hat{y} = b_0 \quad X$
$+ b_1$
grades (GPA)
X
study time

Regression



y
grades
$\hat{y} = b_0 - b_1 x$
time on face book
- negative relationship
x
time on facebook

Simple Linear Regression Model

Simple linear regression model is a model with a single regressor x that has a linear relationship with a response y.

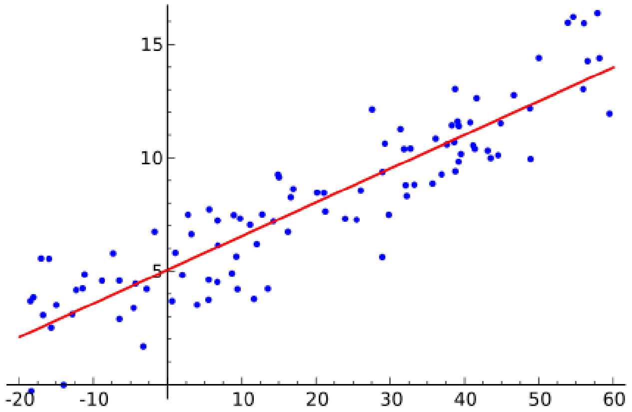
Simple linear regression model:

Intercept **Slope** **Random error component**

$$y = b_0 + b_1 x + \varepsilon$$

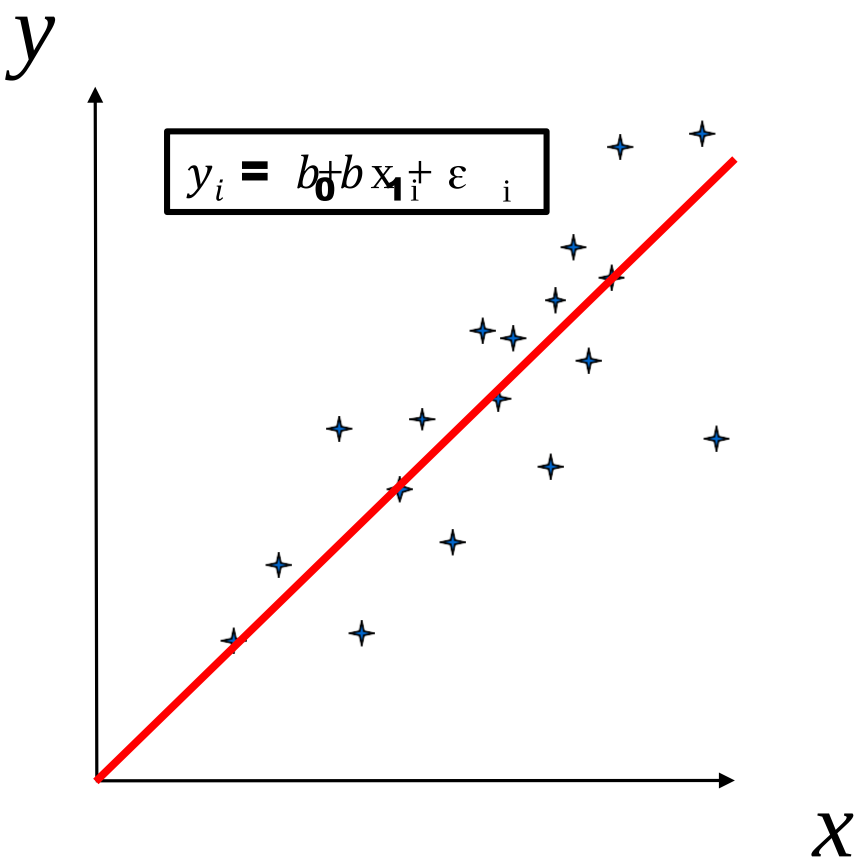
Response variable **Regressor variable**

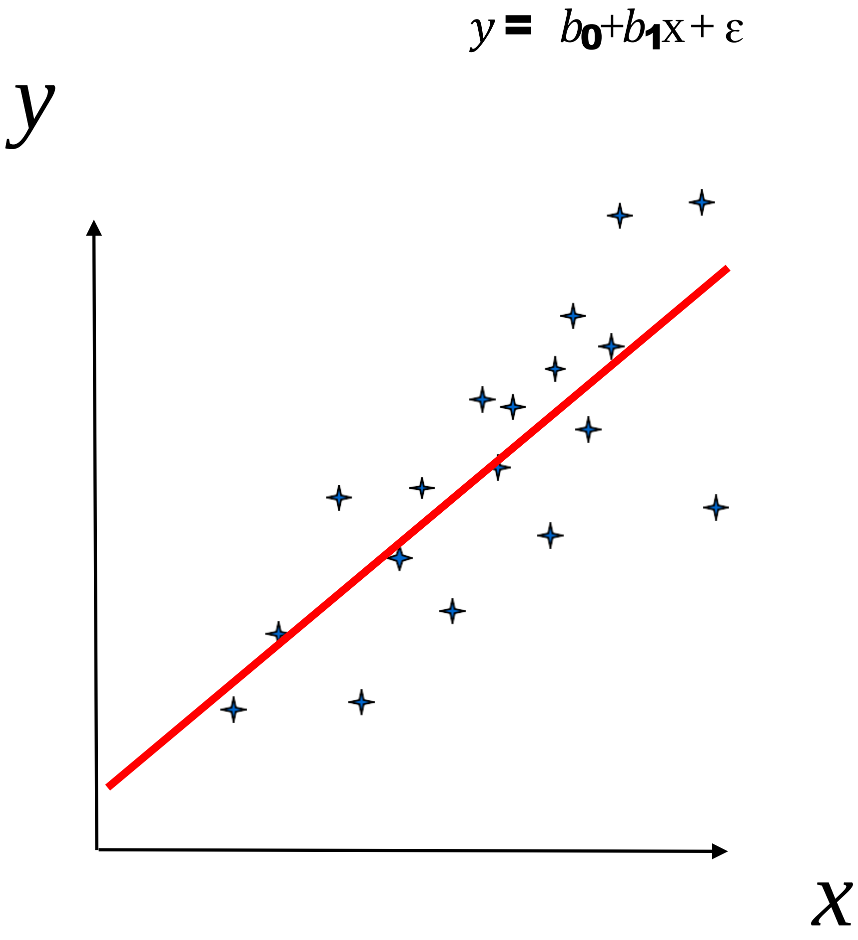
In this technique, the dependent variable is continuous and random variable, independent variable(s) can be continuous or discrete but it is not a random variable, and nature of regression line is linear.



Least Square Estimation for Parameters

The parameters b_0 and b_1 are unknown and must be estimates using sample data:
$(x_1,y_1), (x_2,y_2),\ldots\ldots(x_n,y_n)$





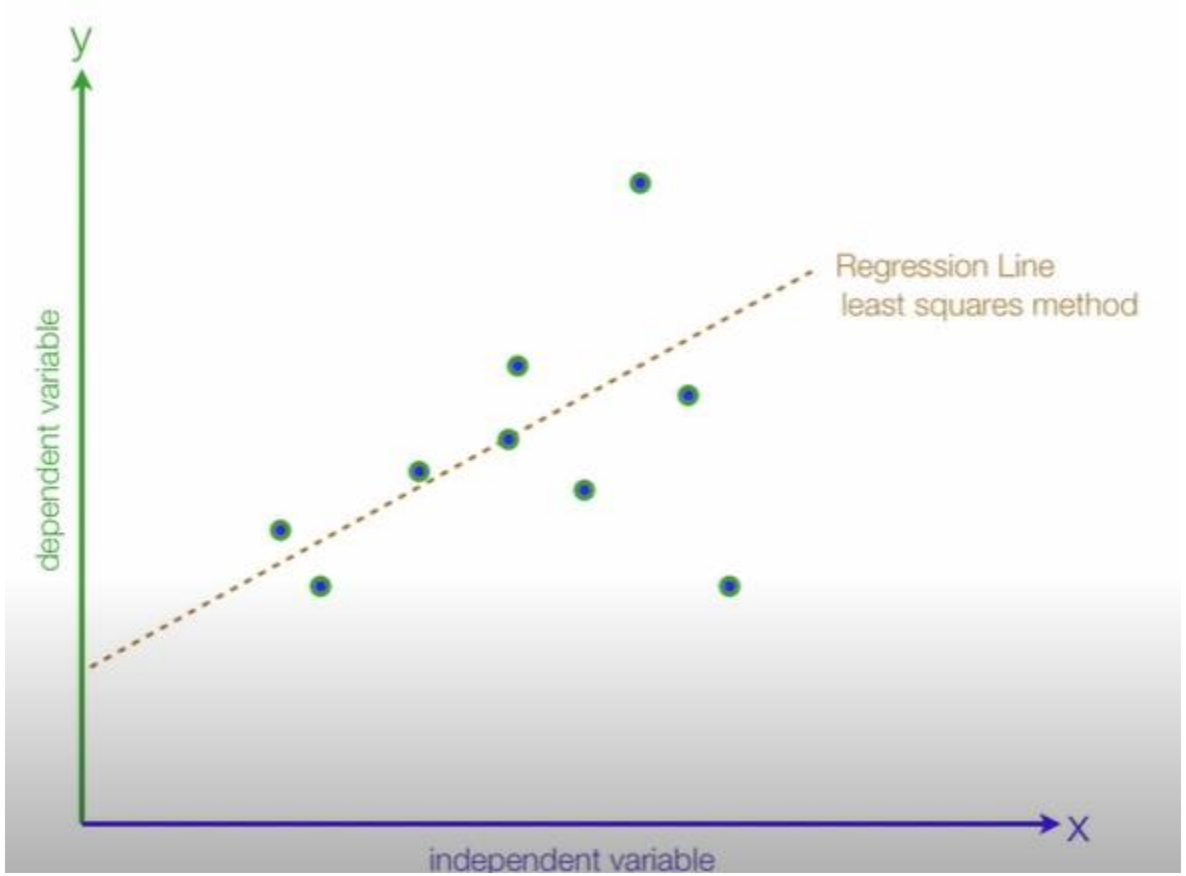
Regression Problem

- A regression problem is the problem of determining a relation between one or more independent variables and an output variable which is a real continuous variable, given a set of observed values of the set of independent

- Let us say we want to have a system that can predict the price of a used car. Inputs are the car attributes brand, year, engine capacity, mileage, and other information that we believe affect a car’s worth.

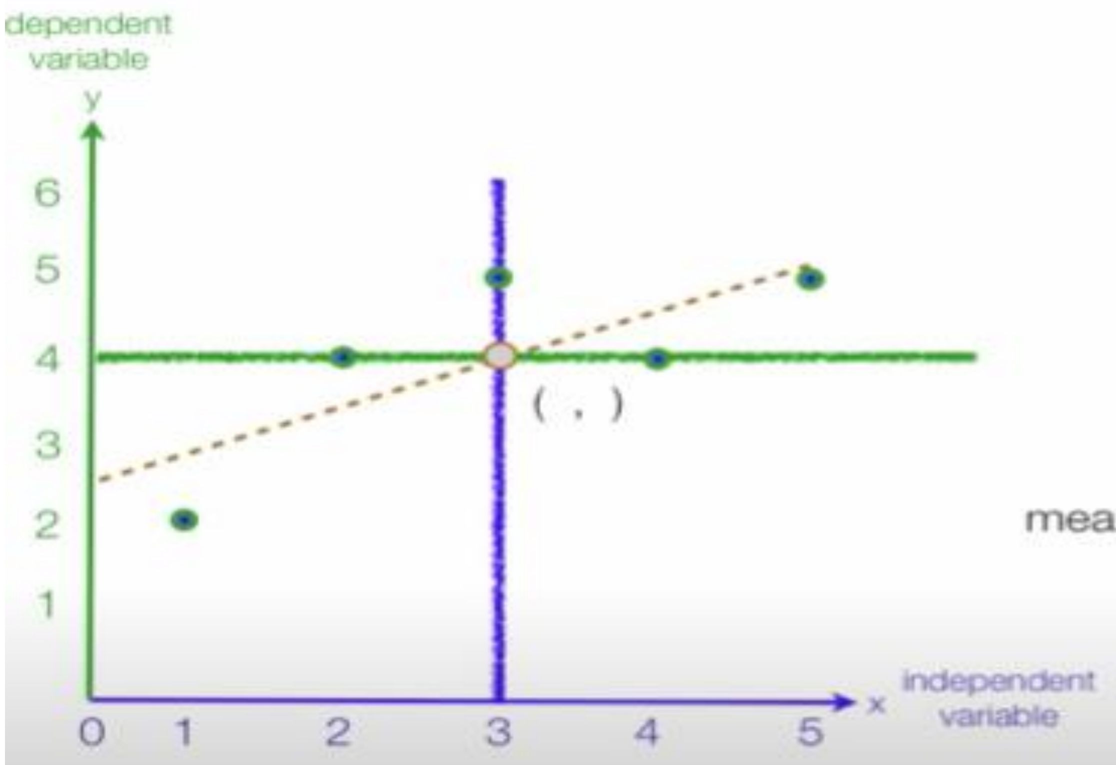
- The output is the price of the car. variables and the corresponding values of the output variable.

Calculate linear regression (using least square method)



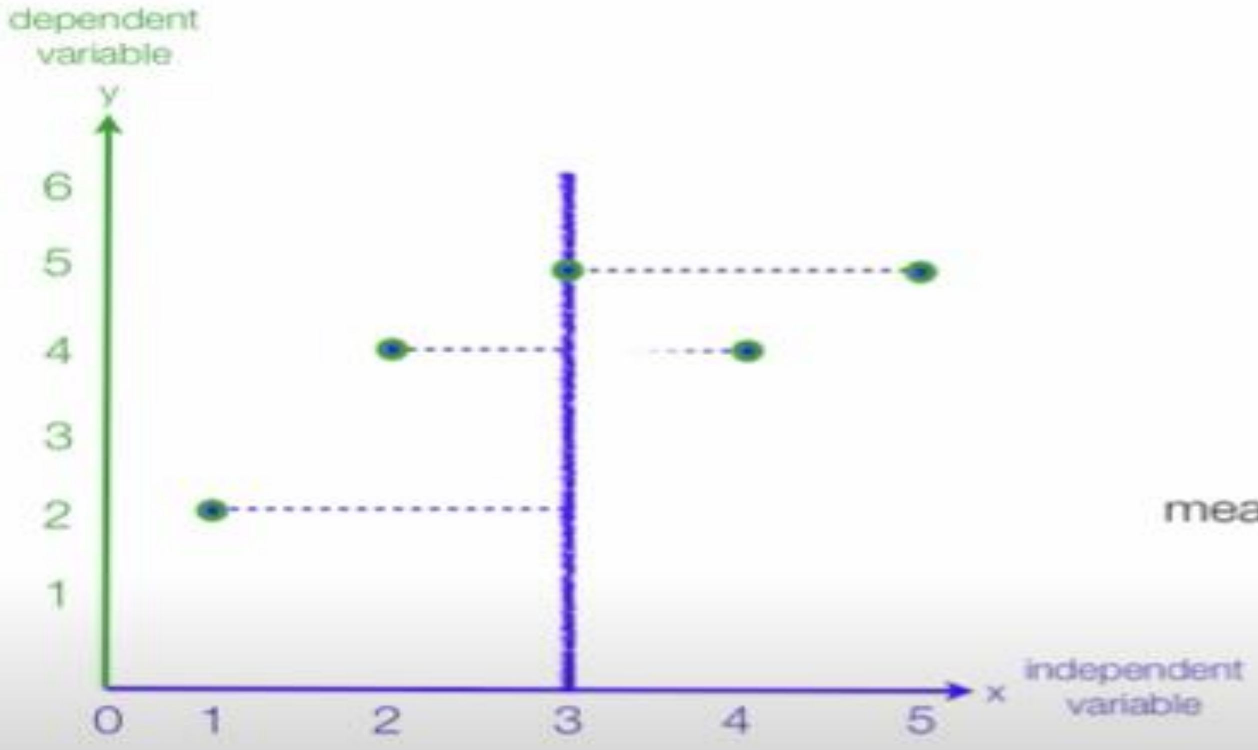
x	y
1	2
2	4
3	5
4	6
5	7

Calculate linear regression using least square method (minimize the distance between the errors)



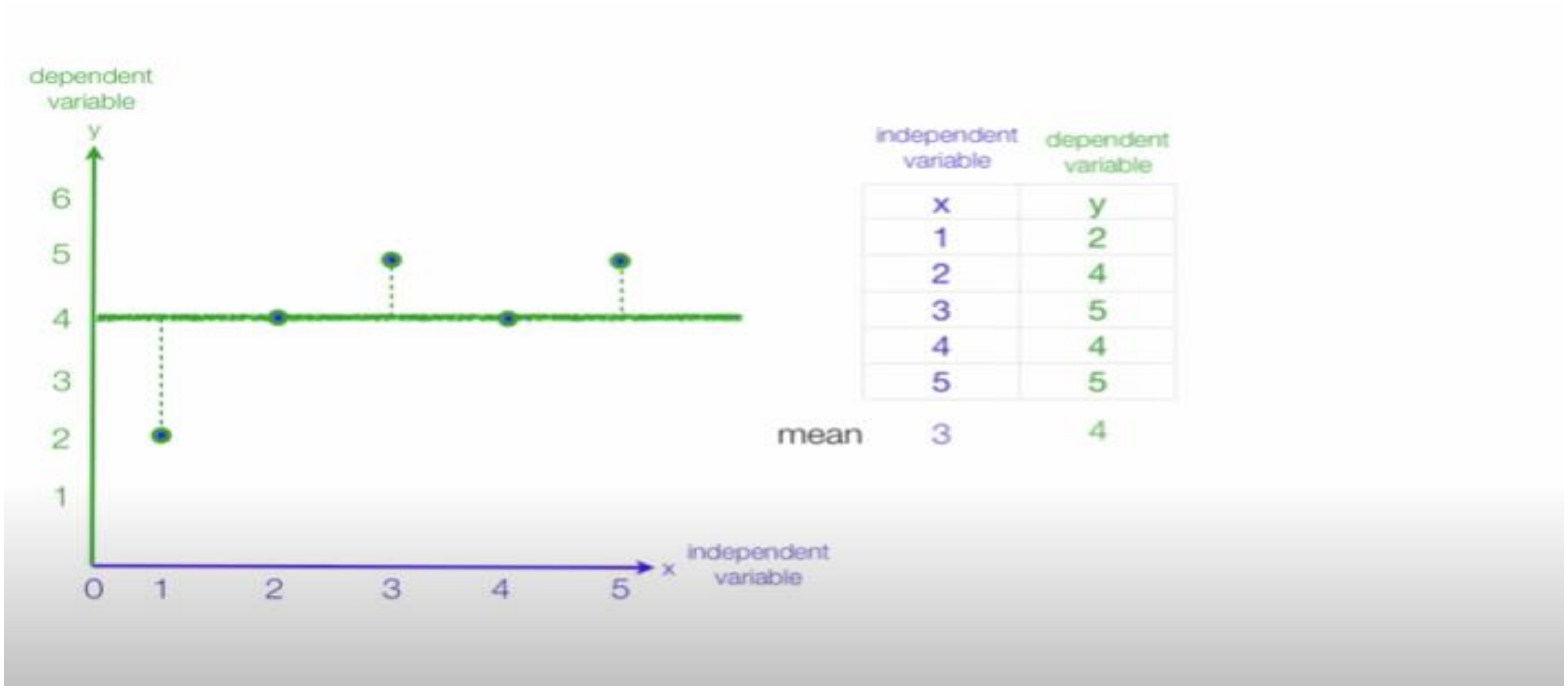
	independent variable	dependent variable
	x	y
	1	2
	2	4
	3	5
	4	4
	5	5
mean	3	4

Distance from the x value to the mean for all observations

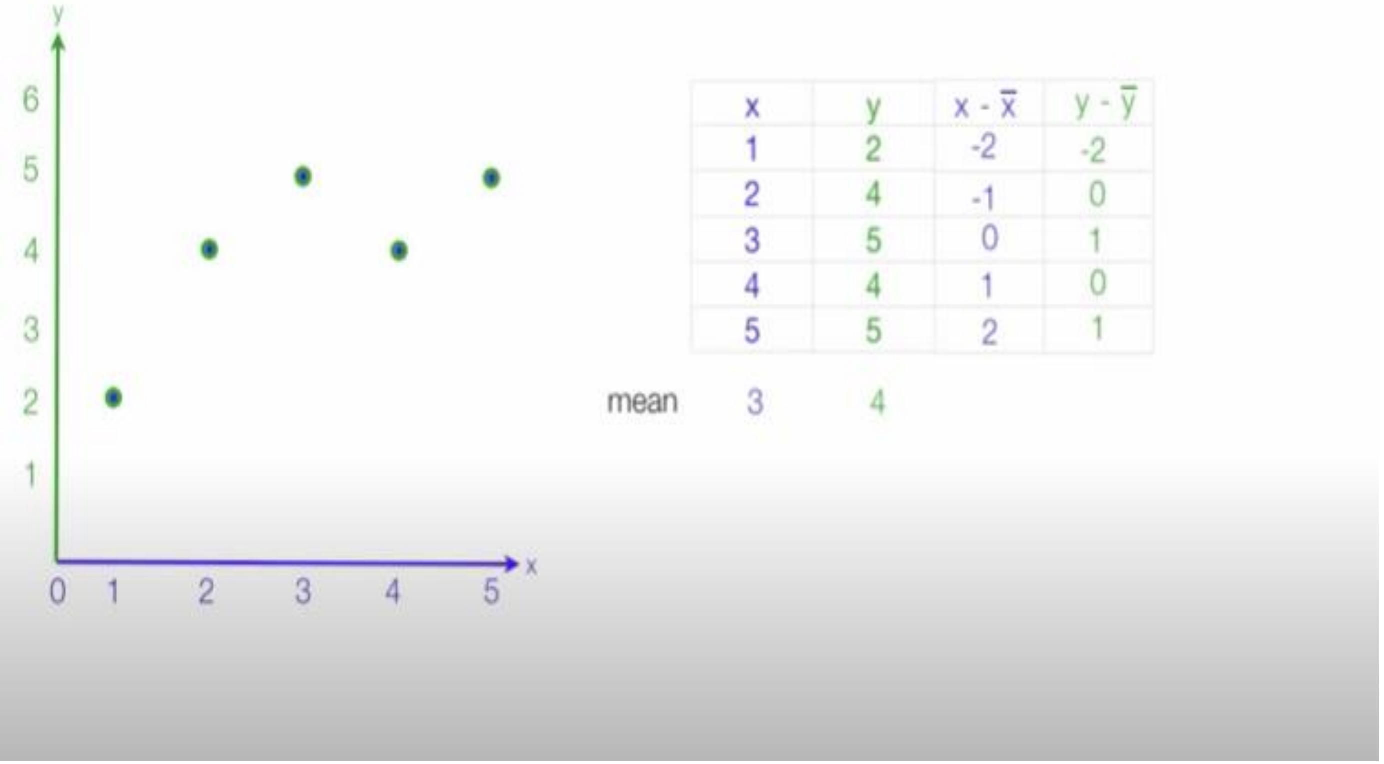


	independent variable	dependent variable
	x	y
	1	2
	2	4
	3	5
	4	4
	5	5
mean	3	4

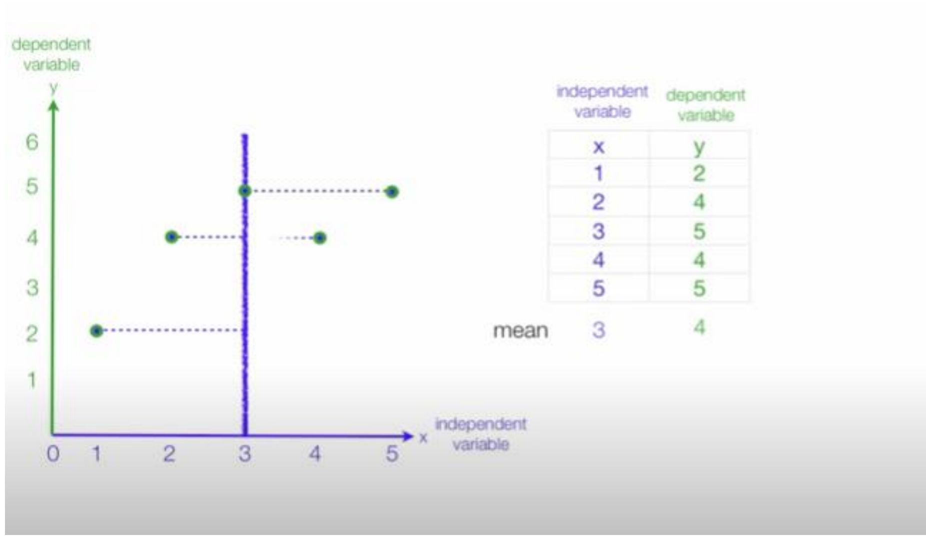
Distance from the y value to the mean for all observations



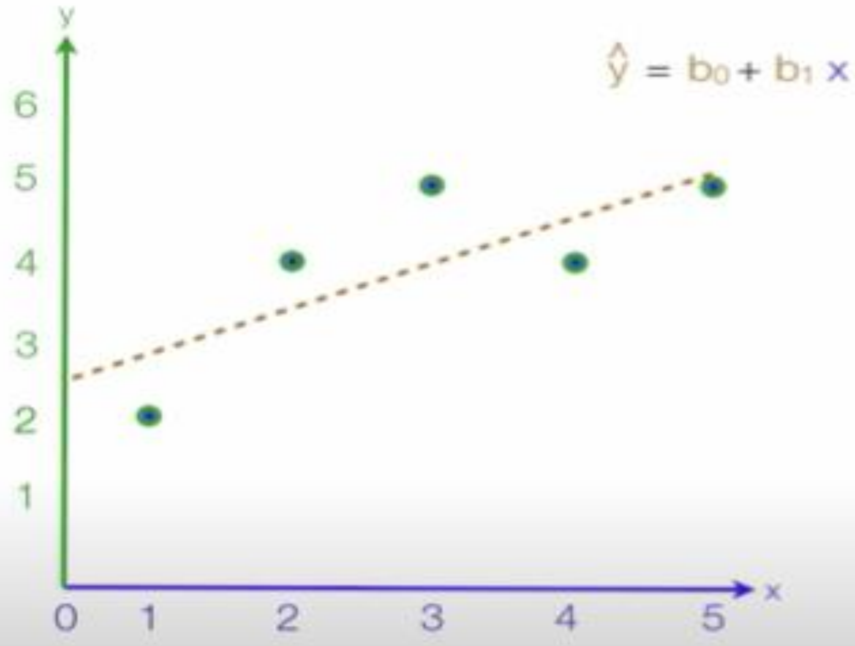
Distance from the x and y value to the mean for all observations



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



To calculate b_1

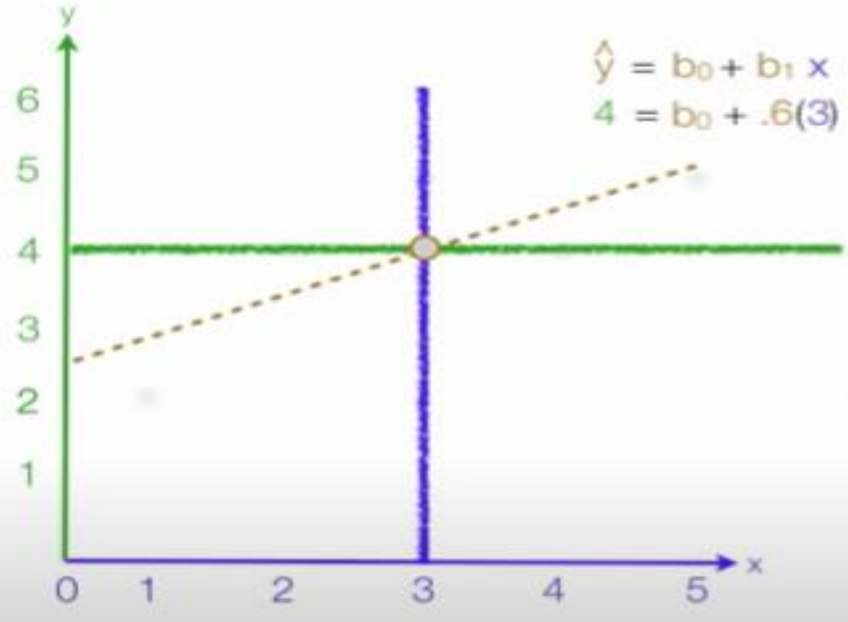


	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	1	2	-2	-2	4	4
	2	4	-1	0	1	0
	3	5	0	1	0	0
	4	4	1	0	1	0
	5	5	2	1	4	2
mean	3	4			10	6

sum

$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

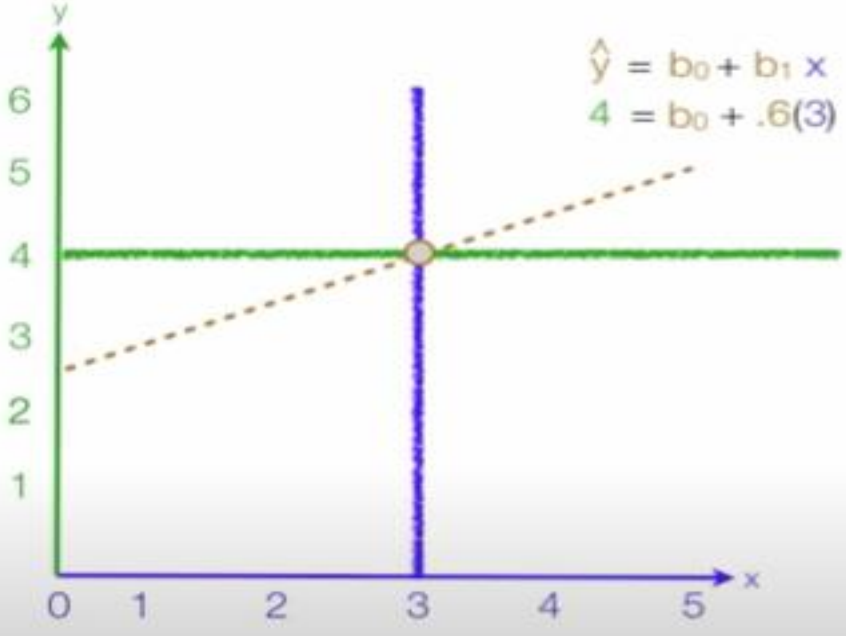
To calculate b_0



x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
				10	6
mean 3	4				

$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

To calculate b_0



x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
				10	6
mean 3	4				

$$4 = b_0 + .6(3)$$

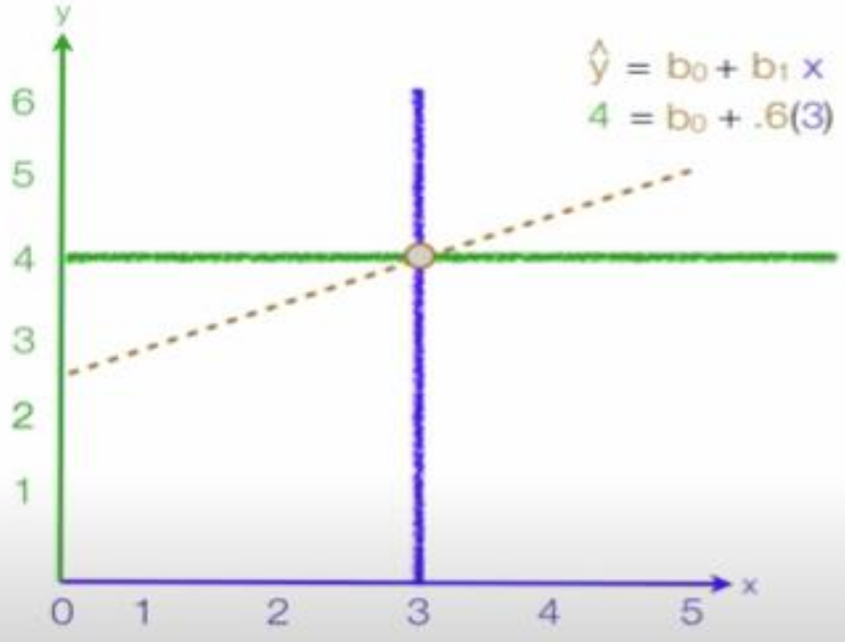
$$4 = b_0 + 1.8$$

$$-1.8 \qquad -1.8$$

$$2.2 = b_0$$

$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

To calculate b_0



$b_0 = 2.2$

$b_1 = .6$

$\hat{y} = 2.2 + .6x$

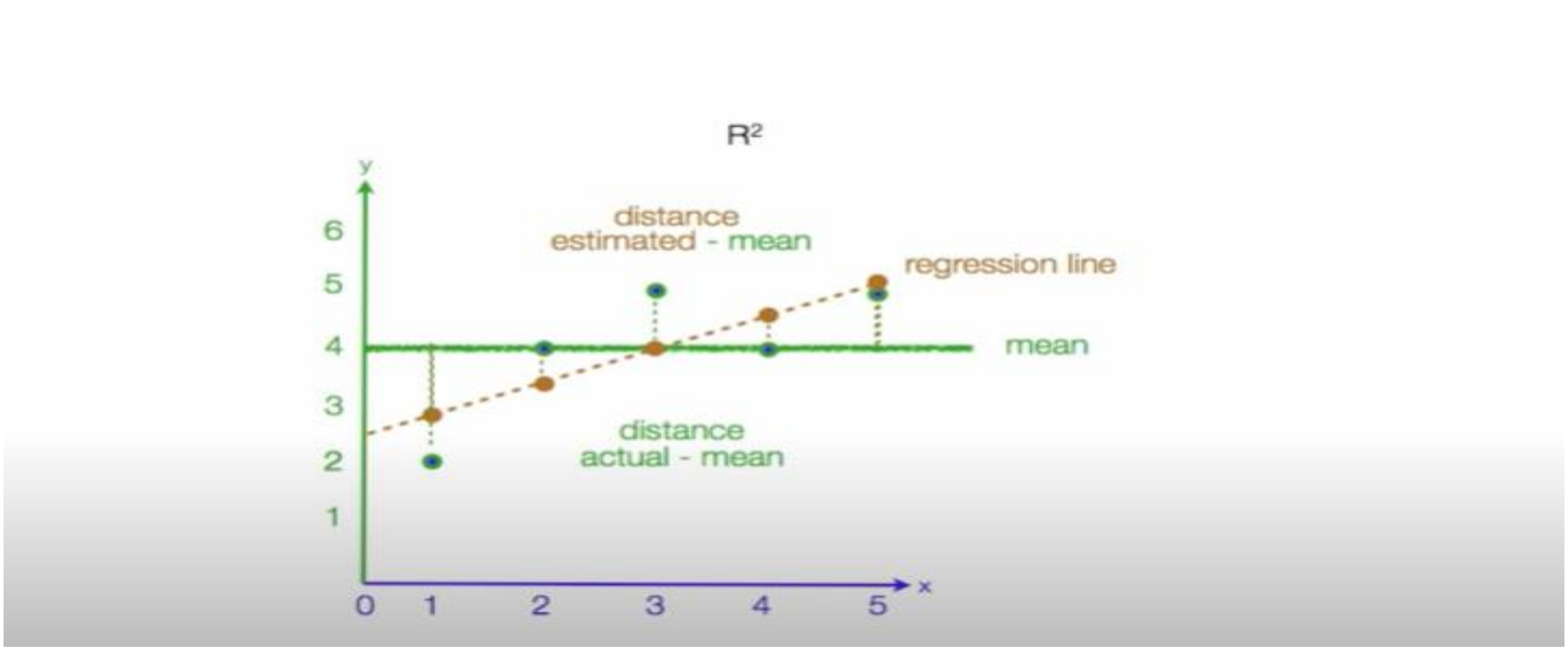
	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	1	2	-2	-2	4	4
	2	4	-1	0	1	0
	3	5	0	1	0	0
	4	4	1	0	1	0
	5	5	2	1	4	2
					10	6
mean	3	4				

$$4 = b_0 + .6(3)$$

$$\begin{array}{r} 4 = b_0 + \cancel{1.8} \\ -1.8 \quad\quad \cancel{-1.8} \\ \hline 2.2 = b_0 \end{array}$$

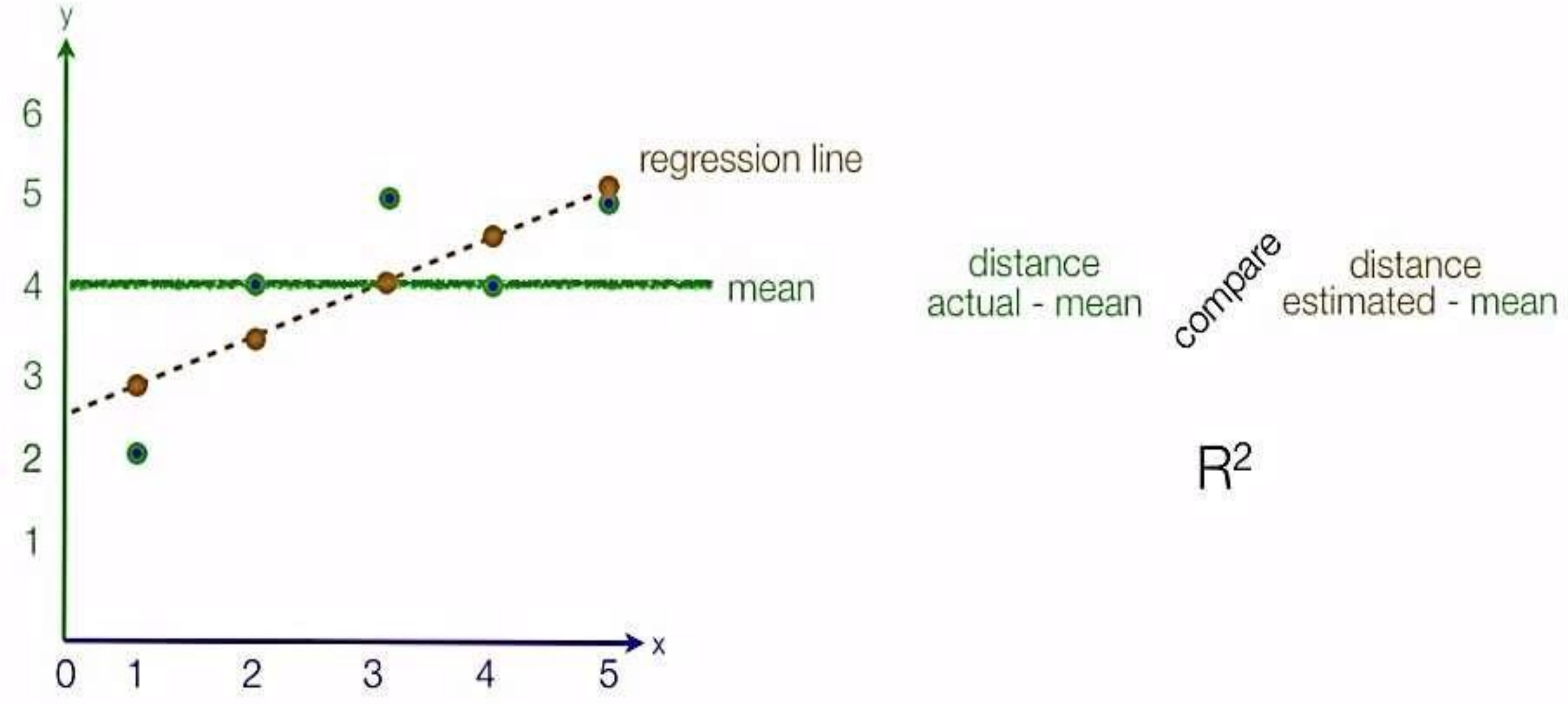
$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

CALCULATING R SQUARED

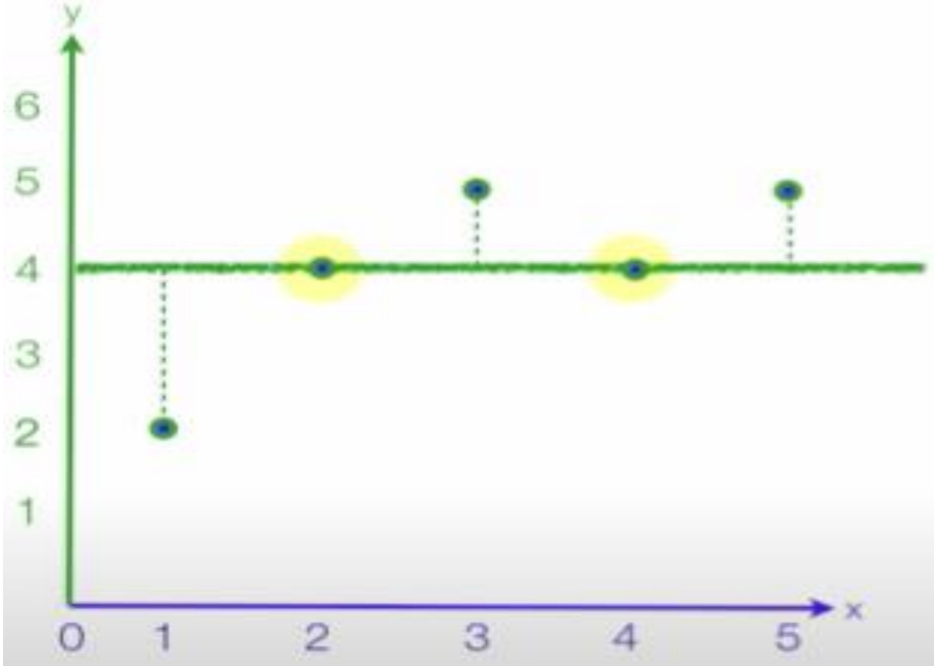


Calculating R^2 Using Regression Analysis

- R-squared is a statistical measure of how close the data are to the fitted regression line(For measuring the goodness of fit). It is also known as the coefficient of determination.
- Firstly we calculate distance between actual values and mean value and also calculate distance between estimated value and mean value.
- Then compare both the distances.



R squared *Predicted*



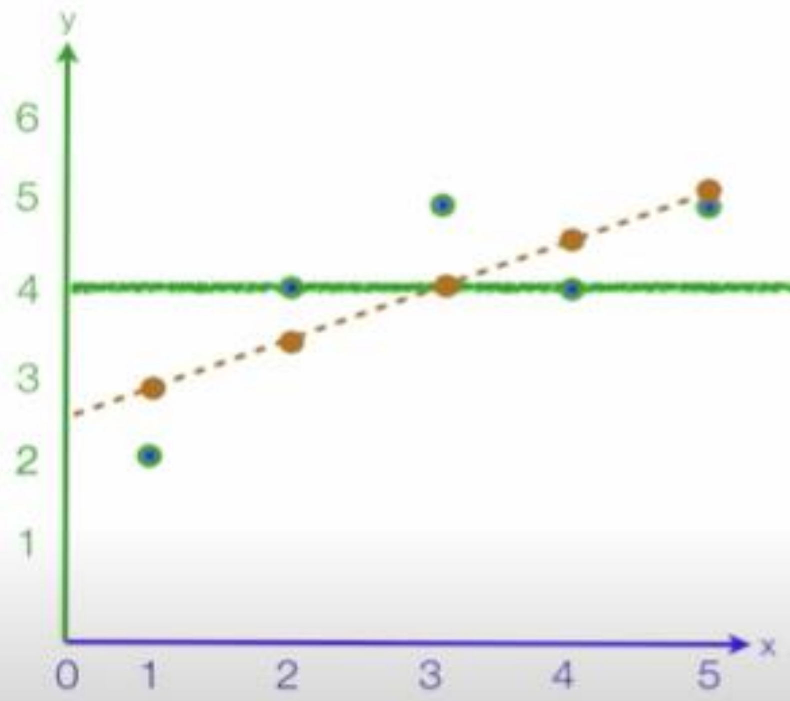
y - mean

x	y	$y - \bar{y}$	$(y - \bar{y})^2$
1	2	-2	4
2	4	0	0
3	5	1	1
4	4	0	0
5	5	1	1
			6
mean	4		

R squared *Estimated (coeff of determination)*

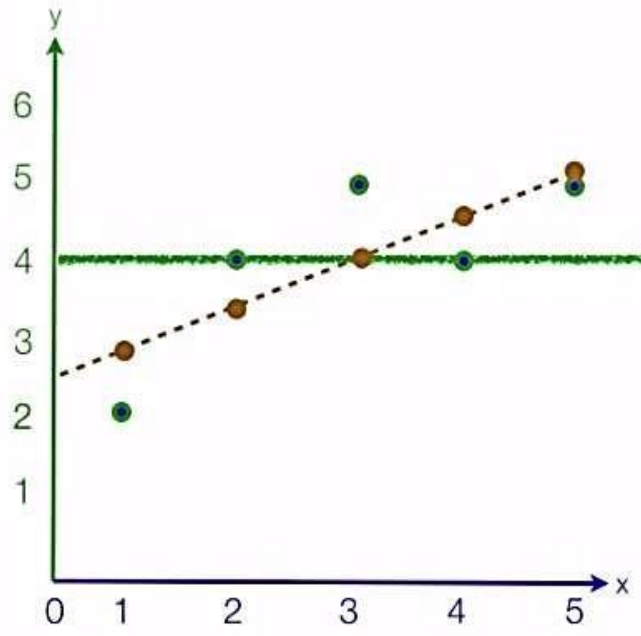
$\hat{y} = 2.2 + .6x$

estimated - mean



x	y	$y - \bar{y}$	$(y - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	2	-2	4	2.8	-1.2	1.44
2	4	0	0	3.4	-.6	.36
3	5	1	1	4	0	0
4	4	0	0	4.6	.6	.36
5	5	1	1	5.2	1.2	1.44
			6			3.6
mean	4					

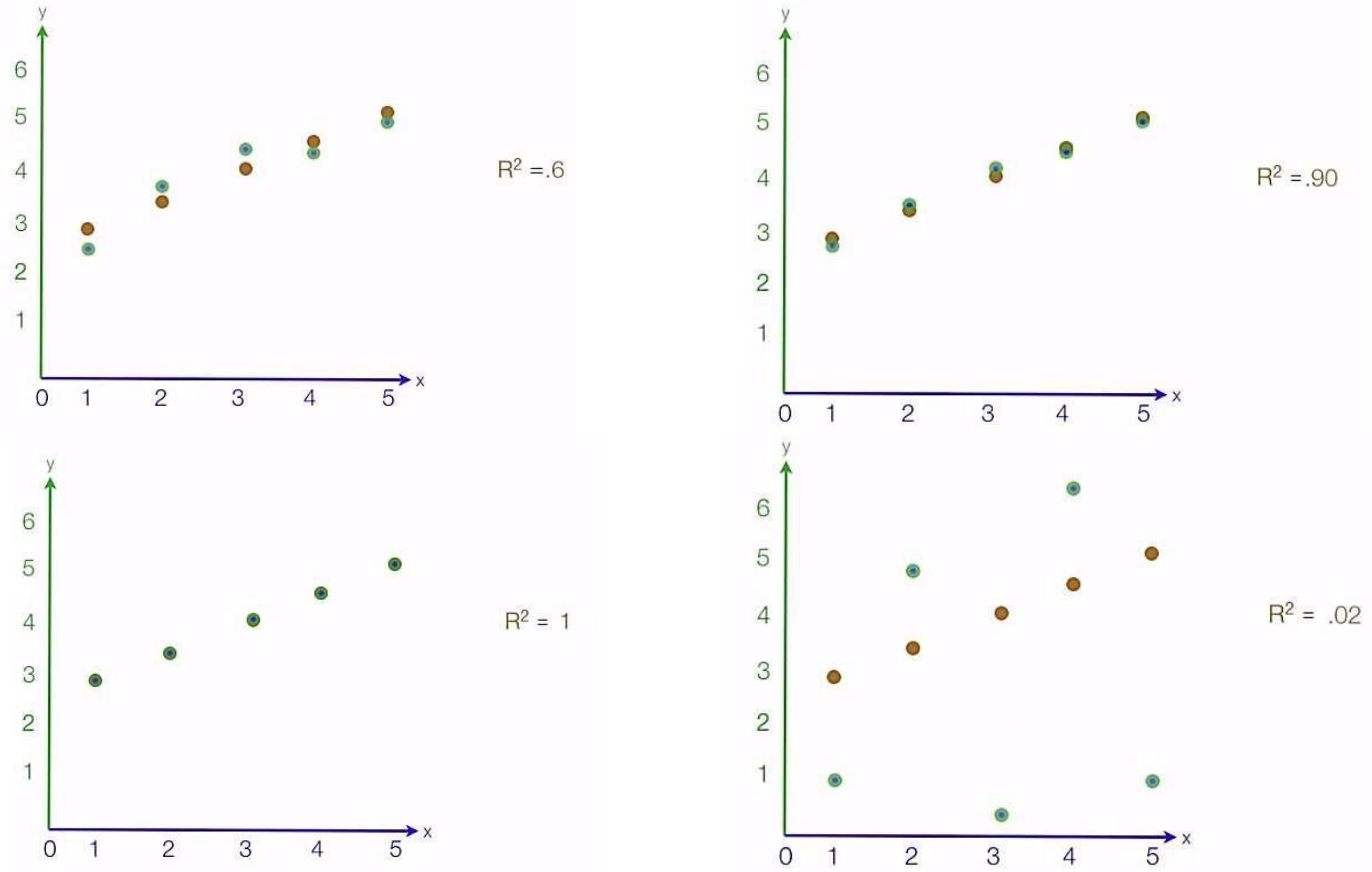
Example



x	y	$y - \bar{y}$	$(y - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	2	-2	4	2.8	-1.2	1.44
2	4	0	0	3.4	-.6	.36
3	5	1	1	4	0	0
4	4	0	0	4.6	.6	.36
5	5	1	1	5.2	1.2	1.44
			6			3.6
mean	4					

$$R^2 = \frac{3.6}{6} = .6 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

Performance of Model



Standard error of the Estimate (Mean square error)

The standard error of the estimate is a measure of the accuracy of predictions.

Note: The regression line is the line that minimizes the sum of squared deviations of prediction (also called the *sum of squares error*).

The standard error of the estimate is closely related to this quantity and is defined below:

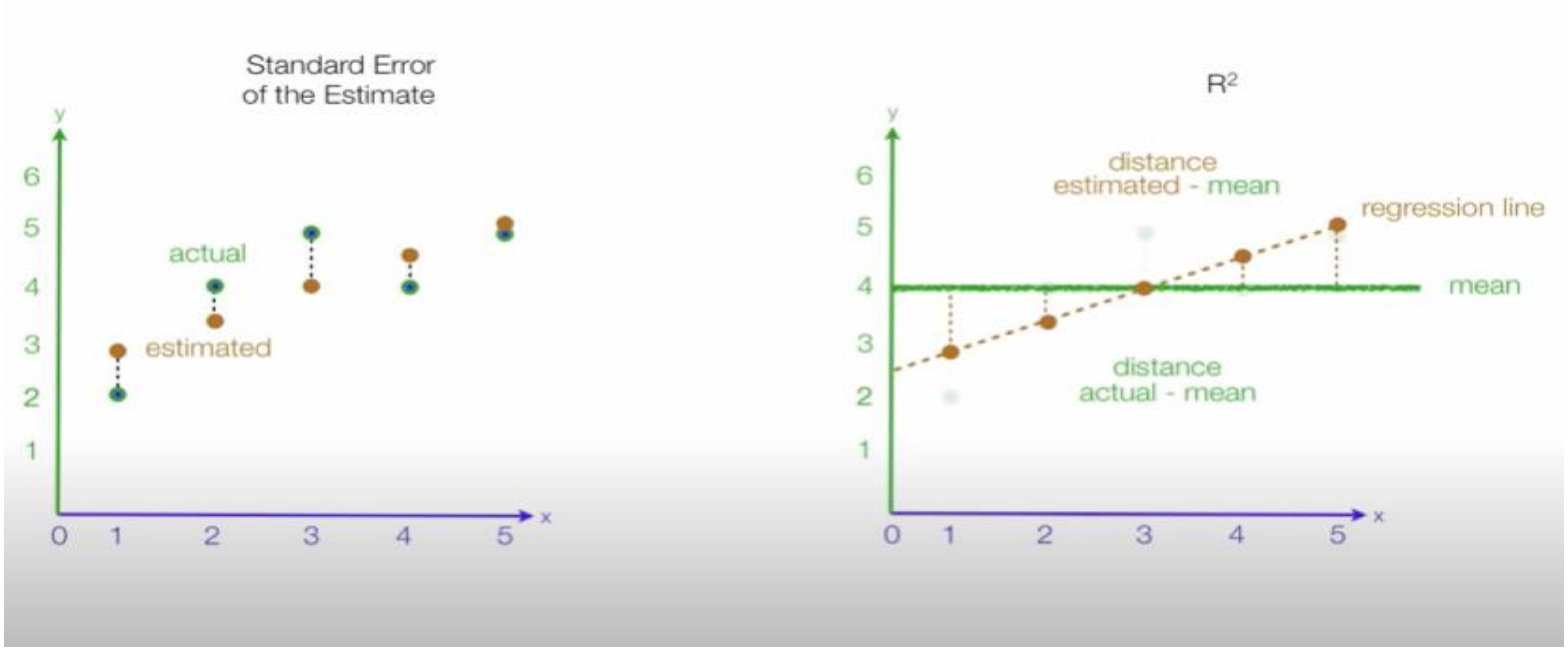
$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

Where Y = actual value

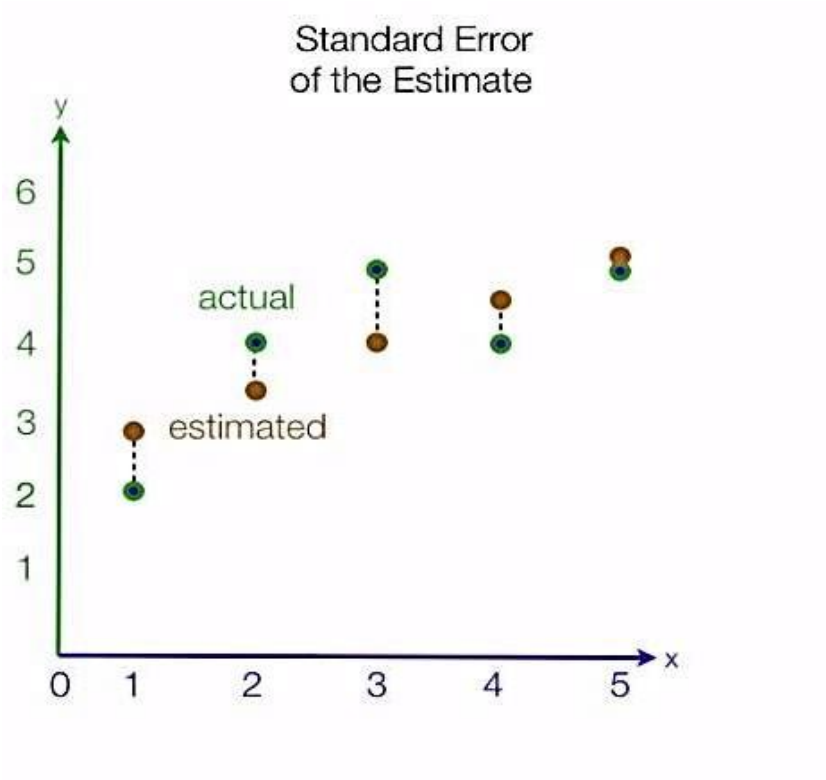
Y'= Estimated Value

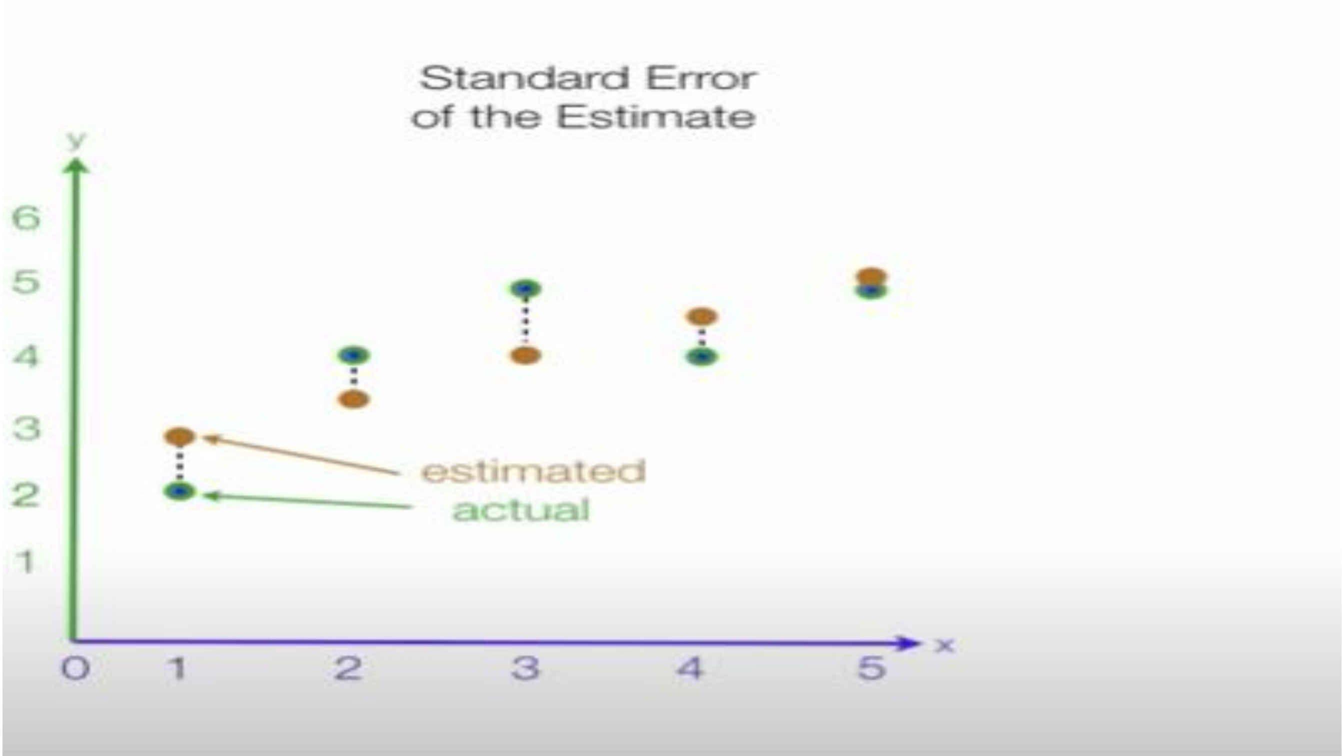
N = No. of observations

Difference between R^2 and Standard error of the Estimate



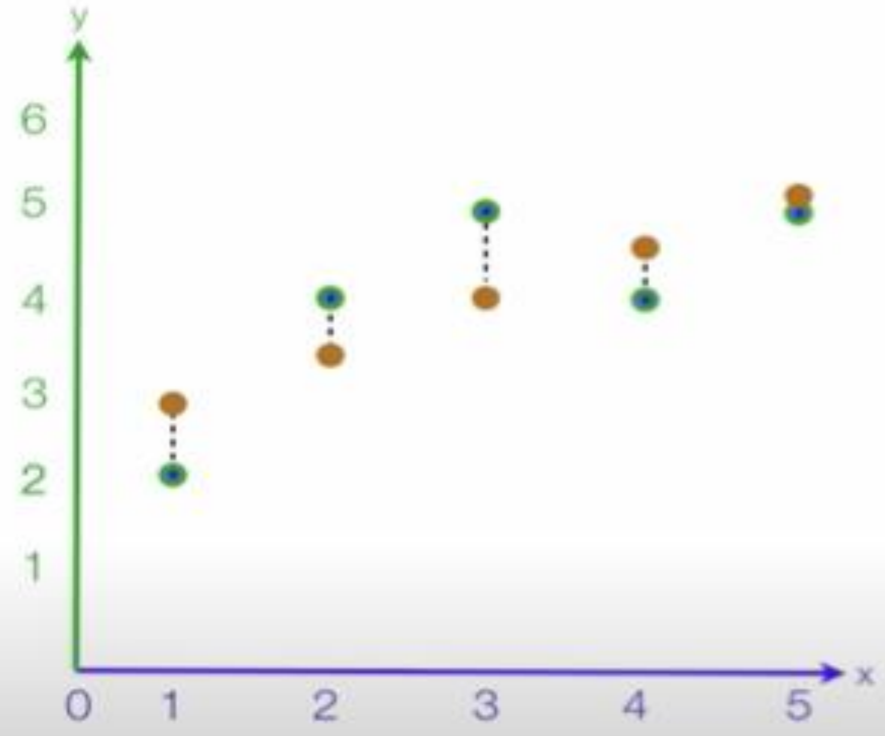
Example : Standard Error of the Estimate used in Regression Analysis (Mean Square Error)





Error: estimated distance between actual

Example : Standard Error of the Estimate used in Regression Analysis (Mean Square Error)

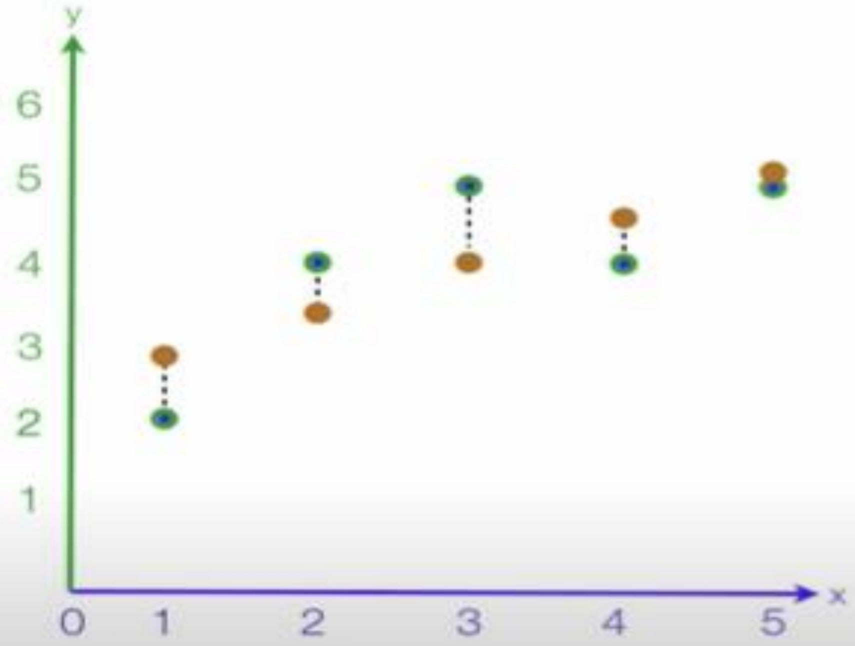


x	y	$\hat{y}$
1	2	2.8
2	4	3.4
3	5	4
4	4	4.6
5	5	5.2

$$\text{Standard Error of the Estimate} = \sqrt{\frac{\Sigma(\hat{y} - y)^2}{n - 2}}$$

Where n is the no of observations

Example : Standard Error of the Estimate used in Regression Analysis (Mean Square Error)

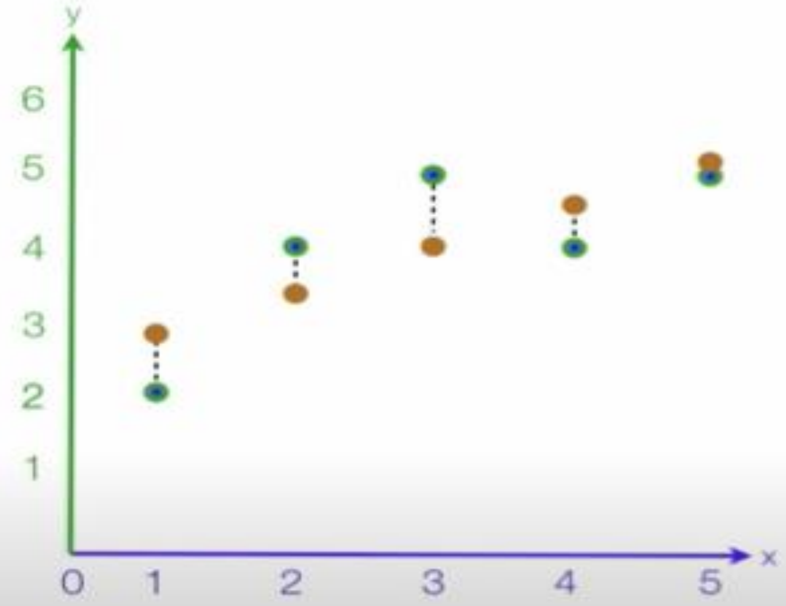


x	y	$\hat{y}$	$\hat{y} - y$
1	2	2.8	.8
2	4	3.4	-.6
3	5	4	-1
4	4	4.6	.6
5	5	5.2	.2

$$\text{Standard Error of the Estimate} = \sqrt{\frac{\Sigma(\hat{y} - y)^2}{n - 2}}$$

Where n is the no of observations

Example : Standard Error of the Estimate used in Regression Analysis (Mean Square Error)

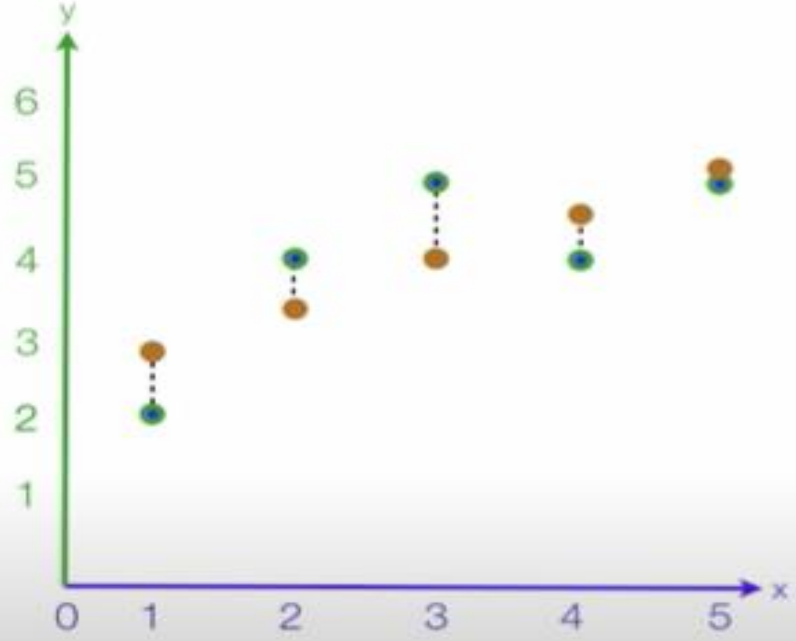


x	y	$\hat{y}$	$\hat{y} - y$	$(\hat{y} - y)^2$
1	2	2.8	.8	.64
2	4	3.4	-.6	.36
3	5	4	-1	1
4	4	4.6	.6	.36
5	5	5.2	.2	.04
				2.4

$$\text{Standard Error of the Estimate} = \sqrt{\frac{\Sigma(\hat{y} - y)^2}{n - 2}}$$

Where n is the no of observations

Example : Standard Error of the Estimate used in Regression Analysis (Mean Square Error)

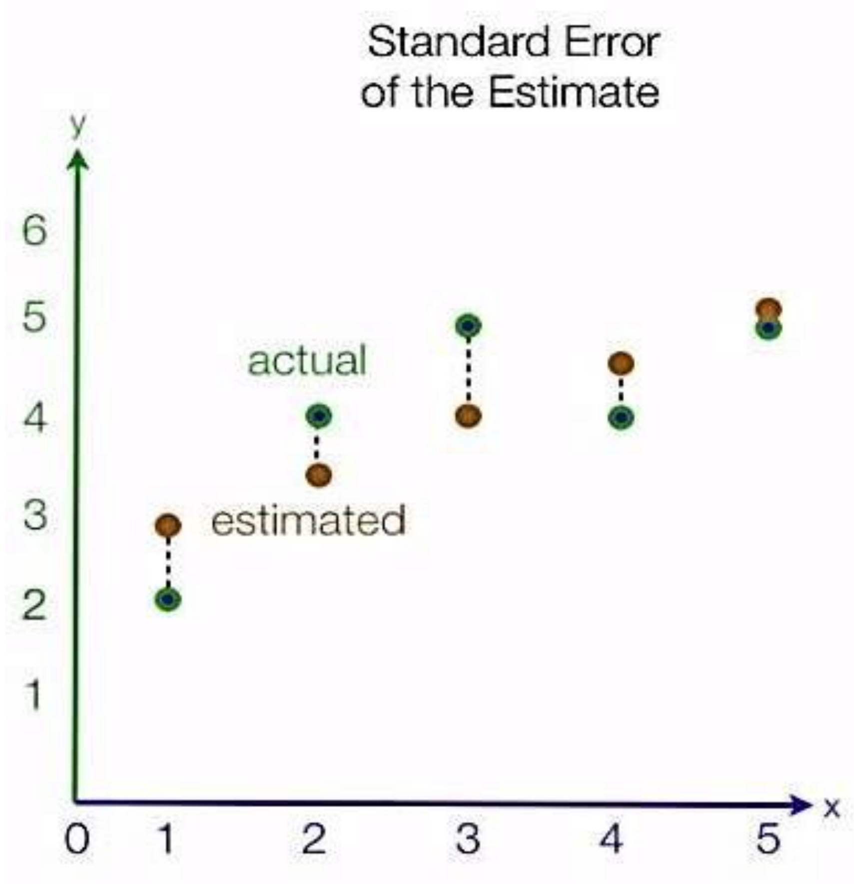


x	y	$\hat{y}$	$\hat{y} - y$	$(\hat{y} - y)^2$
1	2	2.8	.8	.64
2	4	3.4	-.6	.36
3	5	4	-1	1
4	4	4.6	.6	.36
5	5	5.2	.2	.04
				2.4

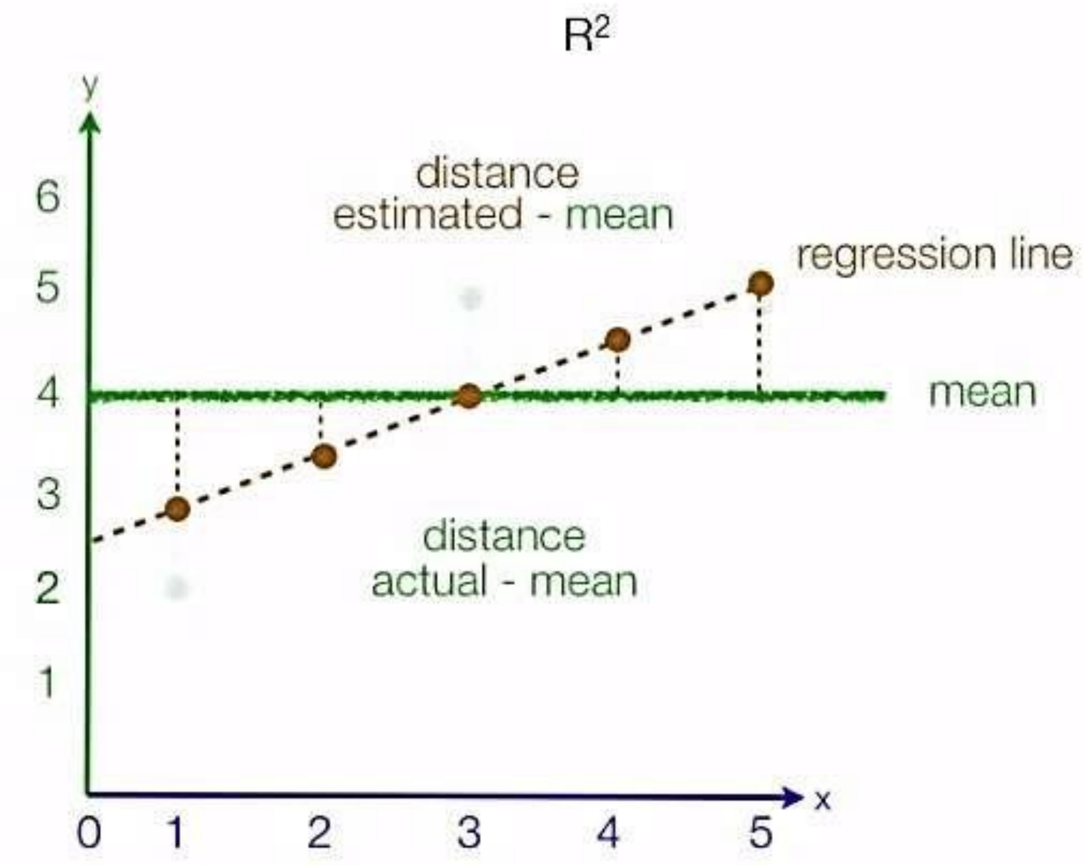
$$\text{Standard Error of the Estimate} = \sqrt{\frac{\Sigma(\hat{y} - y)^2}{n - 2}} = \sqrt{\frac{2.4}{5-2}} = \sqrt{\frac{2.4}{3}} = \sqrt{.8} = .89$$

Where n is the no of observations

Difference



Standard Error
of the Estimate
y
6
5
4
3
2
1
actual
estimated
0
1
2
3
4
5
x



R2
y
6
5
4
3
2
1
distance
estimated - mean
regression line
mean
distance
actual - mean
0
1
2
3
4
5
x

MSE or RMSE

- It measures the **average squared difference between the predicted values and the actual values**, quantifying the discrepancy between the model’s predictions and the true observations.

- Intuitively, the MSE is used to measure the quality of the model based on the predictions made on the entire training dataset.

- In other words, it can be used to represent the **cost associated with the predictions** or the **loss incurred in the predictions**

- **Definition**:
  - **Mean Squared Error (MSE)** is the average of the squares of the errors or deviations. The error is the amount by which the actual values differ from the predicted values.

  - **Root Mean Squared Error (RMSE)** is the square root of MSE.

- **Use in Model Evaluation**:
  - Both metrics are used to measure the quality of a predictor or a regression model; lower values indicate a better fit.
  - RMSE is more commonly reported as it is more interpretable, being in the same units as the dependent variable.

R-Squared

- R-Squared, also known as the **coefficient of determination**, is another statistical metric used to evaluate the performance of regression models. It measures the proportion of the total variation in the dependent variable (output) that can be explained by the independent variables (inputs) in the model.

- Mathematically, that can be represented as the **ratio of the sum of squares regression (SSR) and the sum of squares total (SST)**.

- Sum of Squares Regression (SSR) represents the total variation of all the predicted values found on the regression line or plane from the mean value of all the values of response variables.

- The sum of squares total (SST) represents the total variation of actual values from the mean value of all the values of response variables.
- R-squared value is used to measure the **goodness of fit or best-fit line.** The greater the value of R-Squared, the better is the regression model as most of the variation of actual values from the mean value get explained by the regression model.

MSE or RMSE

- It measures the **average squared difference between the predicted values and the actual values**, quantifying the discrepancy between the model's predictions and the true observations.

- Intuitively, the MSE is used to measure the quality of the model based on the predictions made on the entire training dataset.

- In other words, it can be used to represent the **cost associated with the predictions** or the **loss incurred in the predictions**

- **Definition**:
  - **Mean Squared Error (MSE)** is the average of the squares of the errors or deviations. The error is the amount by which the actual values differ from the predicted values.

  - **Root Mean Squared Error (RMSE)** is the square root of MSE.

- **Use in Model Evaluation**:
  - Both metrics are used to measure the quality of a predictor or a regression model; lower values indicate a better fit.
  - RMSE is more commonly reported as it is more interpretable, being in the same units as the dependent variable.

R-Squared

- R-Squared, also known as the **coefficient of determination**, is another statistical metric used to evaluate the performance of regression models. It measures the proportion of the total variation in the dependent variable (output) that can be explained by the independent variables (inputs) in the model.

- Mathematically, that can be represented as the **ratio of the sum of squares regression (SSR) and the sum of squares total (SST)**.

- Sum of Squares Regression (SSR) represents the total variation of all the predicted values found on the regression line or plane from the mean value of all the values of response variables.

- The sum of squares total (SST) represents the total variation of actual values from the mean value of all the values of response variables.
- R-squared value is used to measure the **goodness of fit or best-fit line.** The greater the value of R-Squared, the better is the regression model as most of the variation of actual values from the mean value get explained by the regression model.

- **Mean Absolute Error (MAE)**

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=0}^{n-1}\left|y_i - \hat{y}_i\right|$$

$$\mathrm{MAE} = \frac{1}{2}\times\left(\left|80-75\right| + \left|75-85\right|\right) = \frac{15}{2} = 7.5$$

x_i Items	y_j (Sales)
I1	80
I2	90
I3	100
I4	110
I5	120

Test Items	Actual Value	Predicted Value
I6	80	75
I7	75	85

- **Mean Squared Error (MSE)**

$$\text{MSE} = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2$$

$$\text{MSE} = \frac{1}{2} \times \left(|80 - 75|^2 + |75 - 85|^2\right) = \frac{125}{2} = 62.5$$

x_i Items	y_j (Sales)
I1	80
I2	90
I3	100
I4	110
I5	120

Test Items	Actual Value	Predicted Value
I6	80	75
I7	75	85

- Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{62.5} = 7.91$$

x_i Items	y_j (Sales)
I1	80
I2	90
I3	100
I4	110
I5	120

Test Items	Actual Value	Predicted Value
I6	80	75
I7	75	85

- **Relative MSE**

$$\text{RelMSE} = \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y}_i)^2}$$

- For finding **RelMSE** and **CV**, the training table should be used to find the average of y.

x_i Items	y_j (Sales)
I1	80
I2	90
I3	100
I4	110
I5	120

The average of y is $\frac{80 + 90 + 100 + 110 + 120}{5} = \frac{500}{5} = 100$.

Test Items	Actual Value	Predicted Value
I6	80	75
I7	75	85

$$\text{RelMSE} = \frac{(80-75)^2 + (75-85)^2}{(80-100)^2 + (75-100)^2} = \frac{125}{1025} = 0.1219$$

- **Coefficient of Variation**

$$CV = \frac{RMSE}{\bar{y}}$$

$$CV = \frac{\sqrt{62.5}}{100} = 0.08$$

x_i Items	y_j (Sales
I1	80
I2	90
I3	100
I4	110
I5	120

Test Items	Actual Value	Predicted Value
I6	80	75
I7	75	85

Multiple Regression

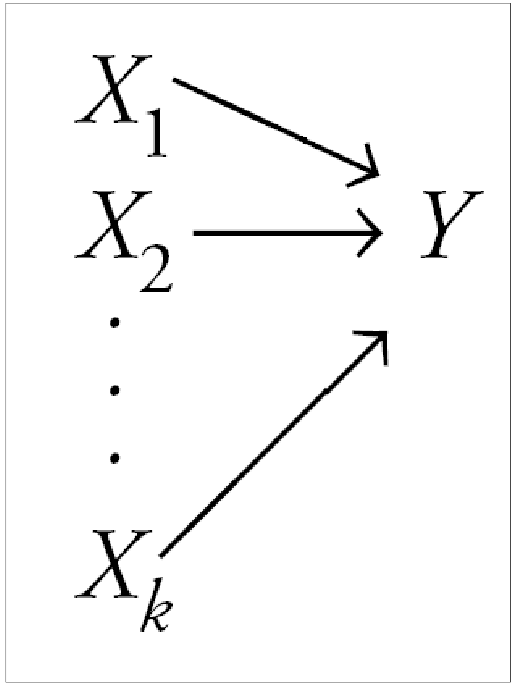
The General Idea

Simple regression considers the relation between a single explanatory variable and response variable

$$X \rightarrow Y$$

The General Idea

Multiple regression simultaneously considers the influence of multiple explanatory variables on a response variable Y



The intent is to look at the independent effect of each variable while “adjusting out” the influence of potential confounders

Simple Regression Model

Regression coefficients are estimated by minimizing $\sum \text{residuals}^2$ (i.e., sum of the squared residuals) to derive this model:

$$\hat{y} = a + bx$$

Multiple Regression Model

Again, **estimates for the *multiple* slope coefficients** are derived by minimizing ∑residuals^2 to derive this multiple regression model:

$$\hat{y} = a + b_1x_1 + b_2x_2$$

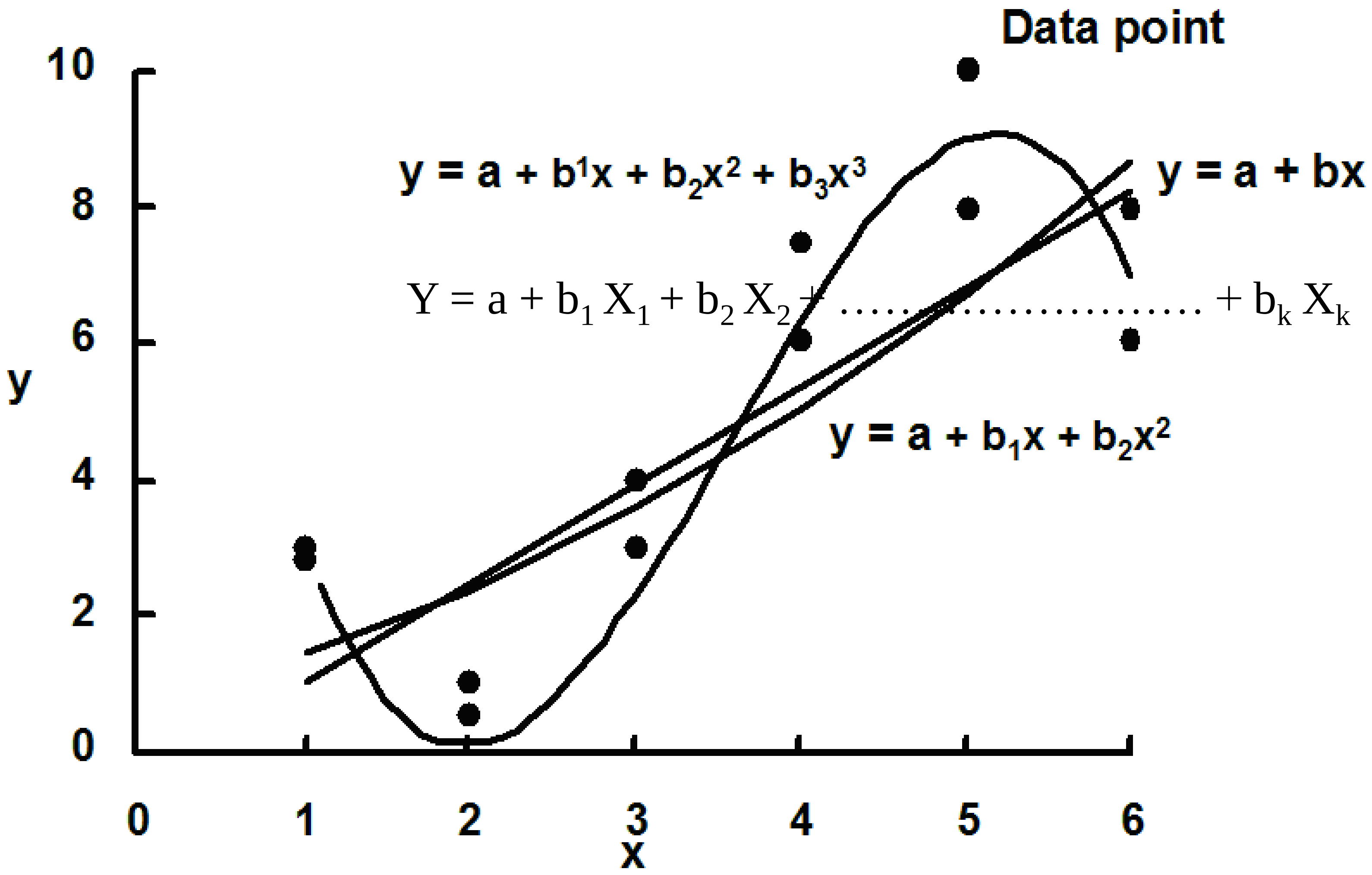
Multiple regression

- Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables.
- It also called as predictors.
- Method used for studying the relationship between a dependent variable and two or more independent variables.
- Purposes:
  - Prediction
  - Explanation
  - Theory building

- The variable whose value is to be predicted is known as the dependent variable.
- The ones whose known values are used for prediction are known Independent (exploratory) variables.

Design Requirements:

- One dependent variable (criterion)
- Two or more independent variables (predictor variables).
- Sample size: ≤ 50 (at least 10 times as many cases as independent variables)



GENERAL EQUATION:

In general, the multiple regression equation of Y on X_1, X_2, …, X_k is given by:

$$Y = a + b_1 X_1 + b_2 X_2 + \dots\dots\dots\dots\dots\dots\dots\dots + b_k X_k$$

Simple vs. Multiple Regression

- One dependent variable Y predicted from a set of independent variables (X1, X2 ….Xk)
- One regression coefficient for each independent variable
- R^2: proportion of variation in dependent variable Y predictable by set of independent variables (X's)

- One dependent variable Y predicted from one independent variable X
- One regression coefficient
- r^2: proportion of variation in dependent variable Y predictable from X

ADVANTAGE:

- Once a multiple regression equation has been constructed, one can check how good it is by examining the coefficient of determination($R2$). R2 always lies between 0 and 1.
- All software provides it whenever regression procedure is run. The closer R_2 is to 1, the better is the model and its prediction.

ASSUMPTIONS:

- Multiple regression technique does not test whether data is linear. On the contrary, it proceeds by assuming that the relationship between the Y and each of X_i's is linear. Hence as a rule, it is prudent to always look at the scatter plots of (Y, X_i), i= 1, 2,…,k. If any plot suggests non linearity, one may use a suitable transformation to attain linearity.

Multi Linear Regression

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

Multi 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.

Multi Linear Regression

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

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x_1}, \boldsymbol{x_2})$$

$$\boldsymbol{y} = \boldsymbol{a_0} + \boldsymbol{a_1 x_1} + \boldsymbol{a_2 x_2}$$

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

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

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

- 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.

Multi Linear Regression

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

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

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

- The coefficient of the multiple regression equation is given as

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

Multi 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

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

- 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

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

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

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

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

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

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

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

- $y = a_0 + a_1 x_1 + a_2 x_2$
- Hence, the constructed model is:
- $y = -1.69 + 3.48x_1 - 0.05x_2$

Idea: Examine the linear relationship between
1 dependent (Y) & 2 or more independent variables (X_i)

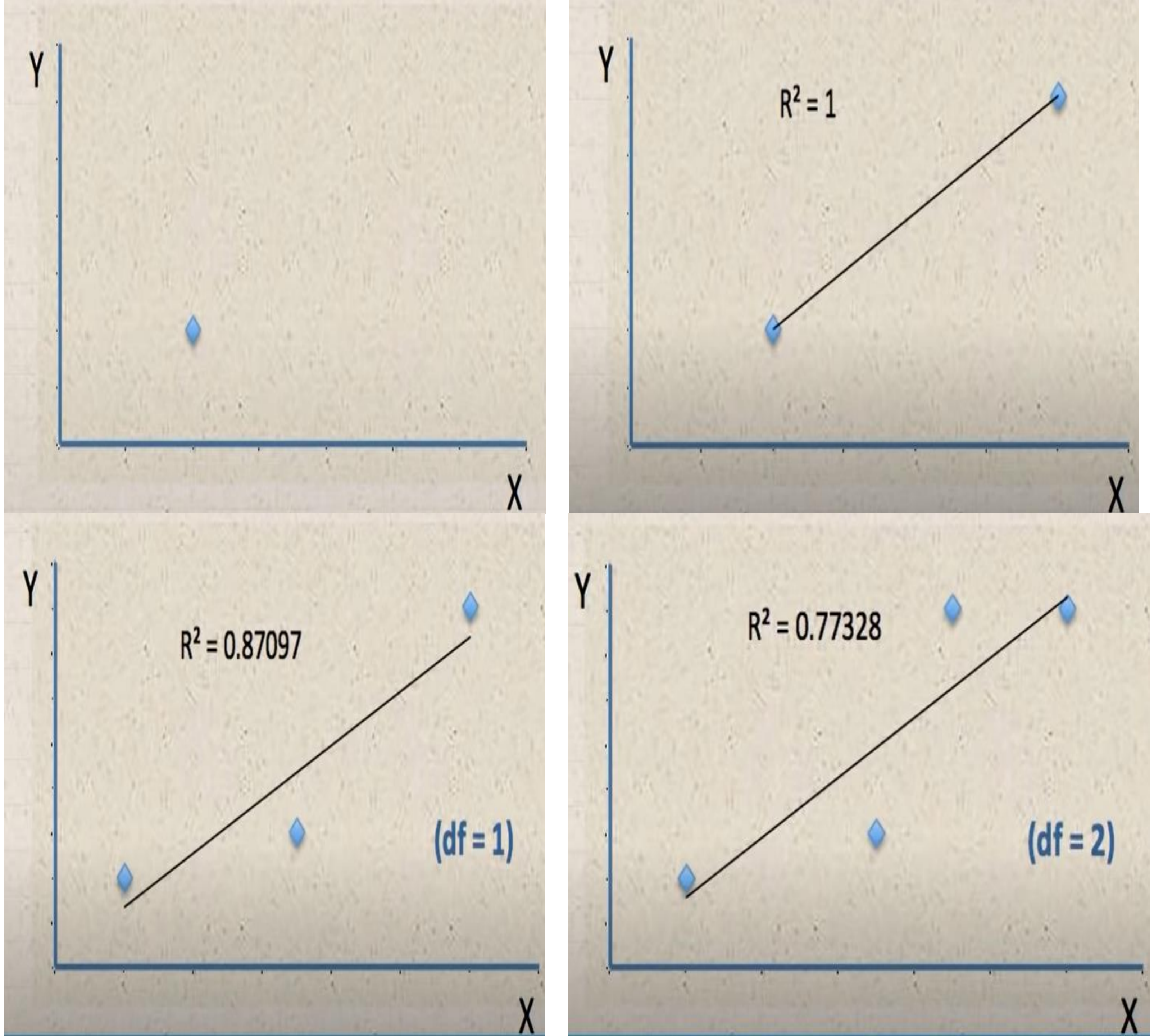
Multiple Regression Model with k Independent Variables:

$$Y = \underbrace{\beta_0}_{\text{Y-intercept}} + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \underbrace{\varepsilon}_{\text{Random Error}}$$

Population slopes

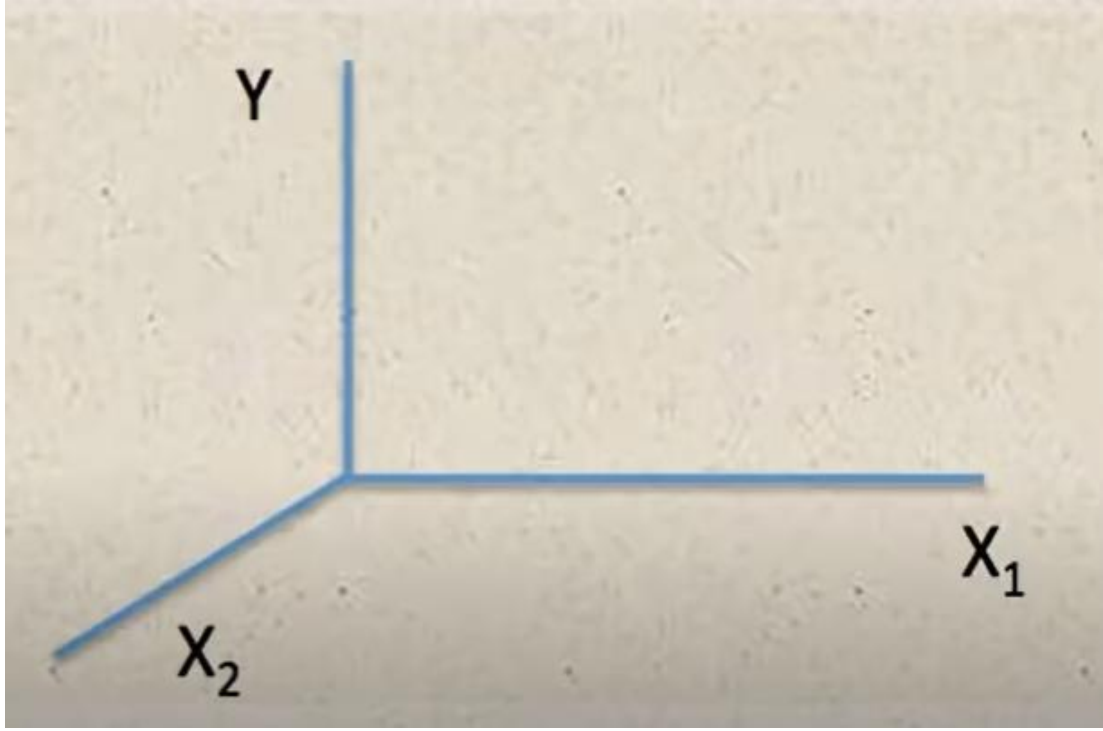
Adjusted R^2

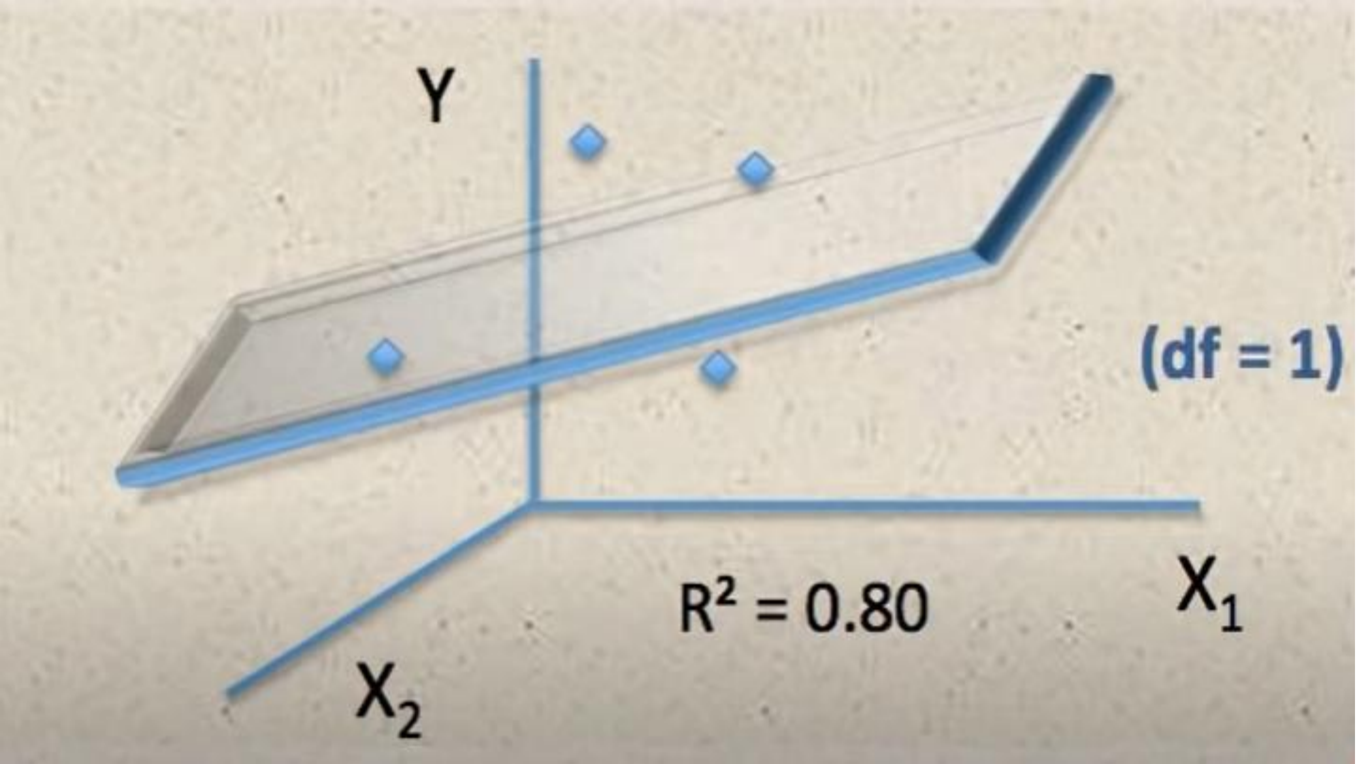
- What is the minimum no of observations required to estimate the regression ?

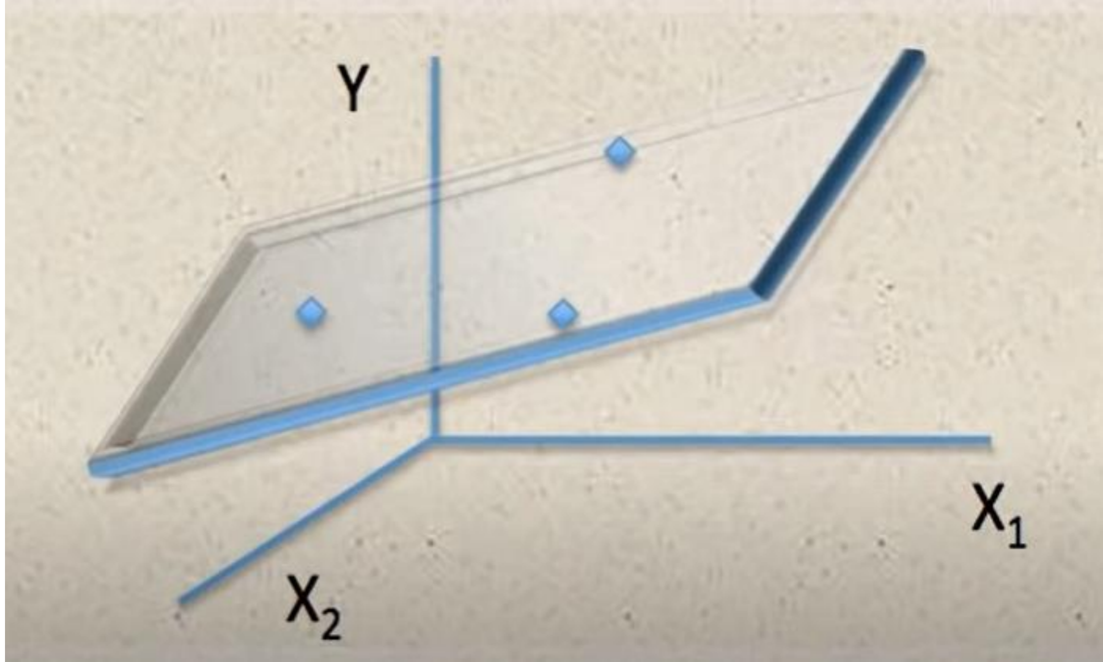


Additional varaiable

$y = b_0 + b_1 x_1 + b_2 x_2 \varepsilon$







Which leads us to the particular equation

df = n-k-1

Where : n is the no of observations

k is the no of variables

if k increasers

df decreases

df= 4 – 2 – 1 = 1 (for two variable)
df= 4 – 3 – 1 = 0 (for three variable)

How does the degree of freedom related to R^2

- As df decreases (i.e more variables are added to the model)
- R^2 only increases
- Because the degree of freedom is decreased so R^2 only increases

Adjusted R^2

$$R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Where K increases adjusted R^2 decreases

R^2

number of observations, n	number of variables, k	R^2
25	4	0.71
25	5	0.76
25	6	0.78
25	7	0.79
10	4	0.71
10	5	0.76
10	6	0.78
10	7	0.79

Adjusted R^2

number of observations, n	number of variables, k	R^2	Adj-R^2
25	4	0.71	0.6520
25	5	0.76	0.6968
25	6	0.78	0.7067
25	7	0.79	0.7035
10	4	0.71	0.4780
10	5	0.76	0.4600
10	6	0.78	0.3400
10	7	0.79	0.0550

Linear and Non Linear

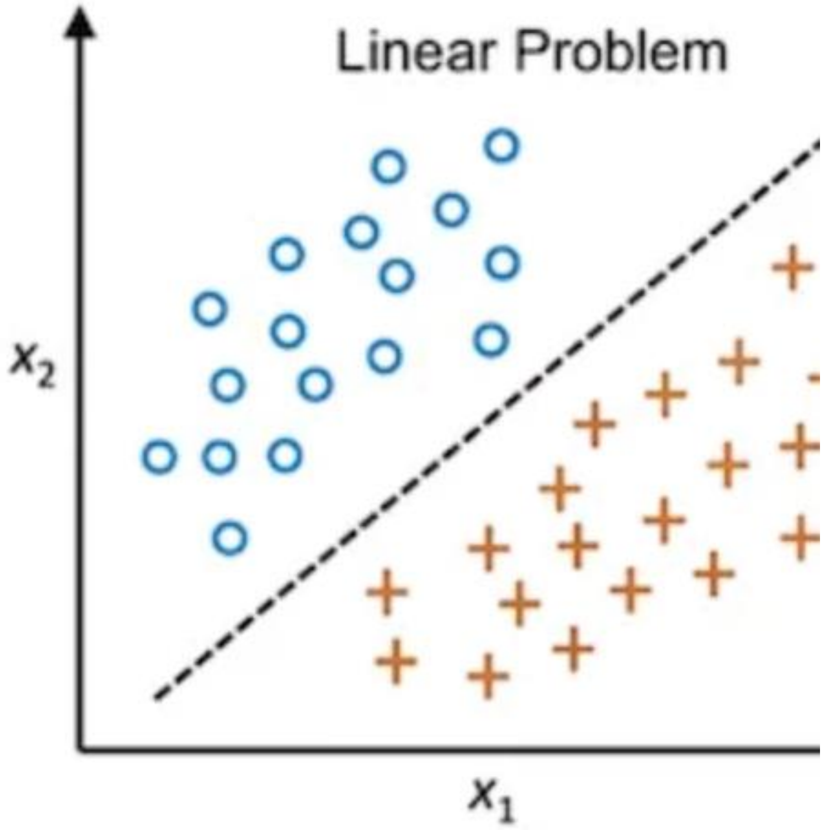
Linear Classification refers to categorizing a set of data points to a discrete class based on a linear combination of its explanatory variables.

On the other hand, Non-Linear Classification refers to separating those instances that are not linearly separable.

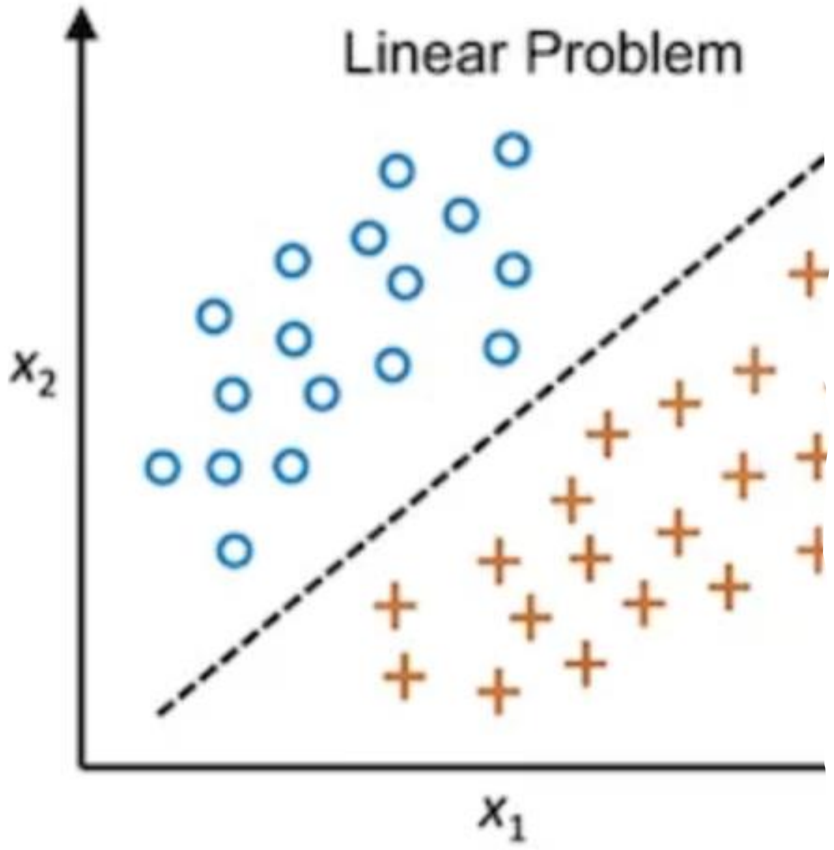
Linear Classification

Linear Classification refers to categorizing a set of data points into a discrete class based on a linear combination of its explanatory variables.

Some of the classifiers that use linear functions to separate classes are *Linear Discriminant Classifier, Naive Bayes, Logistic Regression, Perceptron, SVM (linear kernel).*

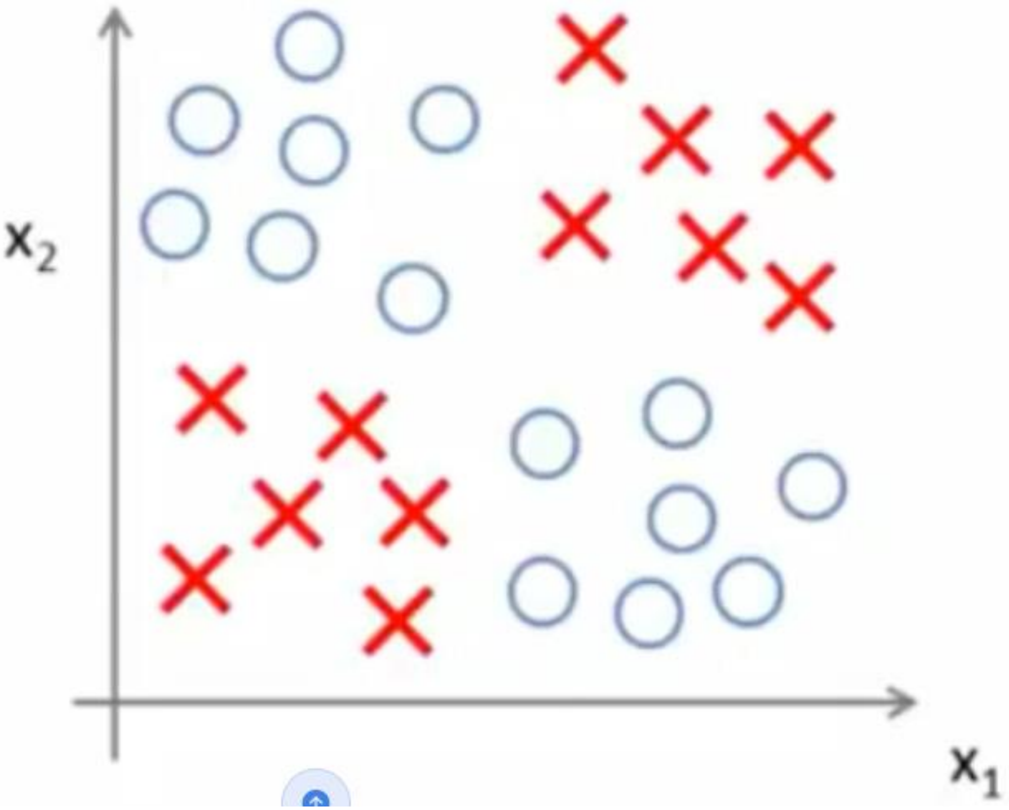


Linear Classification



- In the figure, we have two classes, namely 'O' and '+.' To differentiate between the two classes, an arbitrary line is drawn, ensuring that both the classes are on distinct sides.
- Since we can tell one class apart from the other, these classes are called ‘linearly-separable.’
- However, an infinite number of lines can be drawn to distinguish the two classes.
- The exact location of this plane/hyperplane depends on the type of the linear classifier.

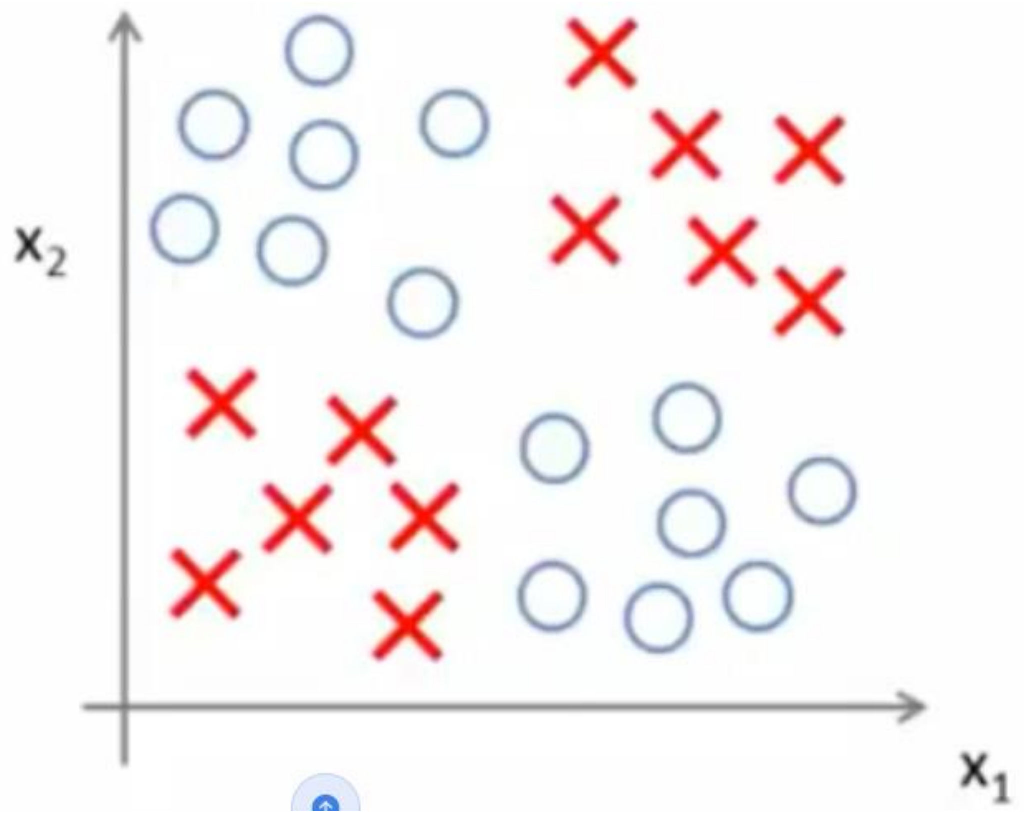
Non Linear Classification



- Non-Linear Classification refers to categorizing those instances that are not linearly separable.

- Some of the classifiers that use non-linear functions to separate classes are Quadratic Discriminant Classifier, Multi-Layer Perceptron (MLP), Decision Trees, Random Forest, and K-Nearest Neighbours (KNN).

Non Linear Classification



- In the above, we have two classes, namely 'O' and 'X.' To differentiate between the two classes, it is impossible to draw an arbitrary straight line to ensure that both the classes are on distinct sides.
- We notice that even if we draw a straight line, there would be points of the first-class present between the data points of the second class.
- In such cases, piece-wise linear or non-linear classification boundaries are required to distinguish the two classes.

Linear Regression

1. Home prices
2. Weather
3. Stock price

Predicted value is continuous

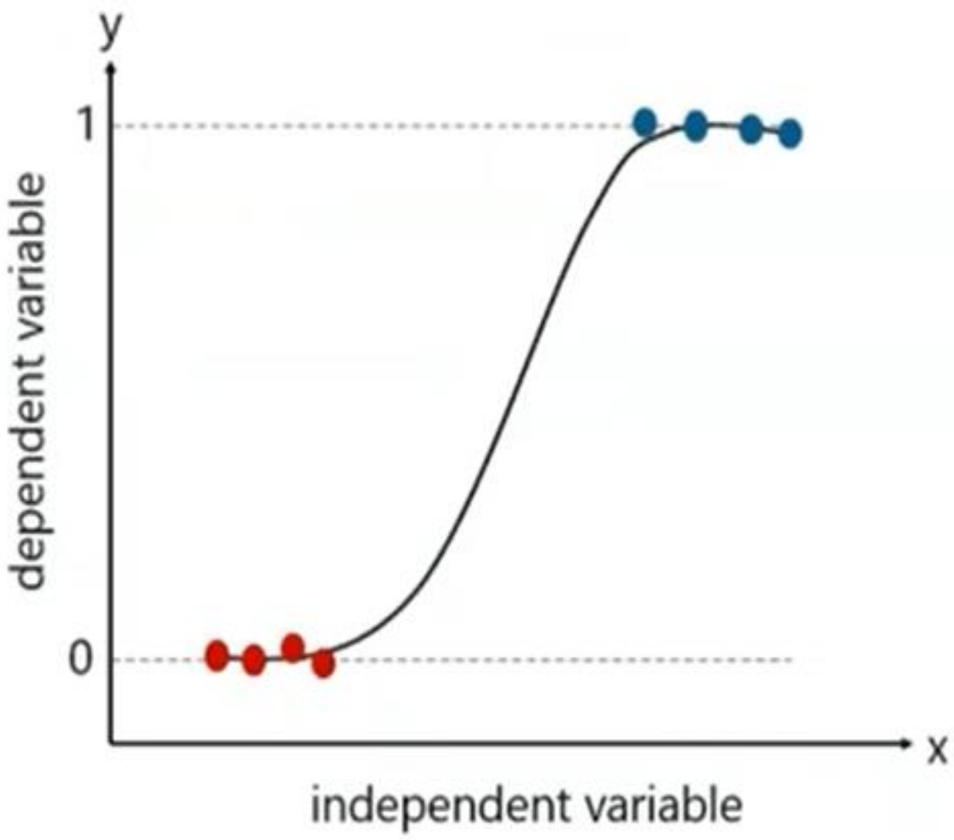
Classification

1. Email is spam or not
2. Will customer buy life insurance?
3. Which party a person is going to vote for?
    1. Democratic
    2. Republican
    3. Independent

Predicted value is categorical

What Is Logistic Regression?

Logistic Regression is a method used to predict a dependent variable, given a set of independent variables, such that the dependent variable is categorical.



- *Dependent variable (Y):*

The response binary variable holding values like 0 or 1, Yes or No, A, B or C

- *Independent variable (X):*

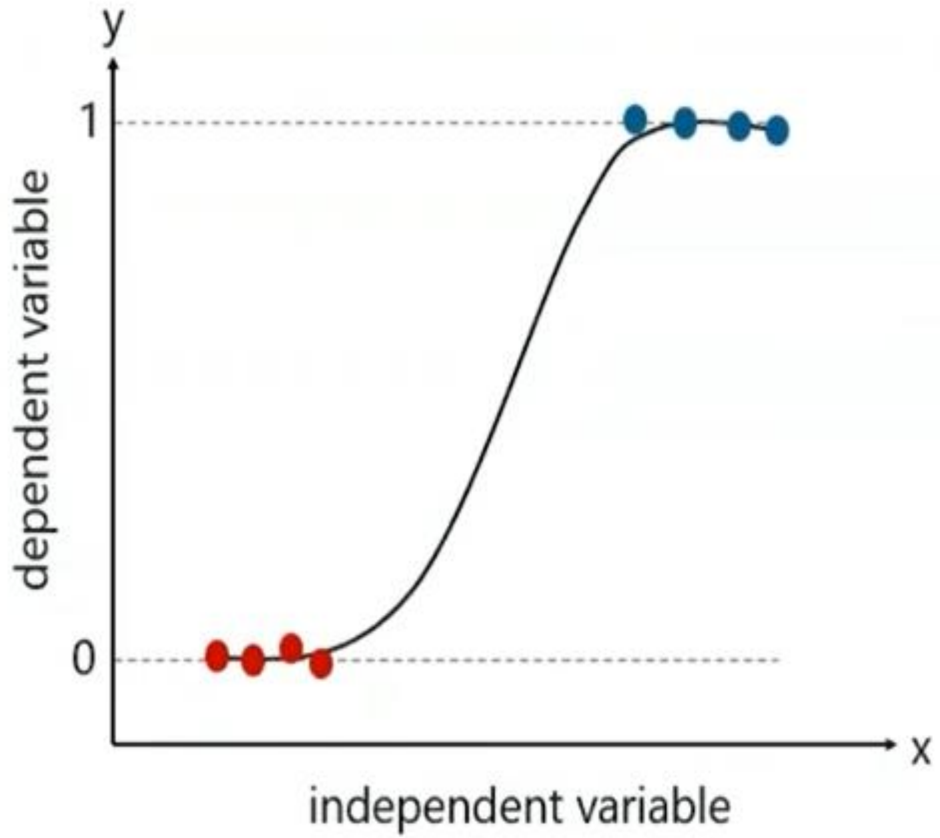
The predictor variable used to predict the response variable.

The following equation is used to represent a linear regression model:

$$\log\left(\frac{Y}{1-Y}\right) = C + B1X1 + B2X2 + ...$$

What Is Logistic Regression?

Logistic Regression is a method used to predict a dependent variable, given a set of independent variables, such that the dependent variable is categorical.



$$\log\left(\frac{Y}{1-Y}\right) = C + B1X1 + B2X2 + \ldots.$$

- Y is the probability of an event to happen which you are trying to predict
- x1, x2 are the independent variables which determine the occurrence of an event i.e. Y
- C is the constant term which will be the probability of the event happening when no other factors are considered

Logistic Regression

- Is not used for regression , **it is used for classification**

- Why the name is **Logistic Regression**

  It uses the underlying principle of simple linear regression

Binary outcomes instead of continuous outcomes

Interested in probability of the outcome for a given value of the independent varaible

Classification Types

Will customer buy life insurance?

1. Yes
2. No

Binary Classification

Which party a person is going to vote for?

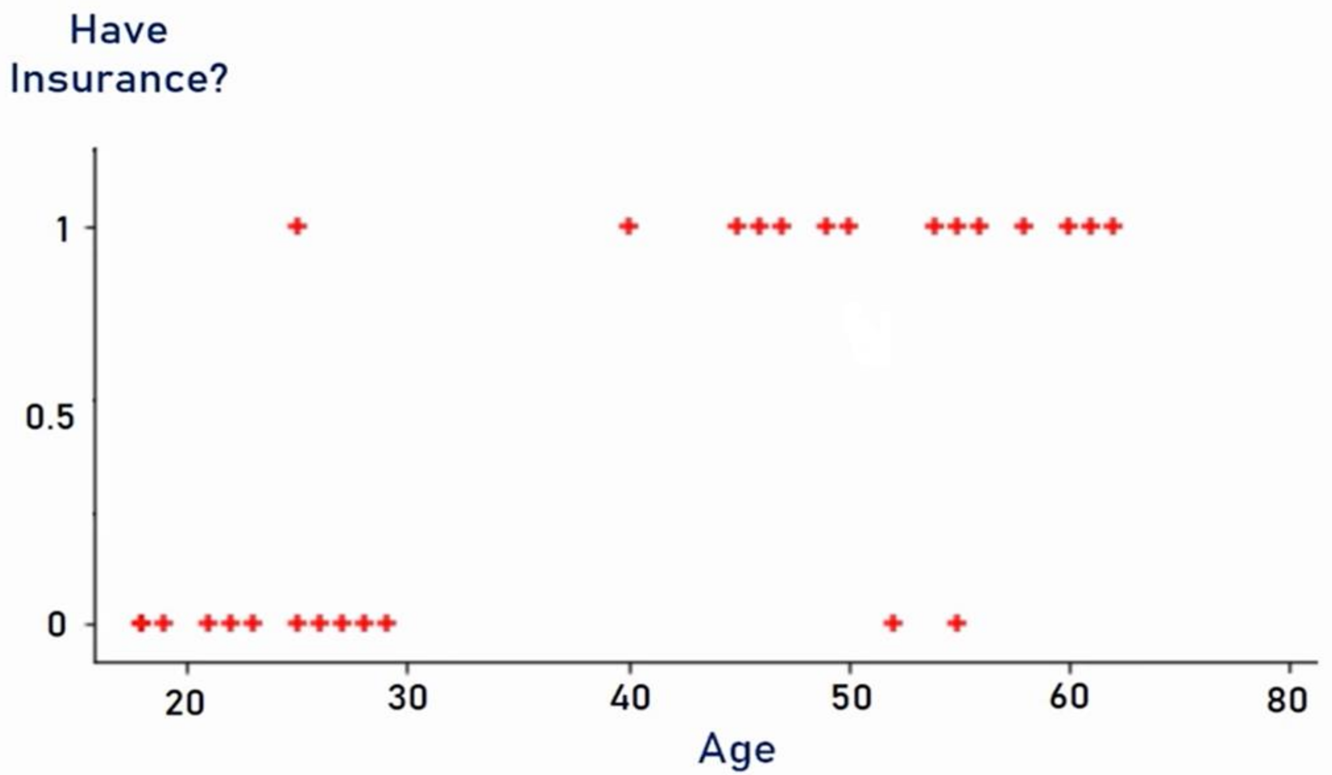
1. Democratic
2. Republican
3. Independent

Multiclass Classification

age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0
49	1

- You want to build a machine learning model that can probably do prediction based on the age of the customer **whether they bought the insurance or not!**

- So you will plot the scatter plots for the given data

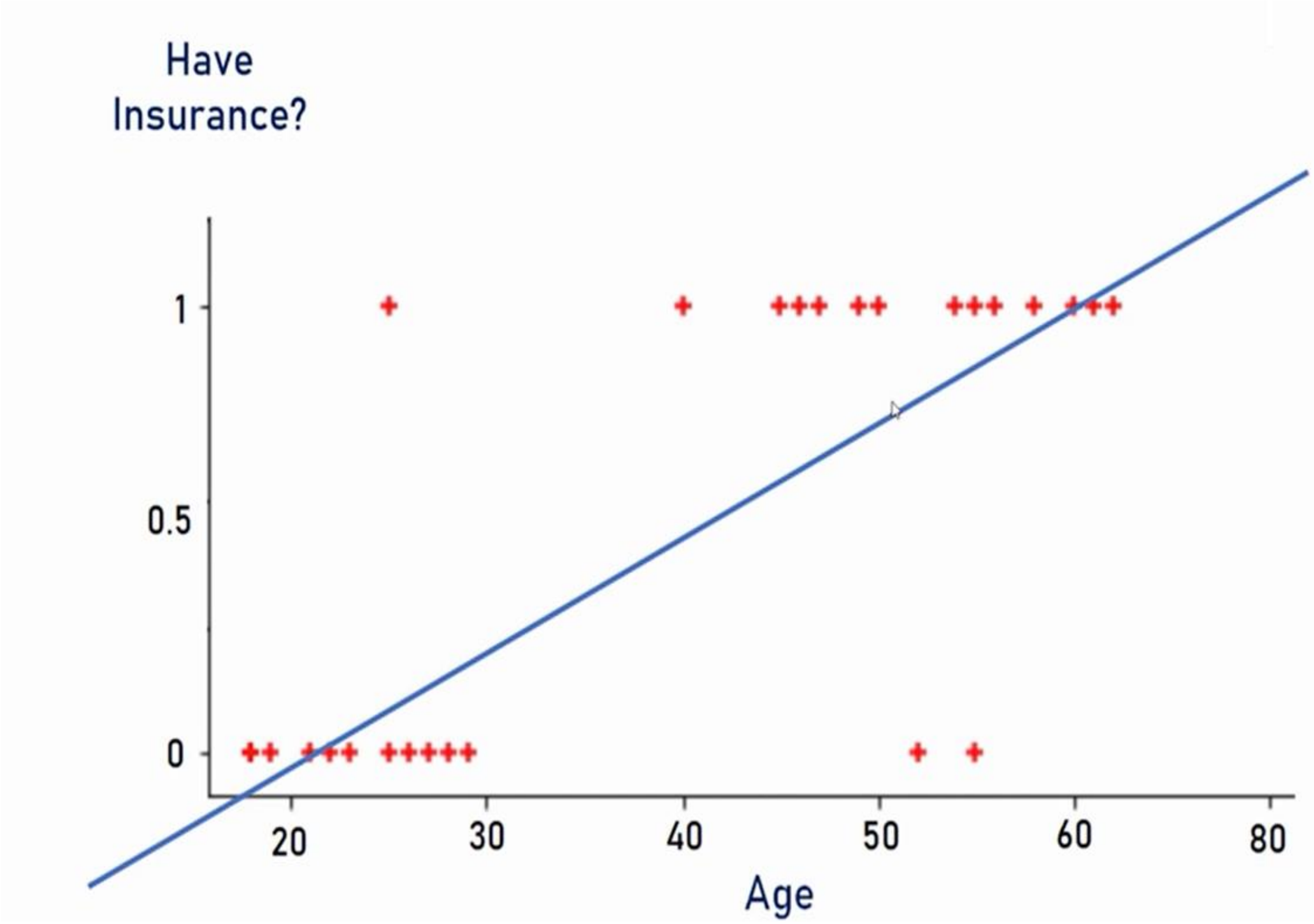


age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0
49	1

- **There is a pattern**
- **Young People – No**
- **Old People – yes**
- **Based on this you are building a Machine learning model based on the age of the customer**

age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0
49	1

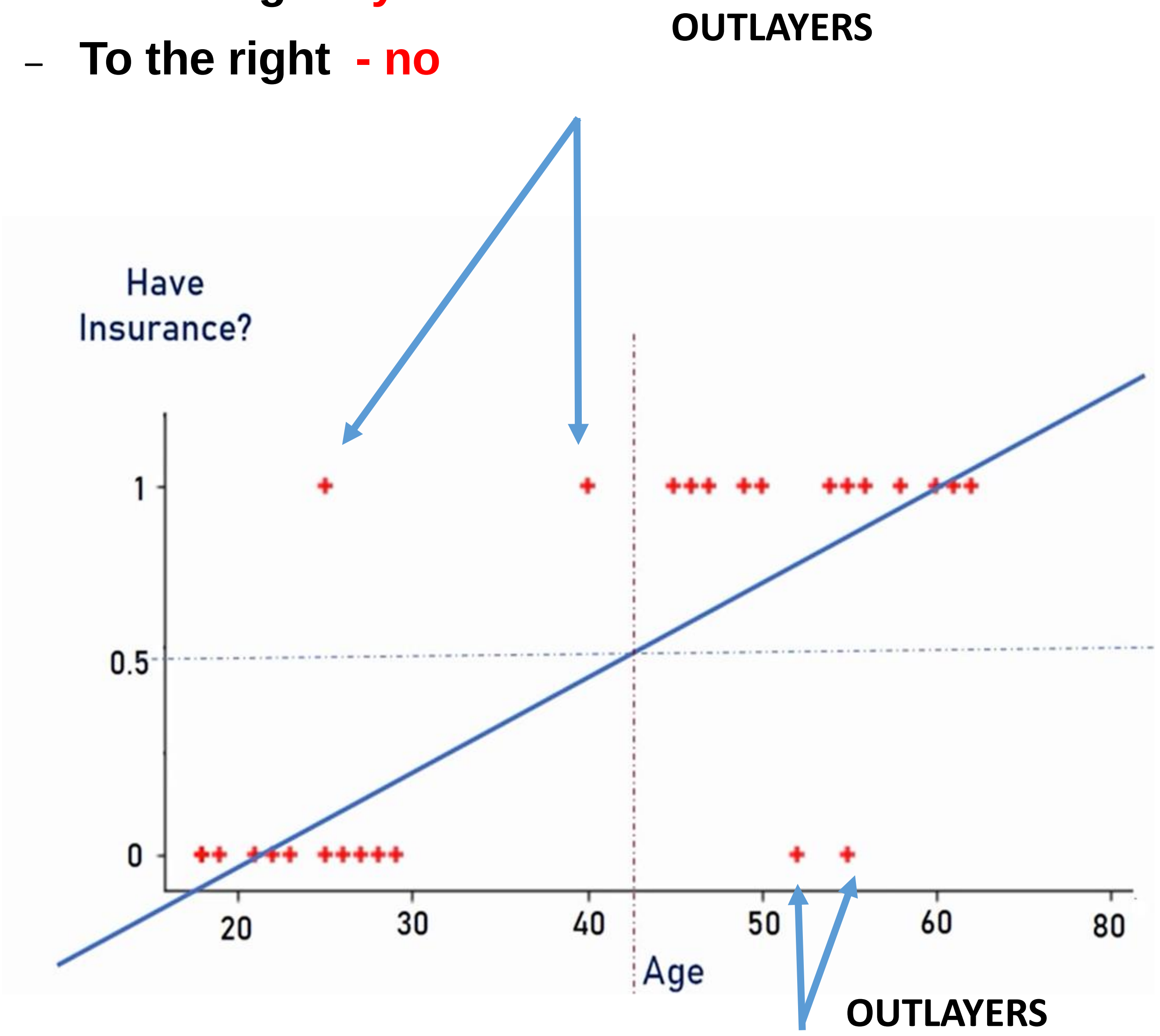
– **Linear Regression**



age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0
49	1

- **Linear Regression**

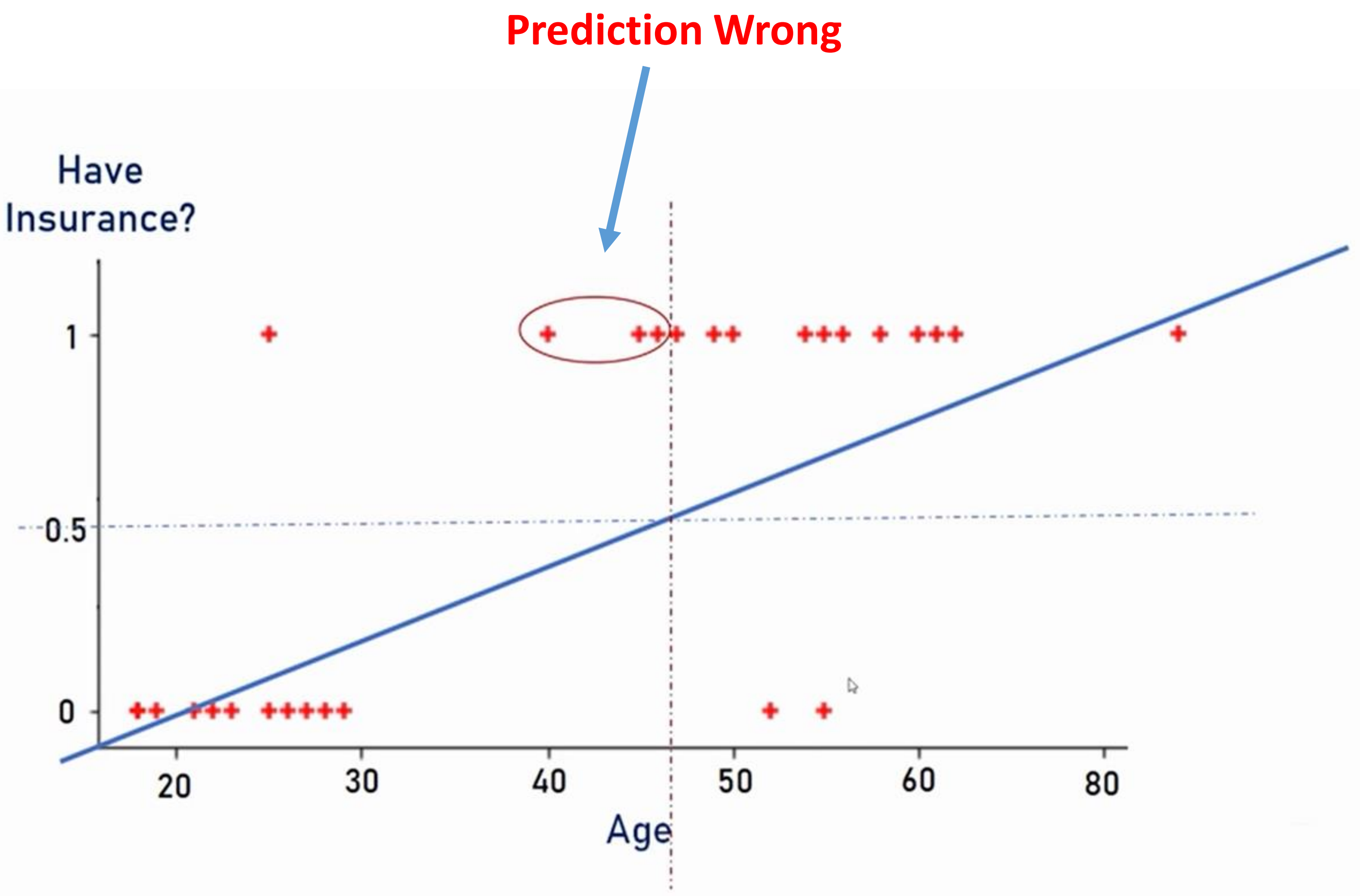
- **To the right - yes**
- **To the right - no**



age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0
49	1

Linear Regression (LR)

- **If there is a customer of age 80 then my linear graph will look as mentioned in below figure**
- **Therefor LR is not a good choice if your data is categorical**

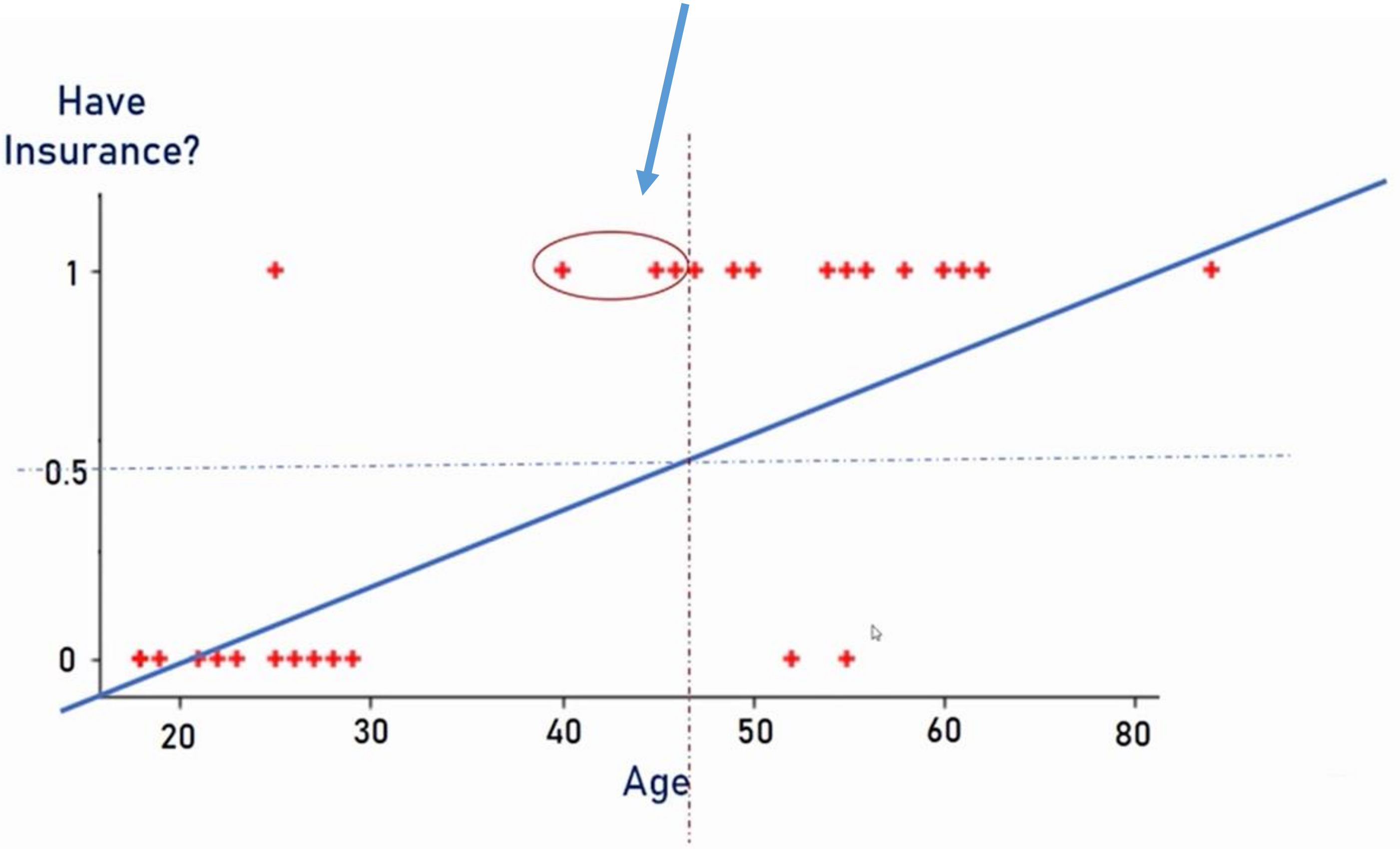


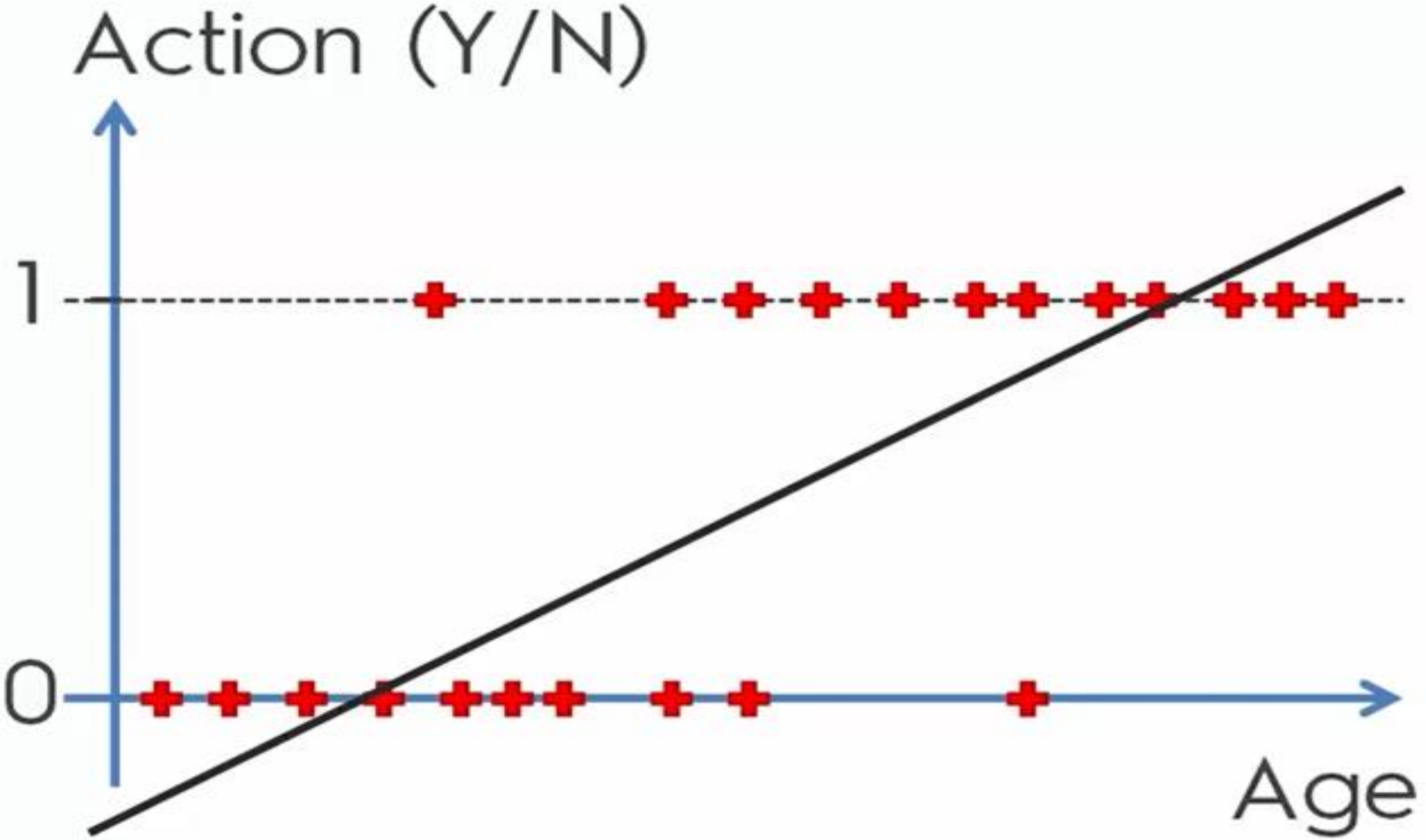
age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0
49	1

Linear models have a draw back of decision boundary from - ∞ to + ∞

- We need to squeeze the linear line in to some function , In Log Reg we use sigmoid funct to squeeze the line such that values fall between 0-1

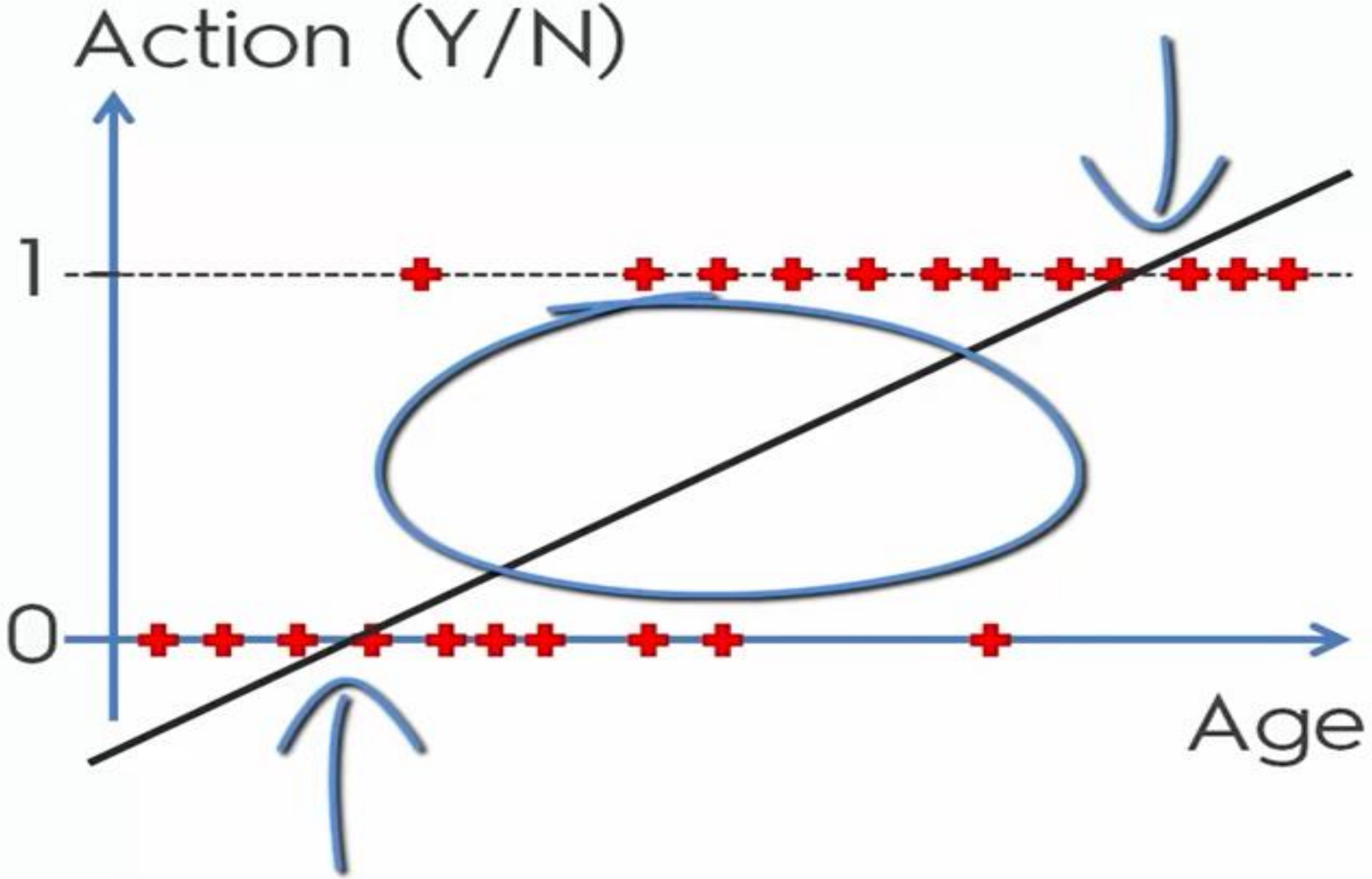
Prediction Wrong



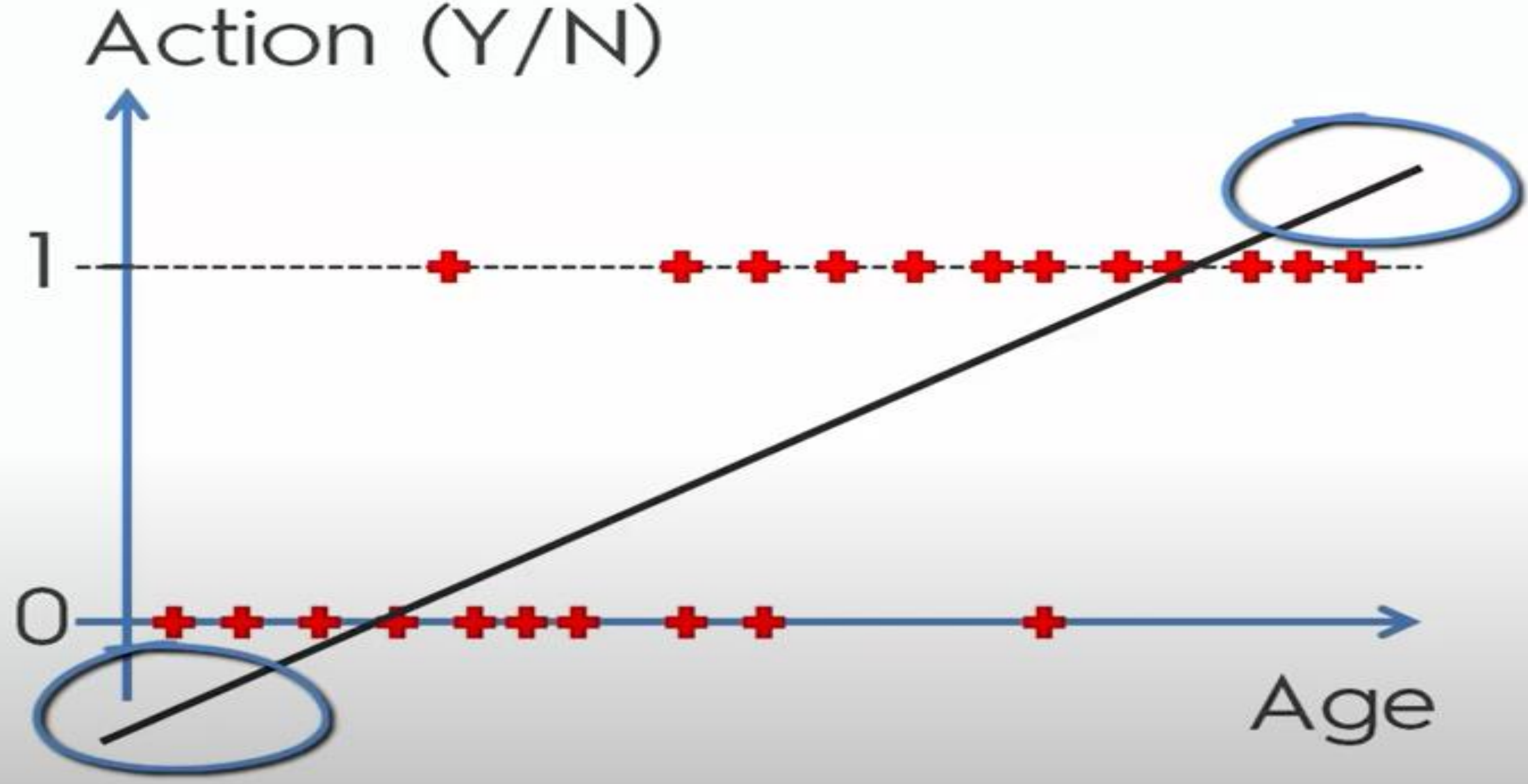


Action (Y/N)
1
0
Age

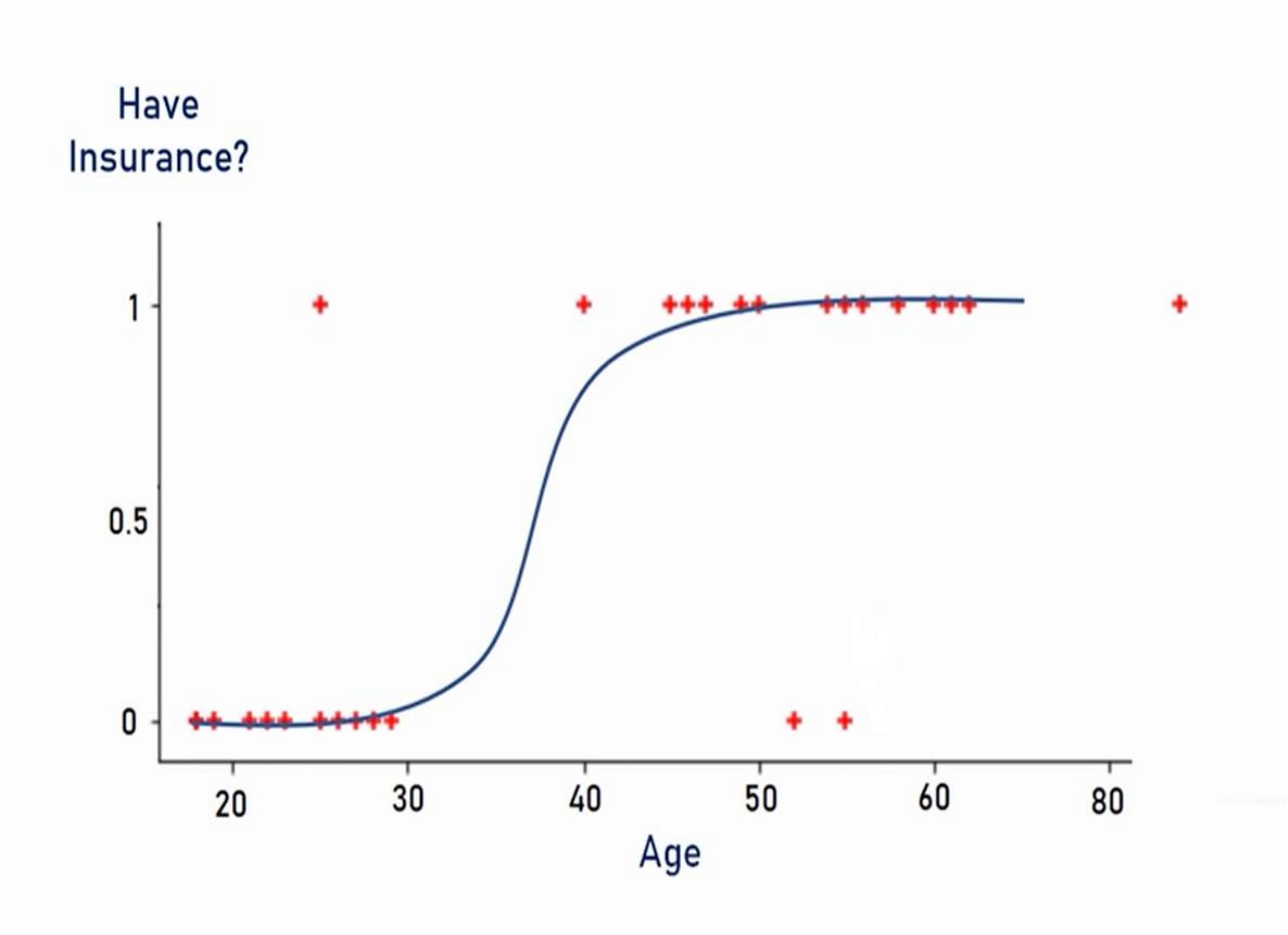
What is the probability between age 35 – 55



Probability can never be between 0 or above 1

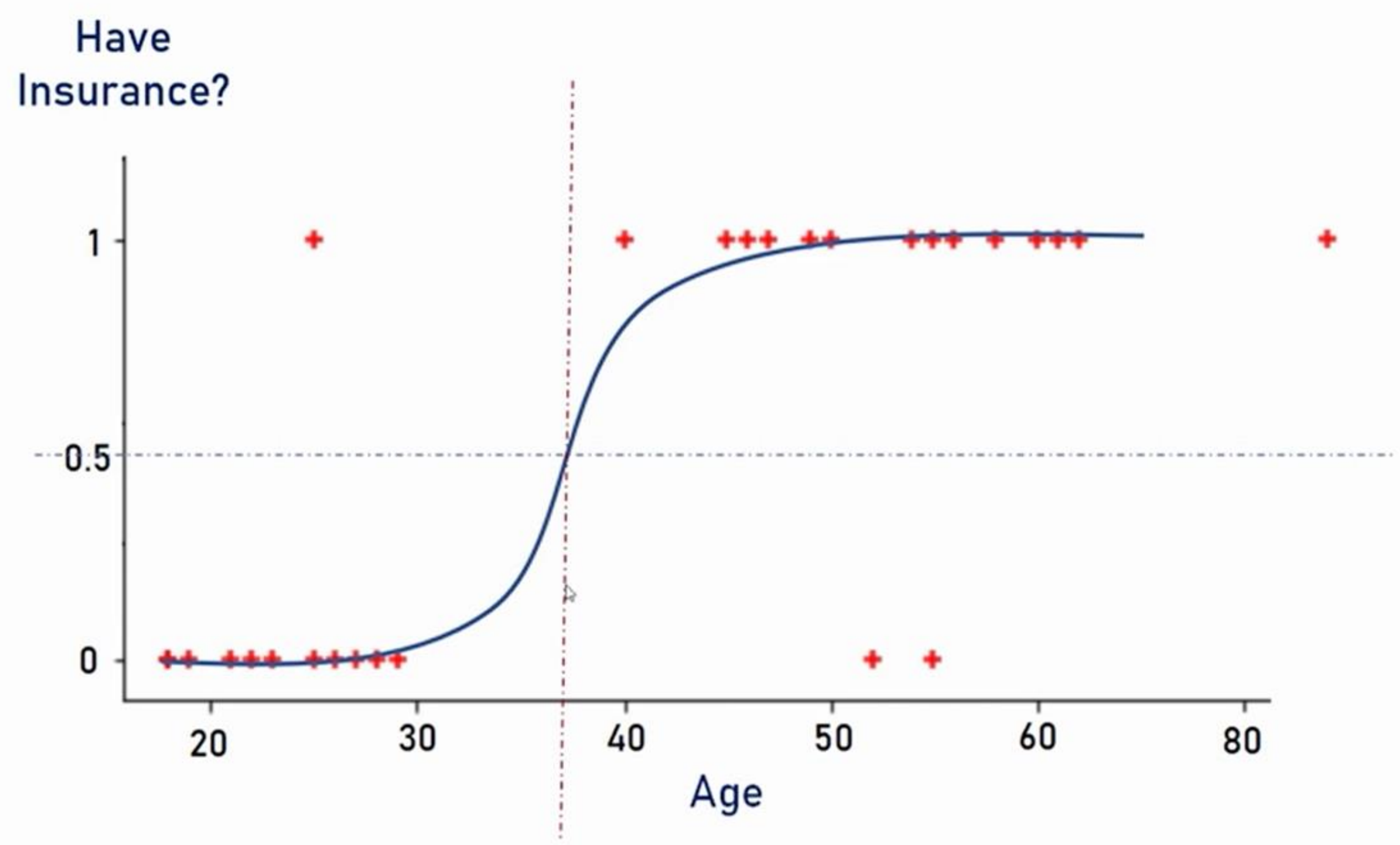


Imagine if you can draw a line like mentioned below



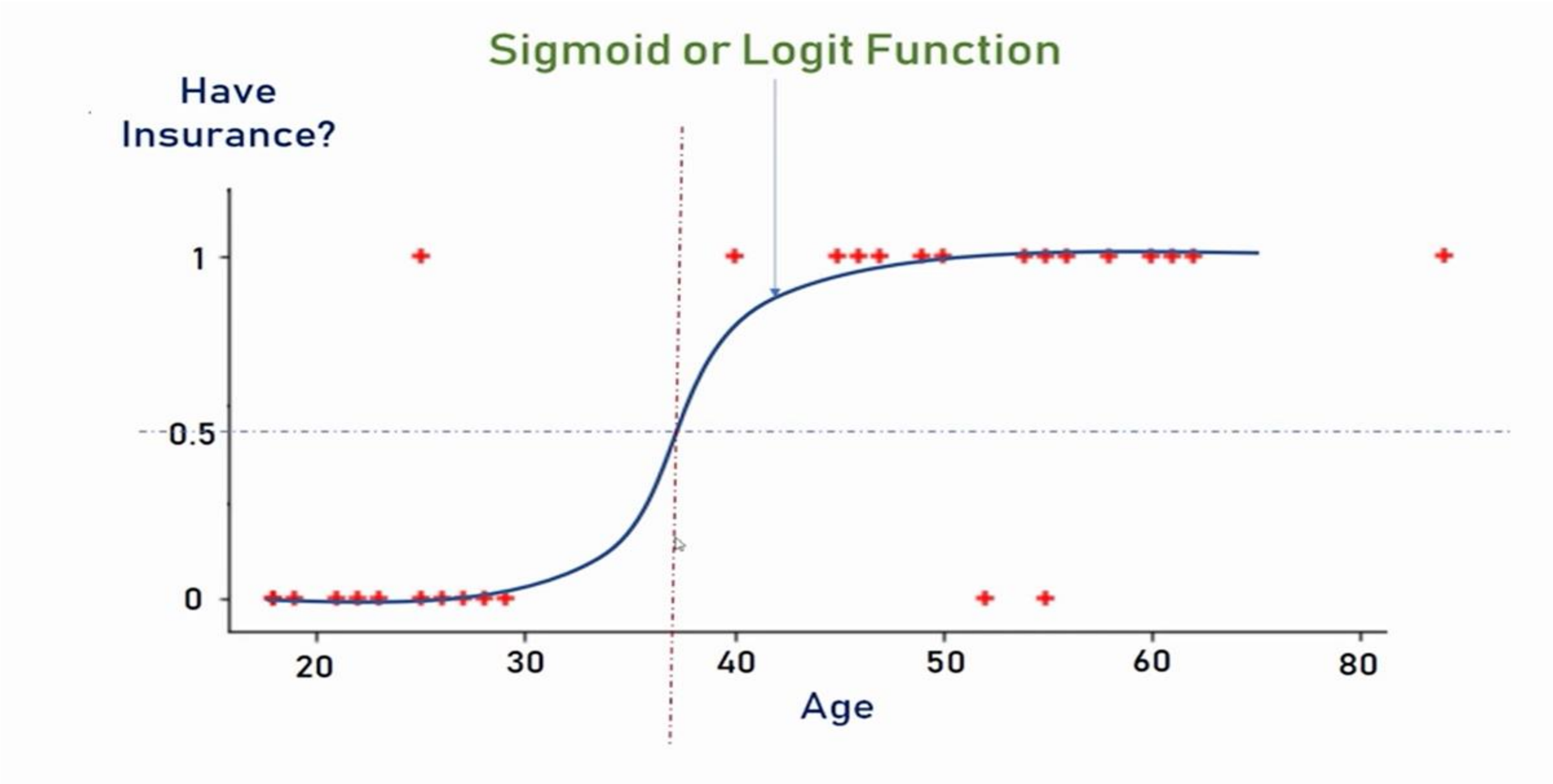
- This is a much **better fit** compared to the linear equation
- This model works better than the previous LR

Imagine if you can draw a line like mentioned below



- How do you come up with this type of **S** curve line that fits the model which has categorical data

Using Sigmoid/Logit



SigmoidFunction

$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

e = Euler's number ~ 2.71828

Sigmoid function converts input into range 0 to 1

Odds: Chances of happening one event over chances of this event not happening

If you give a set of numbers to **sigmoid function** it will **convert that number between 0-1** and the graph you get is S shape

Sigmoid Function

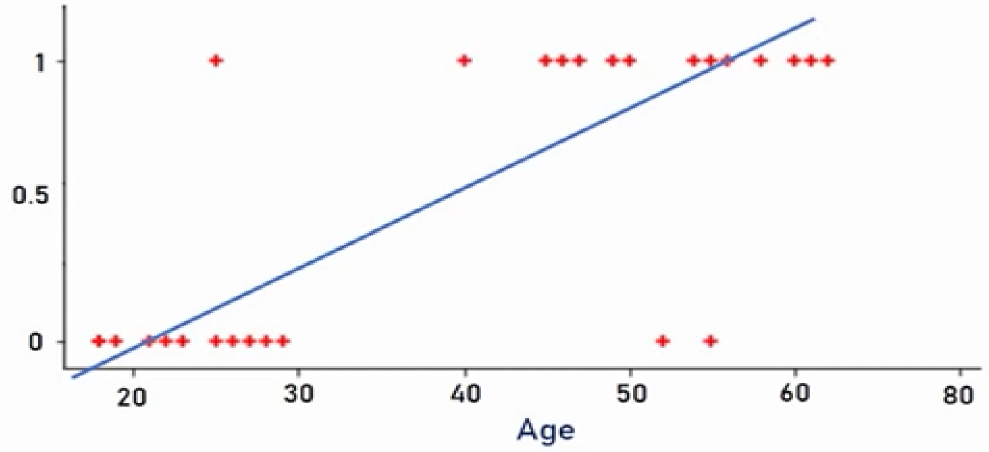
convert that number between 0-1 and the graph you get is S shape , Applied the sigmoid function on top of the linear equation that’s how we get S shape curve that fits better for classification problem

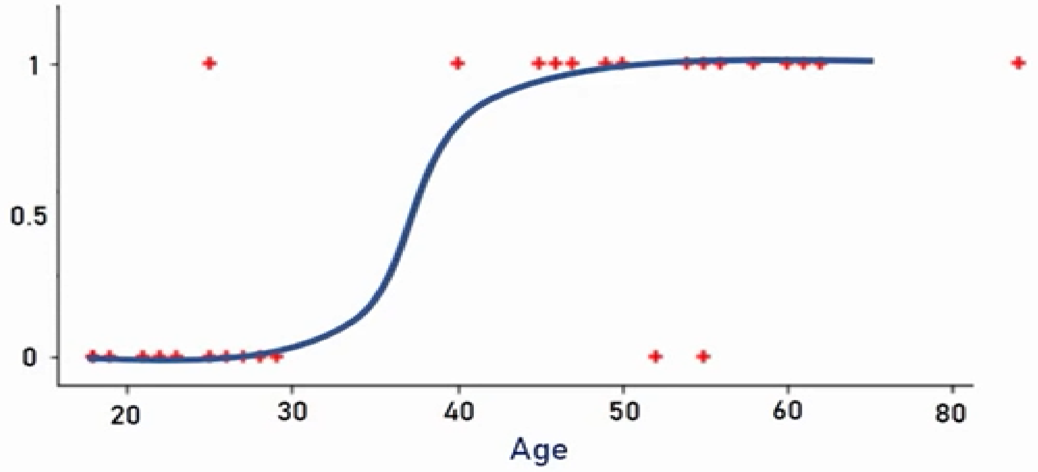
Input(can be continuous) can be from - ∞ to + ∞

output will be from 0 to 1

$$y = m * x + b$$

$$y = \frac{1}{1 + e^{-(m*x+b)}}$$



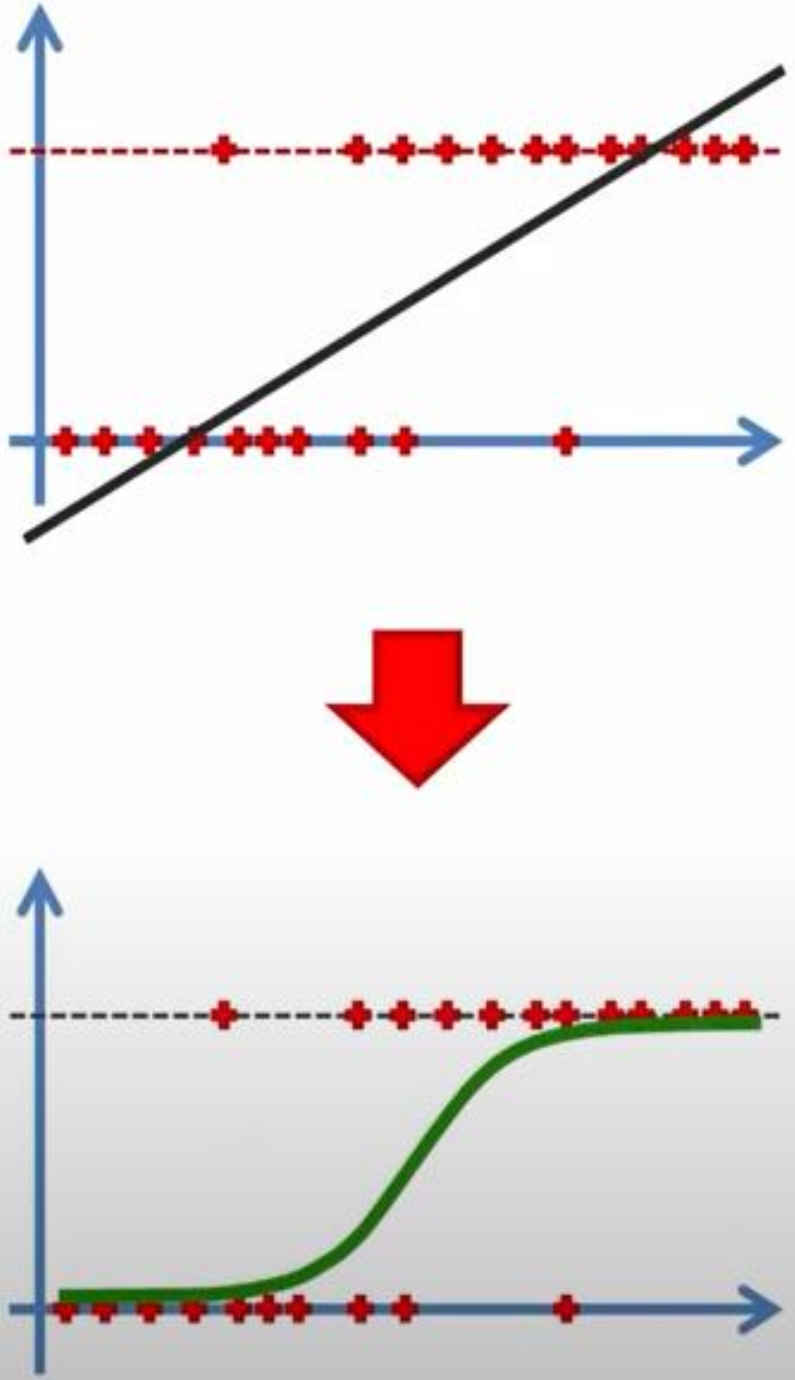


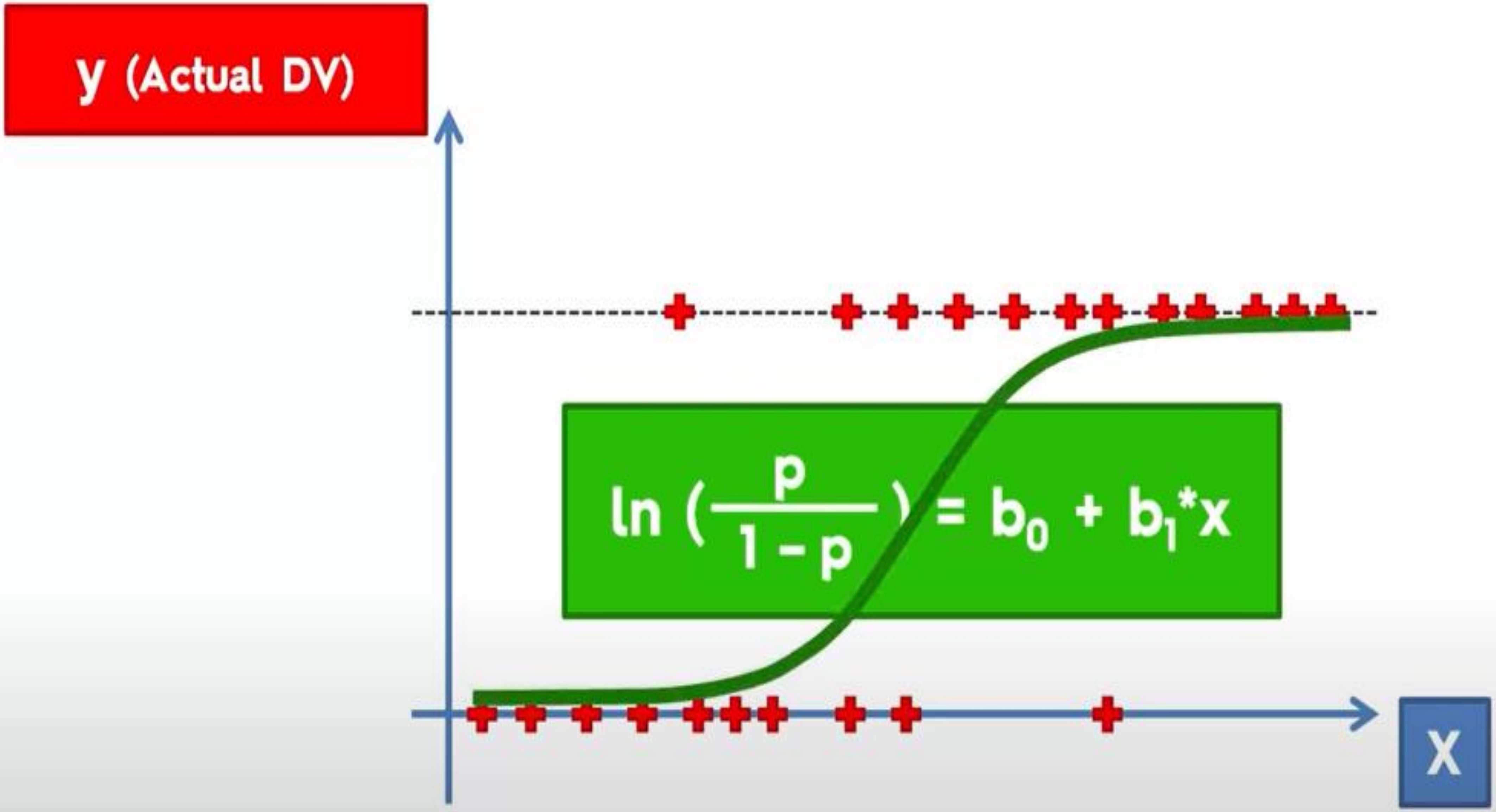
$$y = b_0 + b_1 * x$$

Sigmoid Function

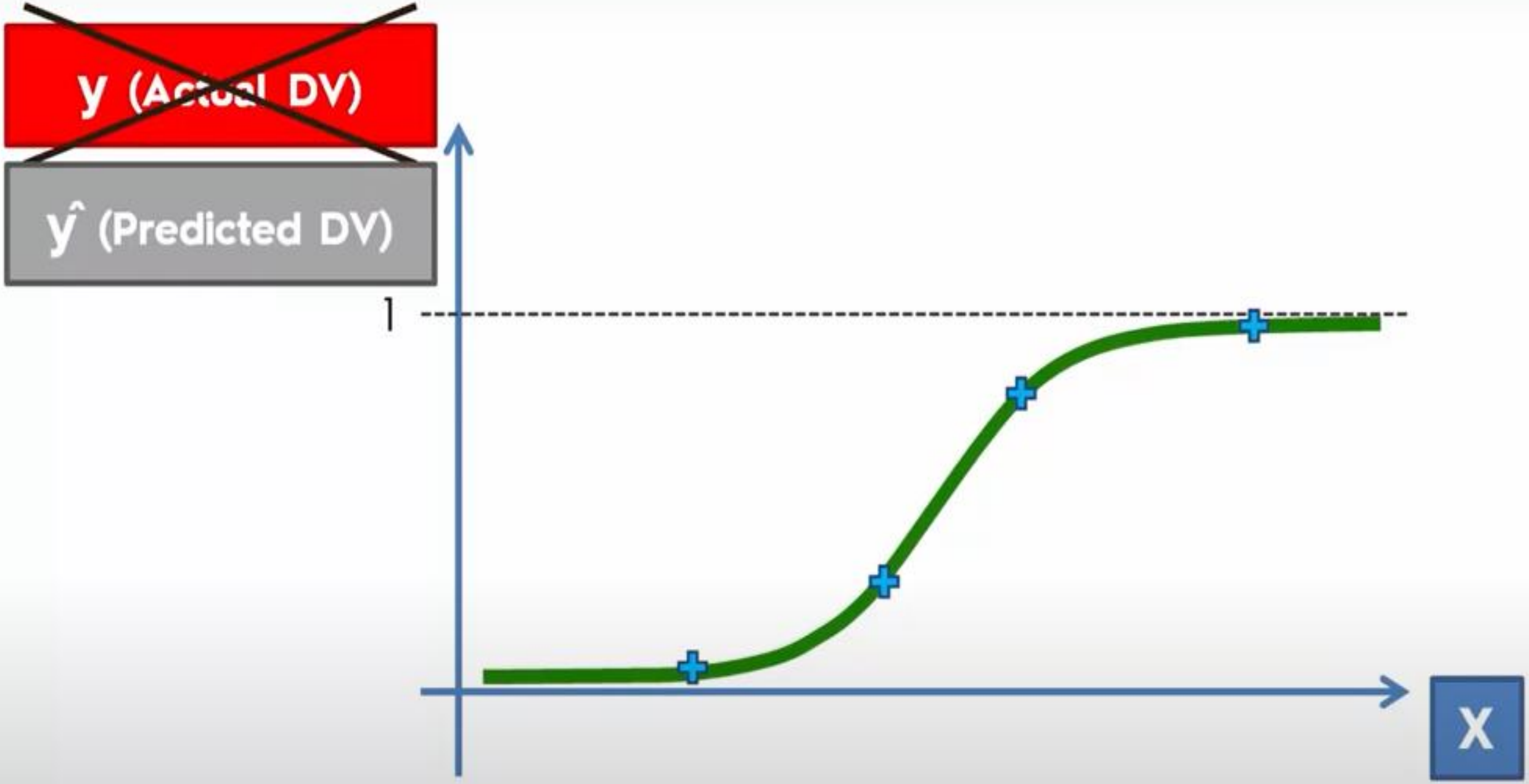
$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$





y (Actual DV)
$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1*x$
X



y (Actual DV)
ŷ (Predicted DV)
1
X

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

- $p = \frac{1}{1+e^{-z}}$

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

- $z = -64 + 2 * 33 = -64 + 66 = 2$
- $p = \frac{1}{1+e^{-2}} = 0.88$
- 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

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

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

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

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

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

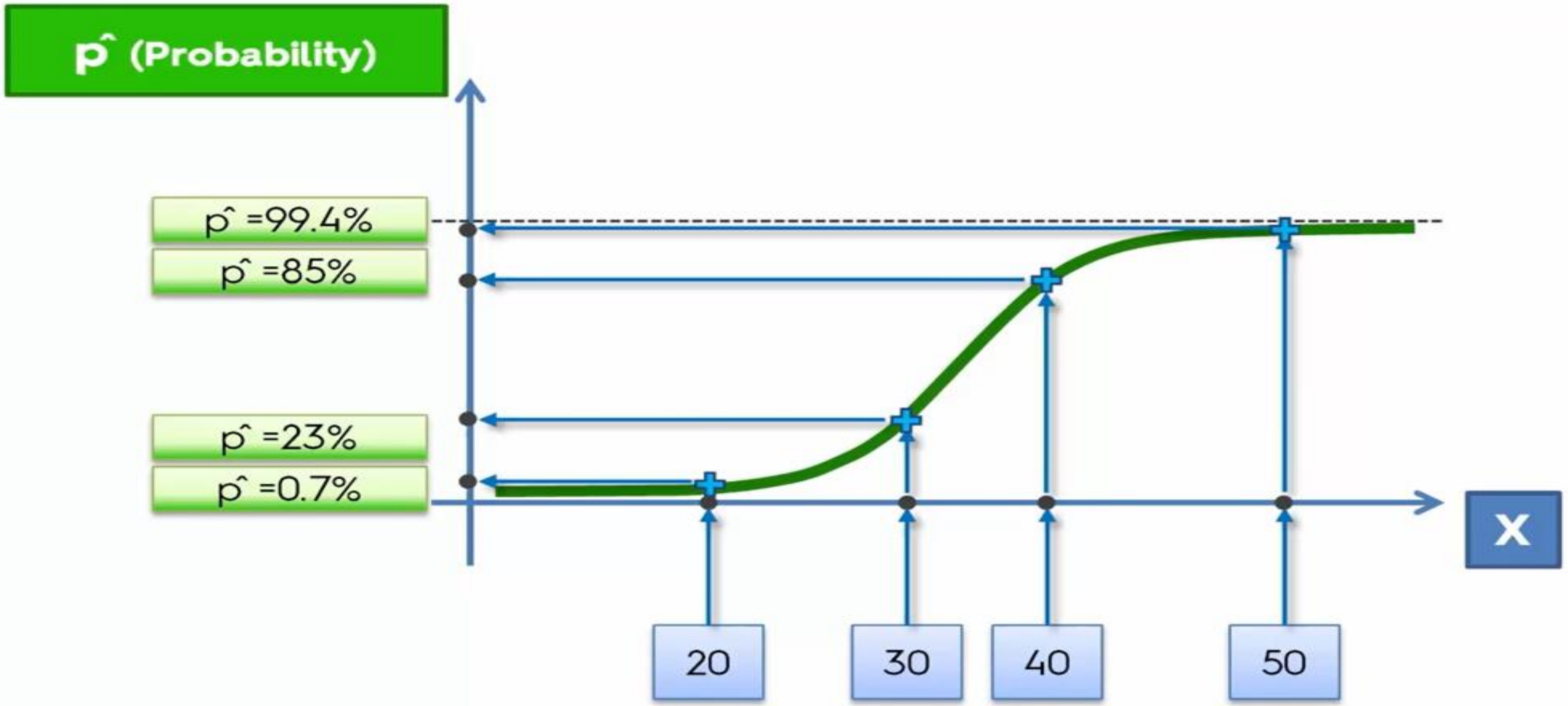
$$\boldsymbol{z = 2.94}$$

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

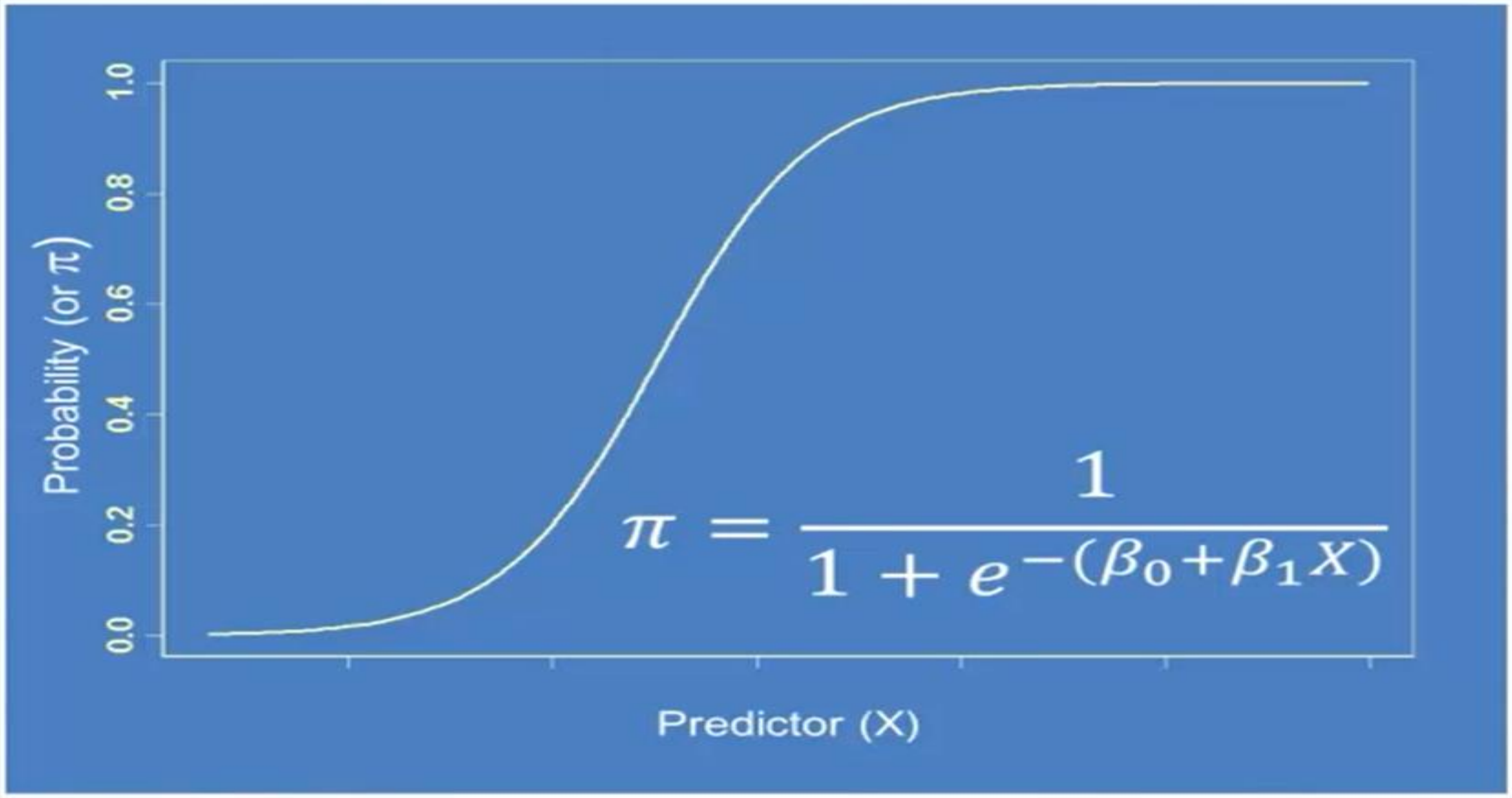
Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

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

X (input)	F(x)=X	Output of Sigmoid function
1	F(1) = 0.73	0.73
10	F(10) = 0.99	0.99
100	F(100) = 1	1
-1	F(-1) = 0.268	0.268



S Curve


1.0
0.8
0.6
0.4
0.2
0.0
Probability (or π)
Predictor (X)
$\pi = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$

S curve to Logit

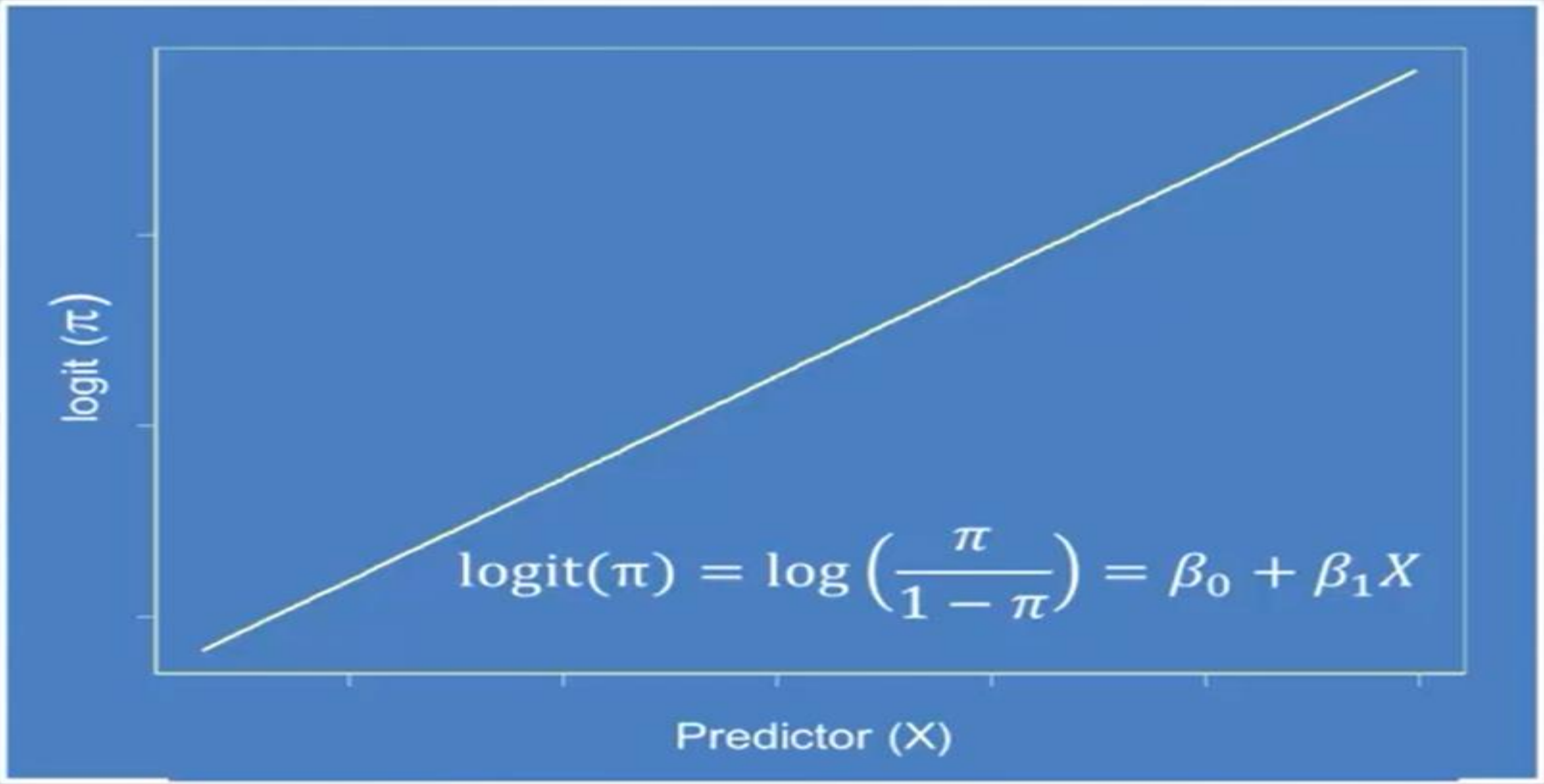
- To transform nonlinear S curve relationship between predictor and outcome to simple linear relationship

- To figure out intercept and slopes for the equation of the straight line

- Convenient way to transform **into linear relationship using logistic function**

- We apply logistic funct to probability values graphed in the figure and plot the new values on the vertical axis vs the values of the predictors on the X axis we obtain Linear in the logit.

Linear in the Logit



Stretching and Movement for the S curve

- Stretching – handled by mx
- Movement – Up and down (is handled by b which is intercept)

Advantages :

Resistant to Overfitting

Extended to multiclass problem