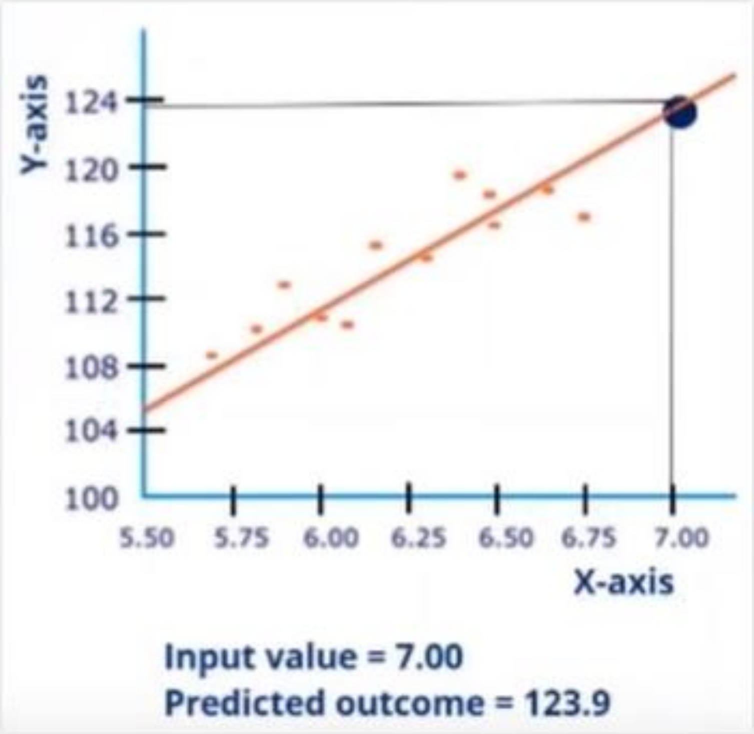
Regression

SANTHOSH KUMAR K P

What is Regression?

Regression Analysis is a **predictive modelling** technique

It estimates the relationship between a **dependent** (target) and an **independent** variable(predictor)



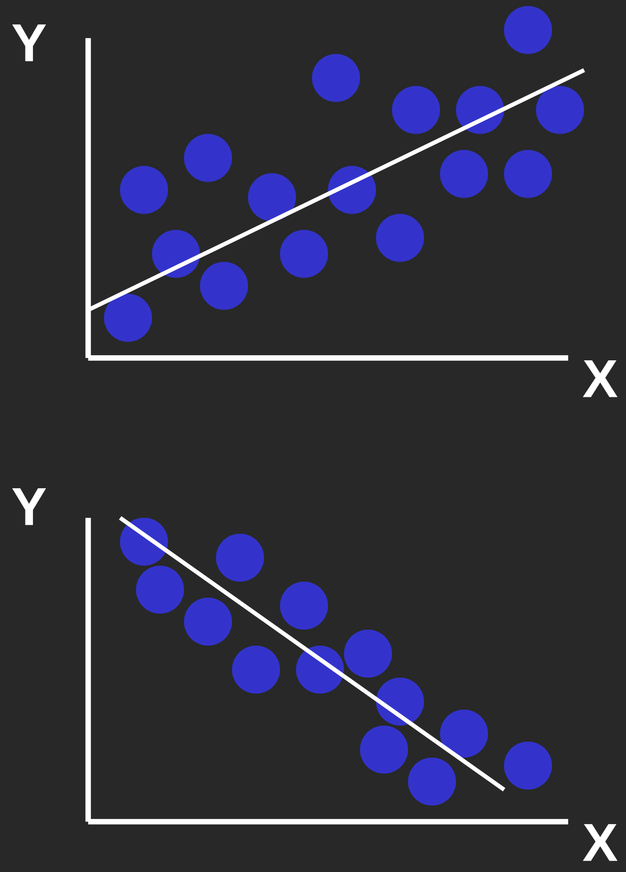
Uses of Regression

Three major uses for regression analysis are

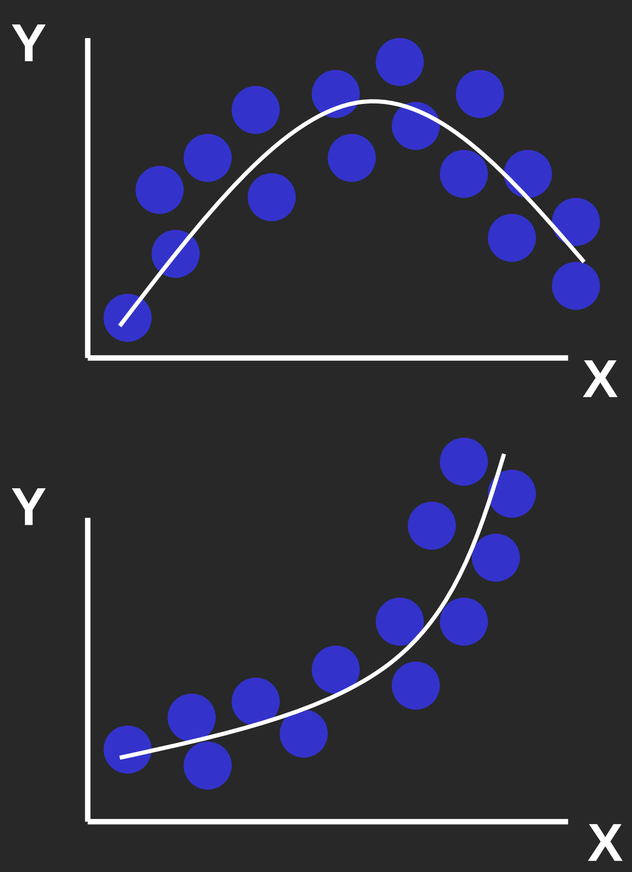
- Determining the strength of predictors
- Forecasting an effect, and
- Trend forecasting

Regression Models

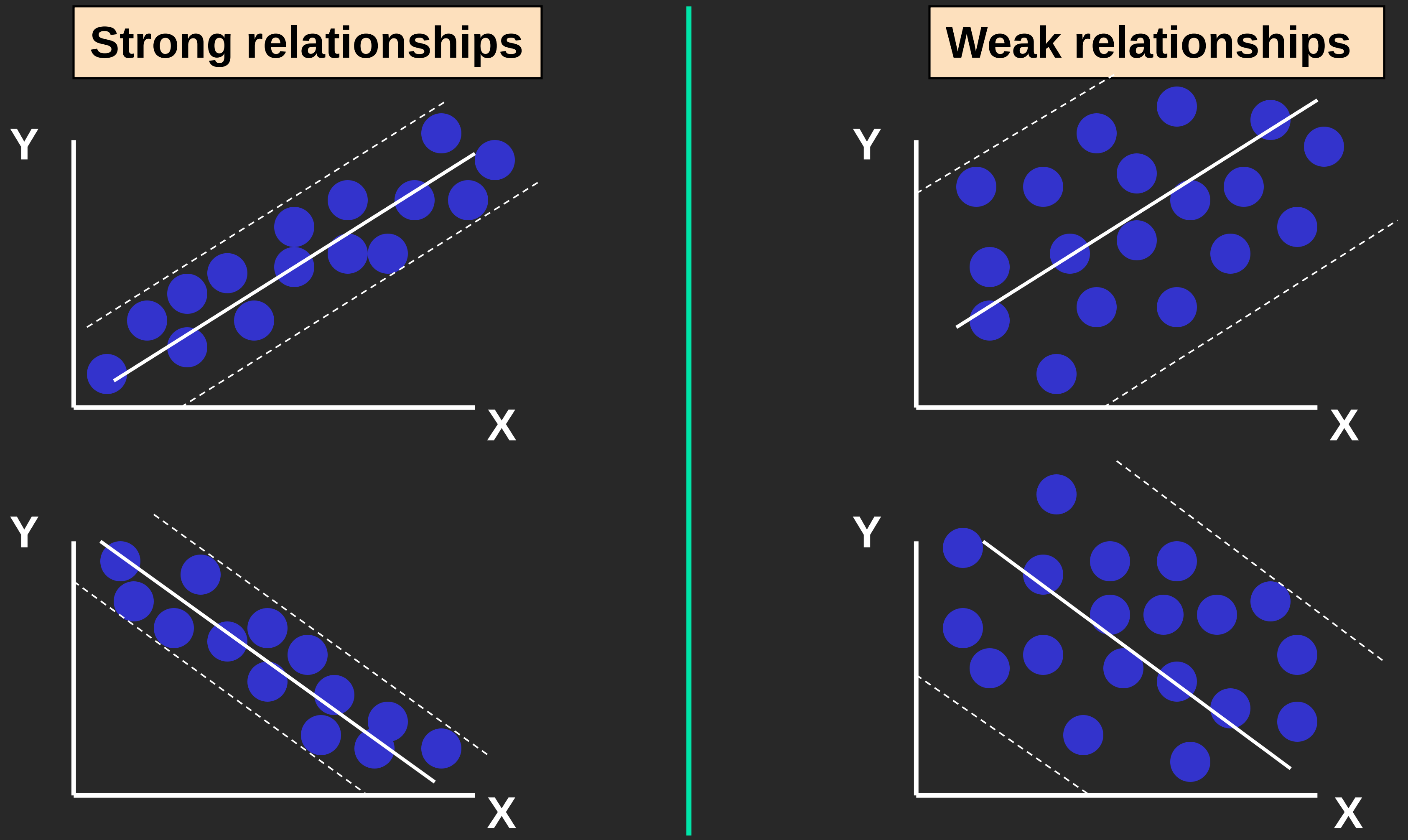
Linear relationships



Curvilinear relationships

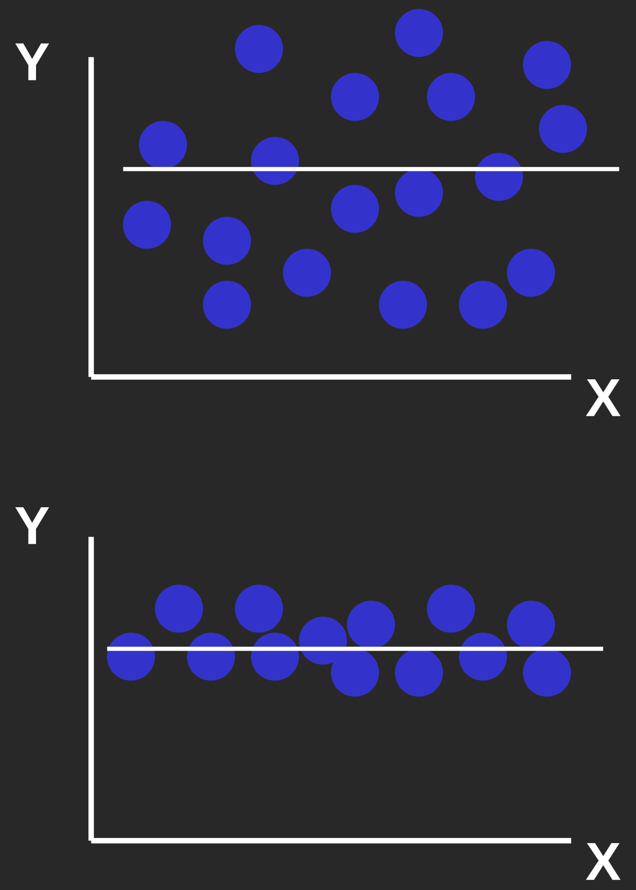


Types of Relationships

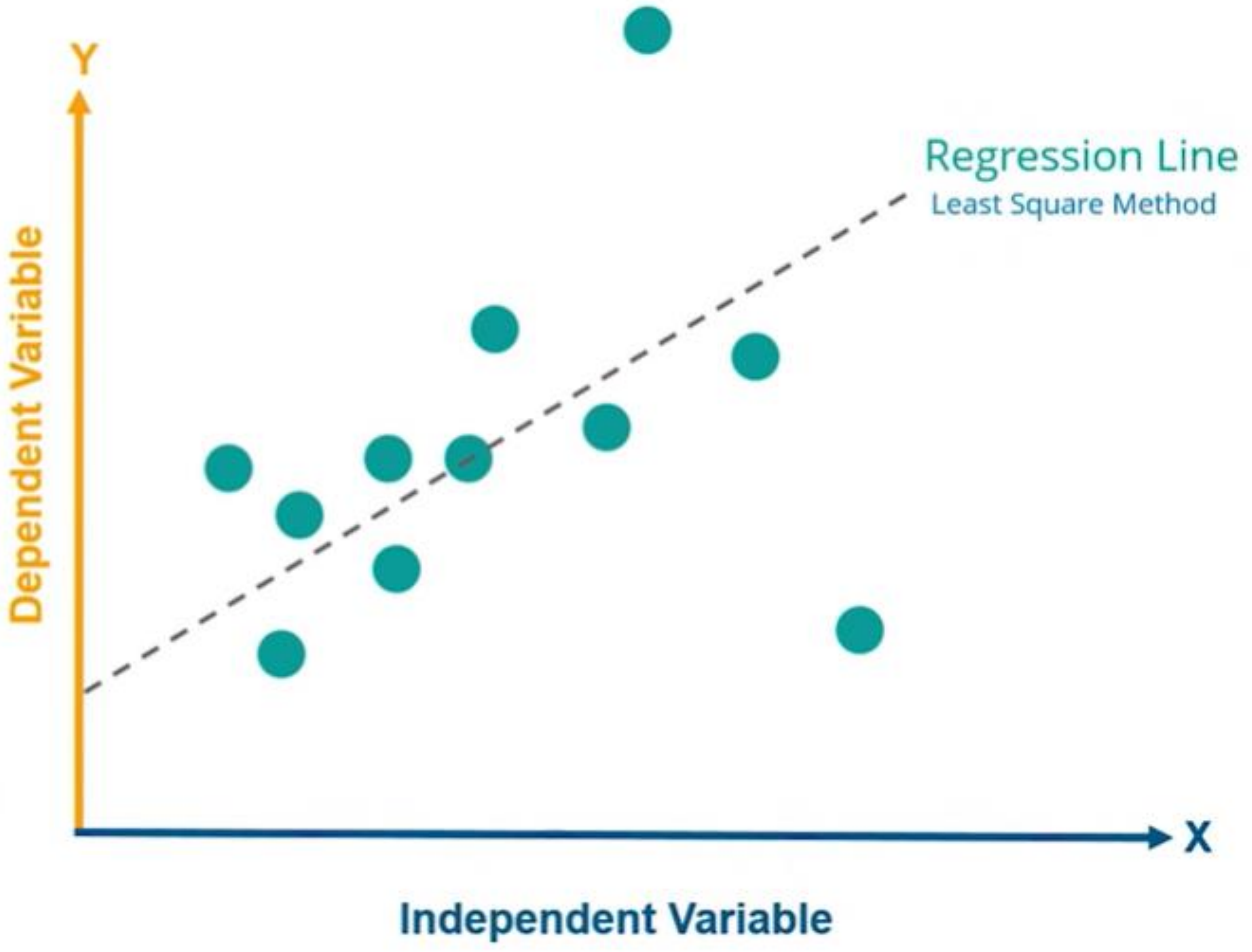


Types of Relationships

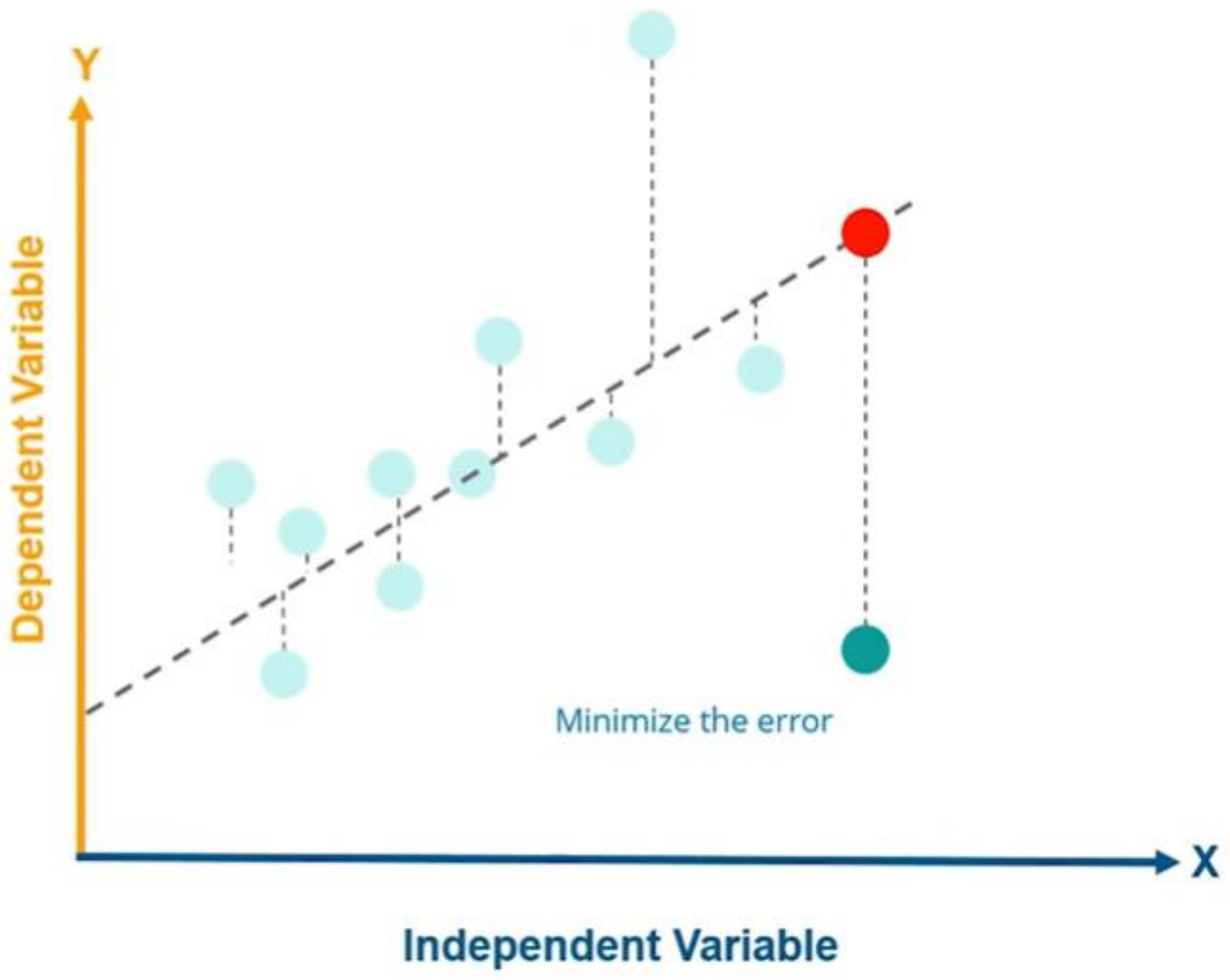
No relationship



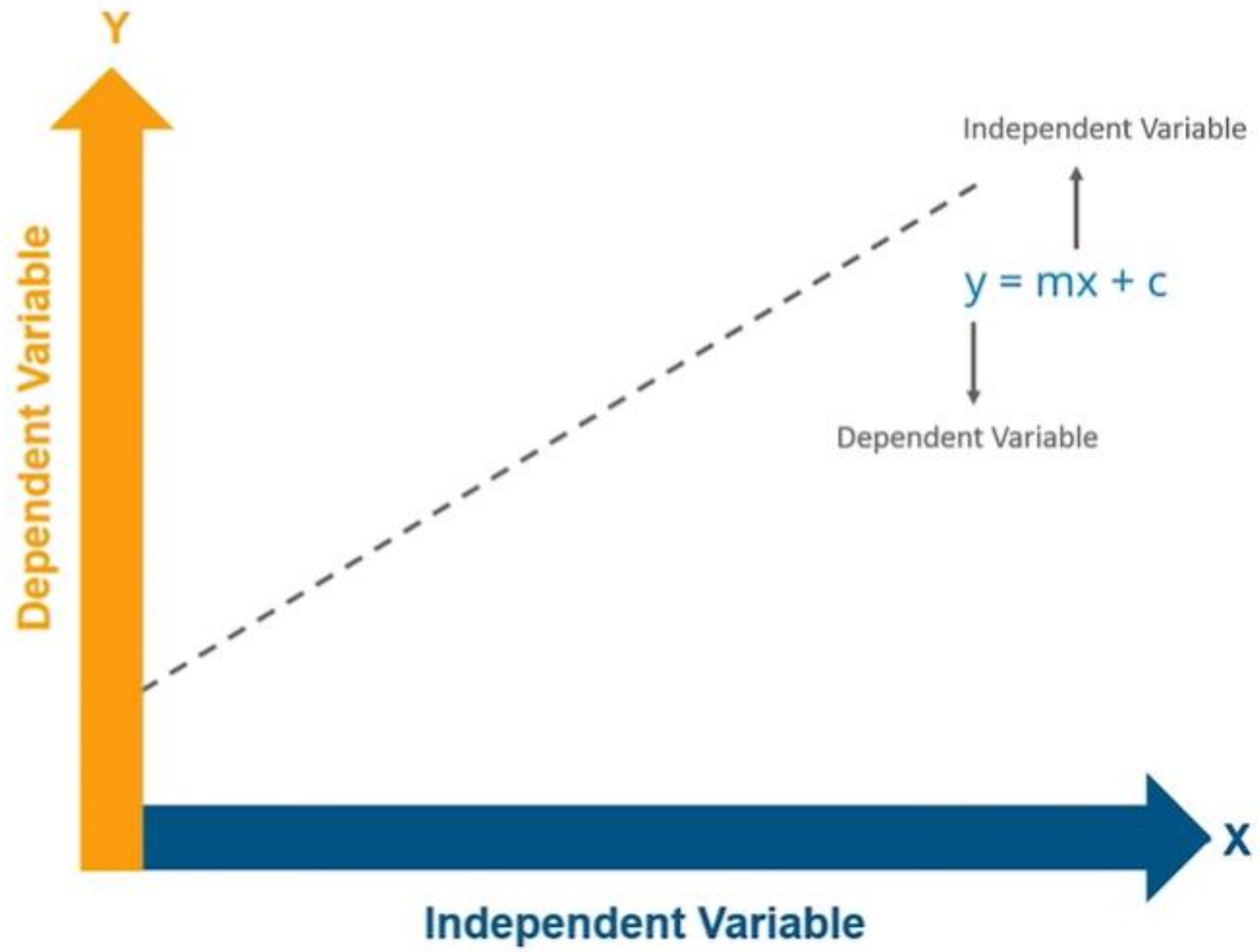
Understanding Linear Regression Algorithm



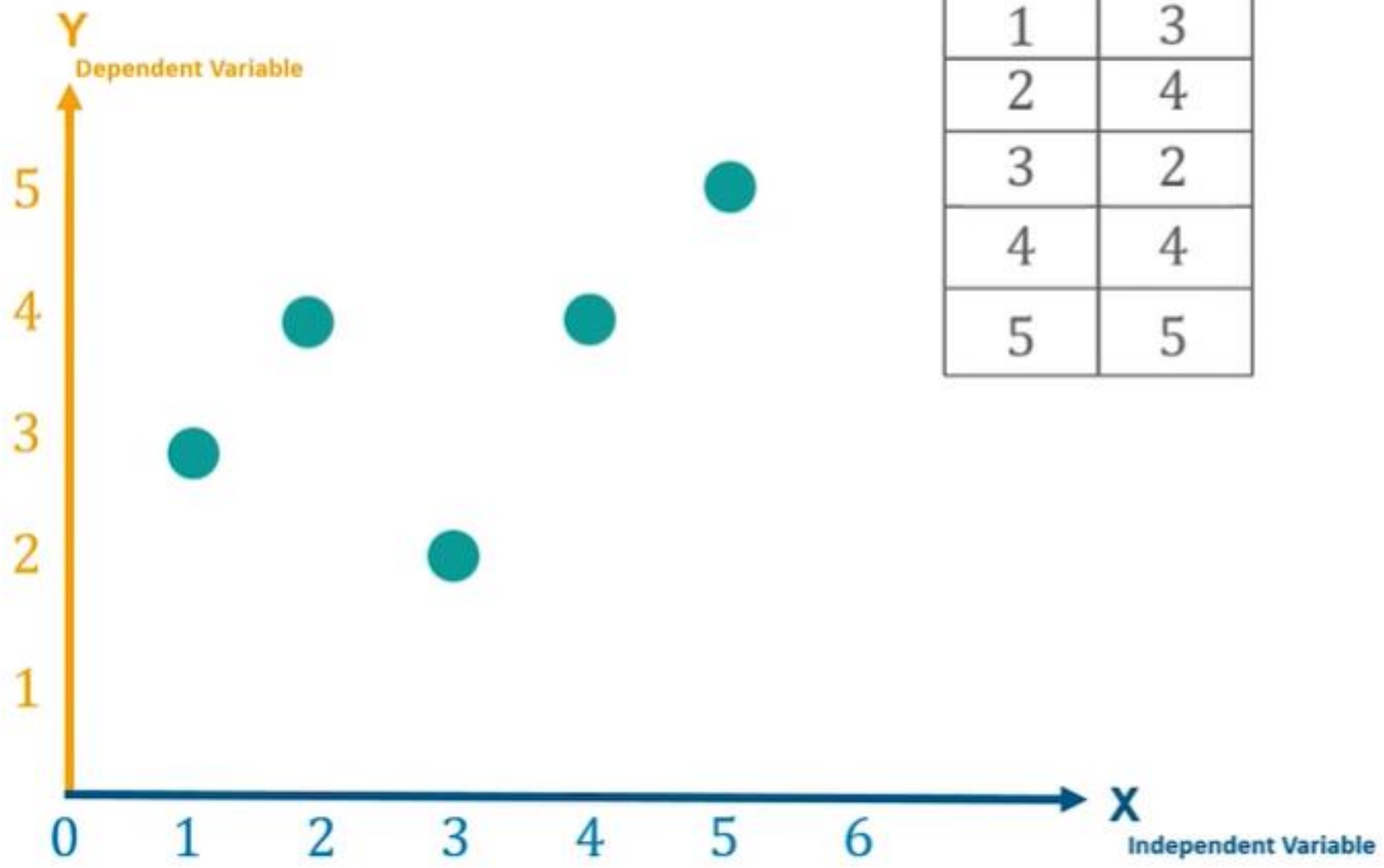
Understanding Linear Regression Algorithm



Understanding Linear Regression Algorithm

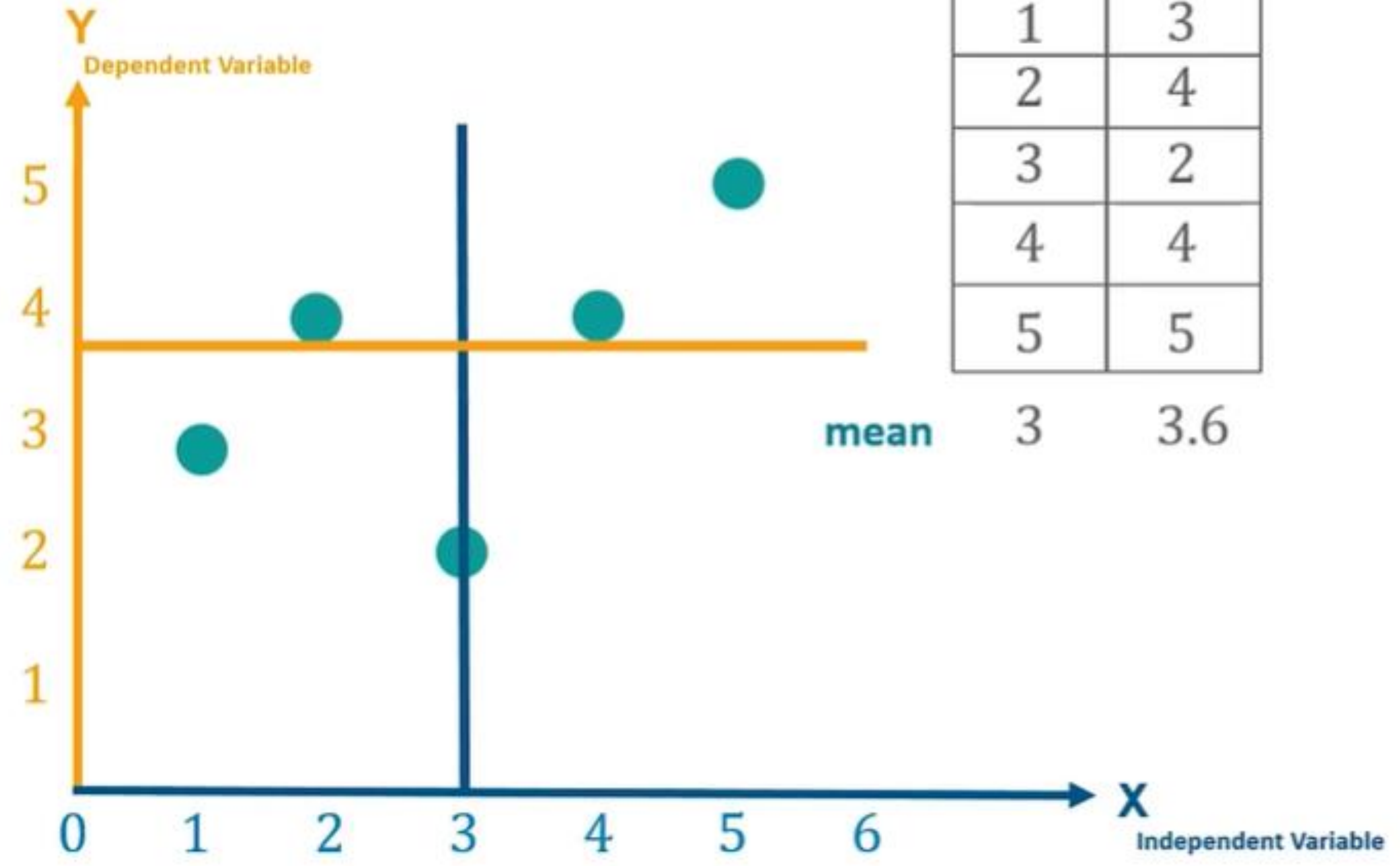


Understanding Linear Regression Algorithm



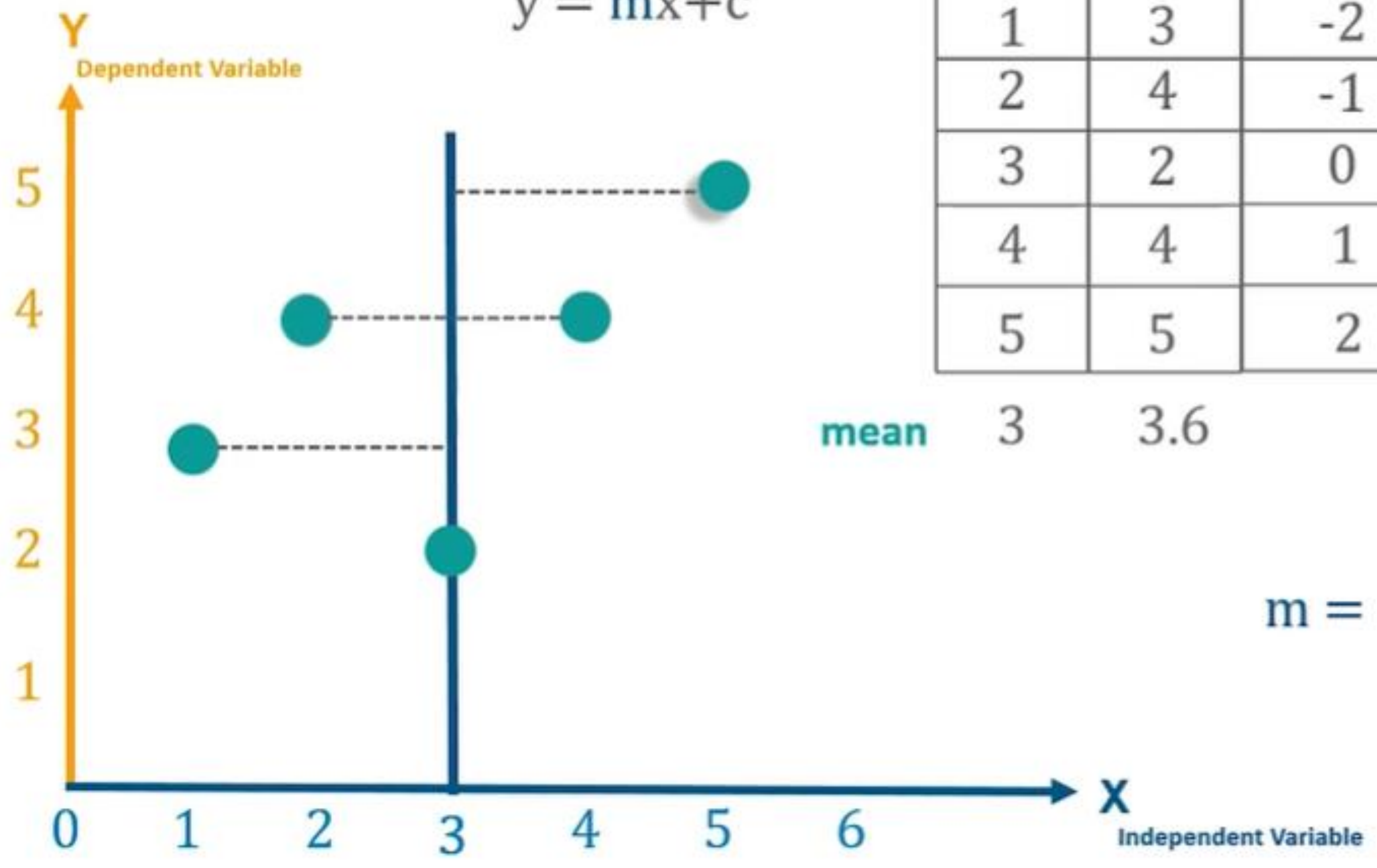
x	y
1	3
2	4
3	2
4	4
5	5

Understanding Linear Regression Algorithm



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

Understanding Linear Regression Algorithm



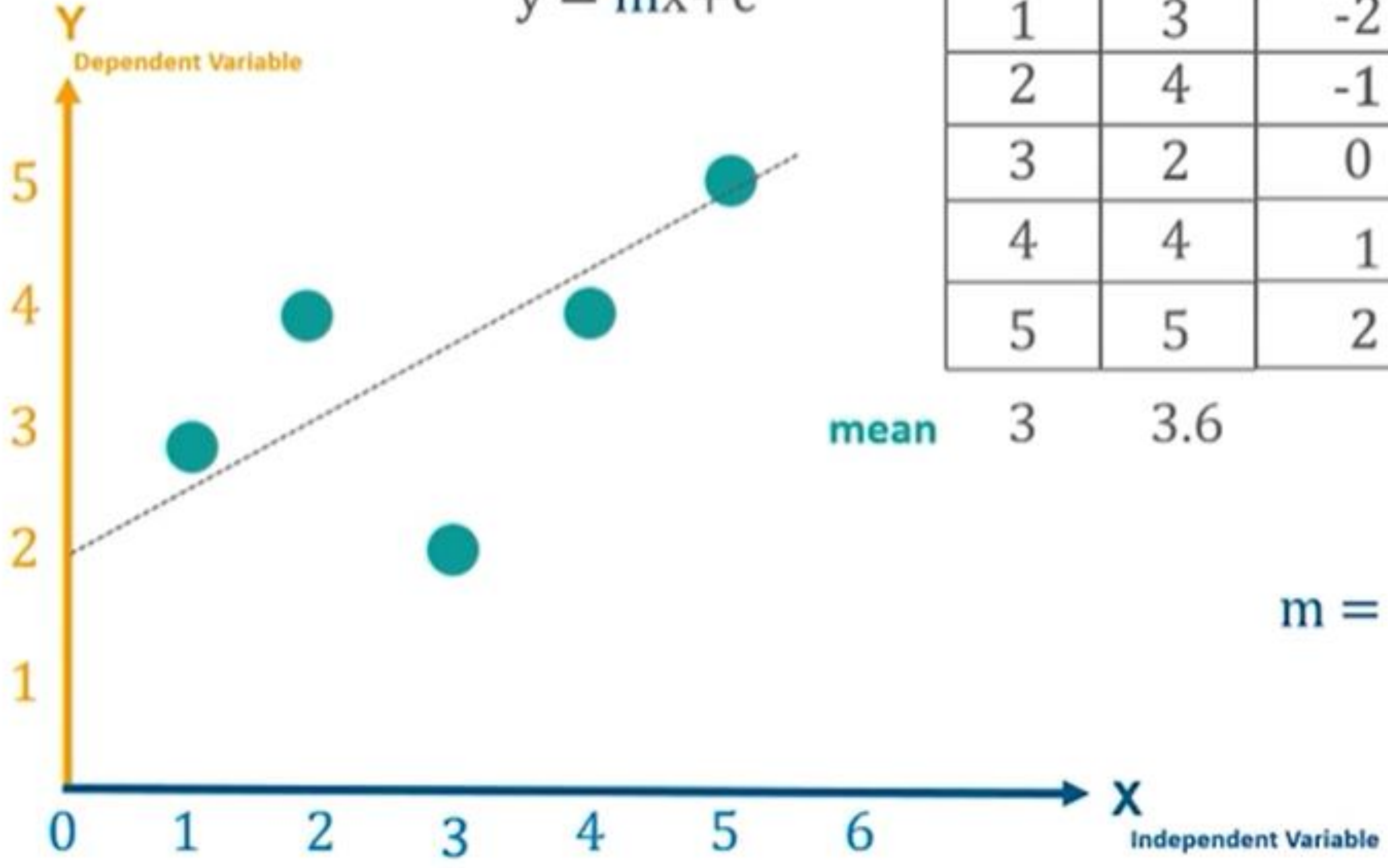
$y = mx + c$

x	y	$x - \bar{x}$
1	3	-2
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

$y = mx + c$



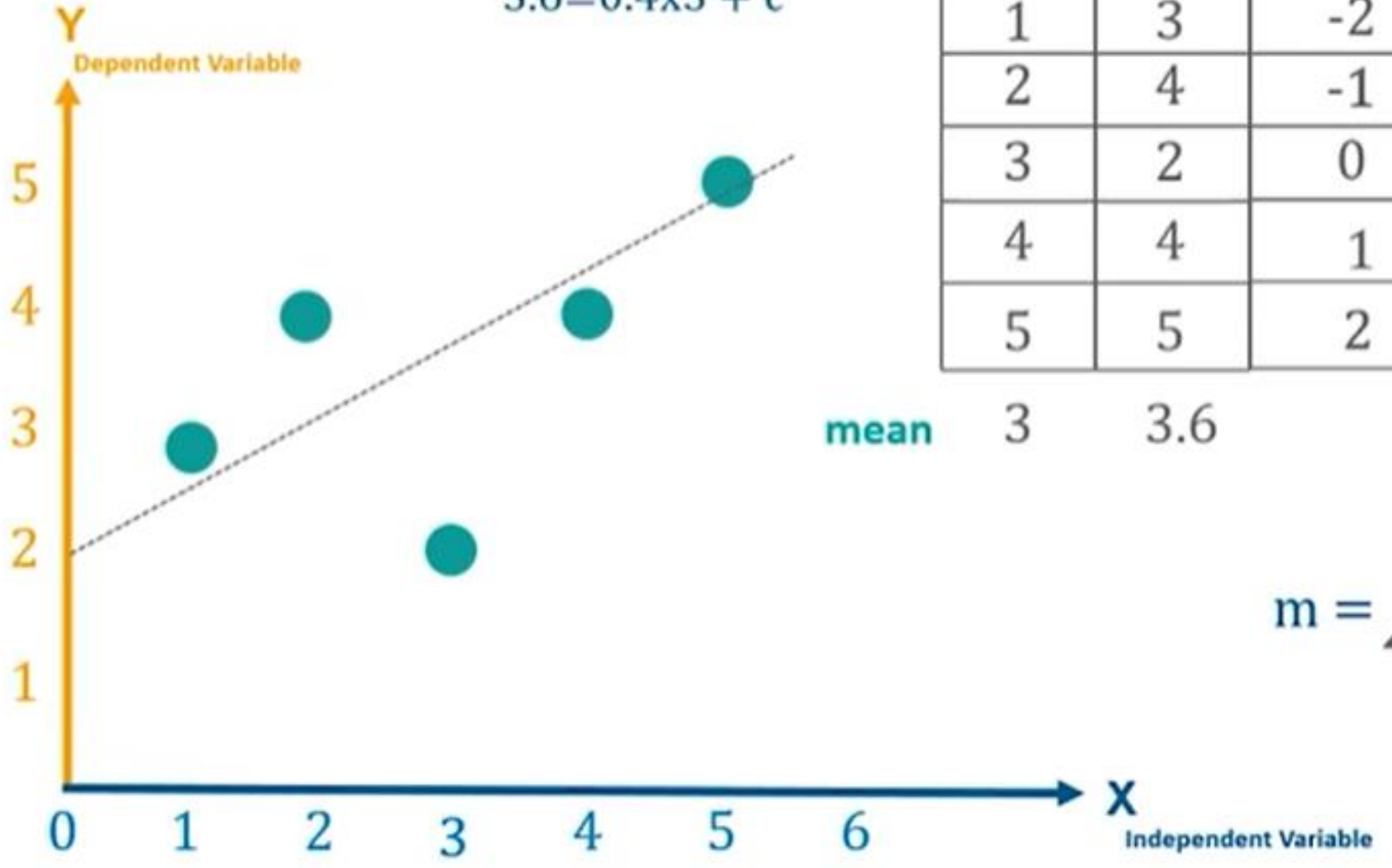
	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	1	3	-2	-0.6	4	1.2
	2	4	-1	0.4	1	-0.4
	3	2	0	-1.6	0	0
	4	4	1	0.4	1	0.4
	5	5	2	1.4	4	2.8
mean	3	3.6			$\Sigma = 10$	$\Sigma = 4$

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

Understanding Linear Regression Algorithm

$$y = mx+c$$

$$3.6=0.4x3 + c$$



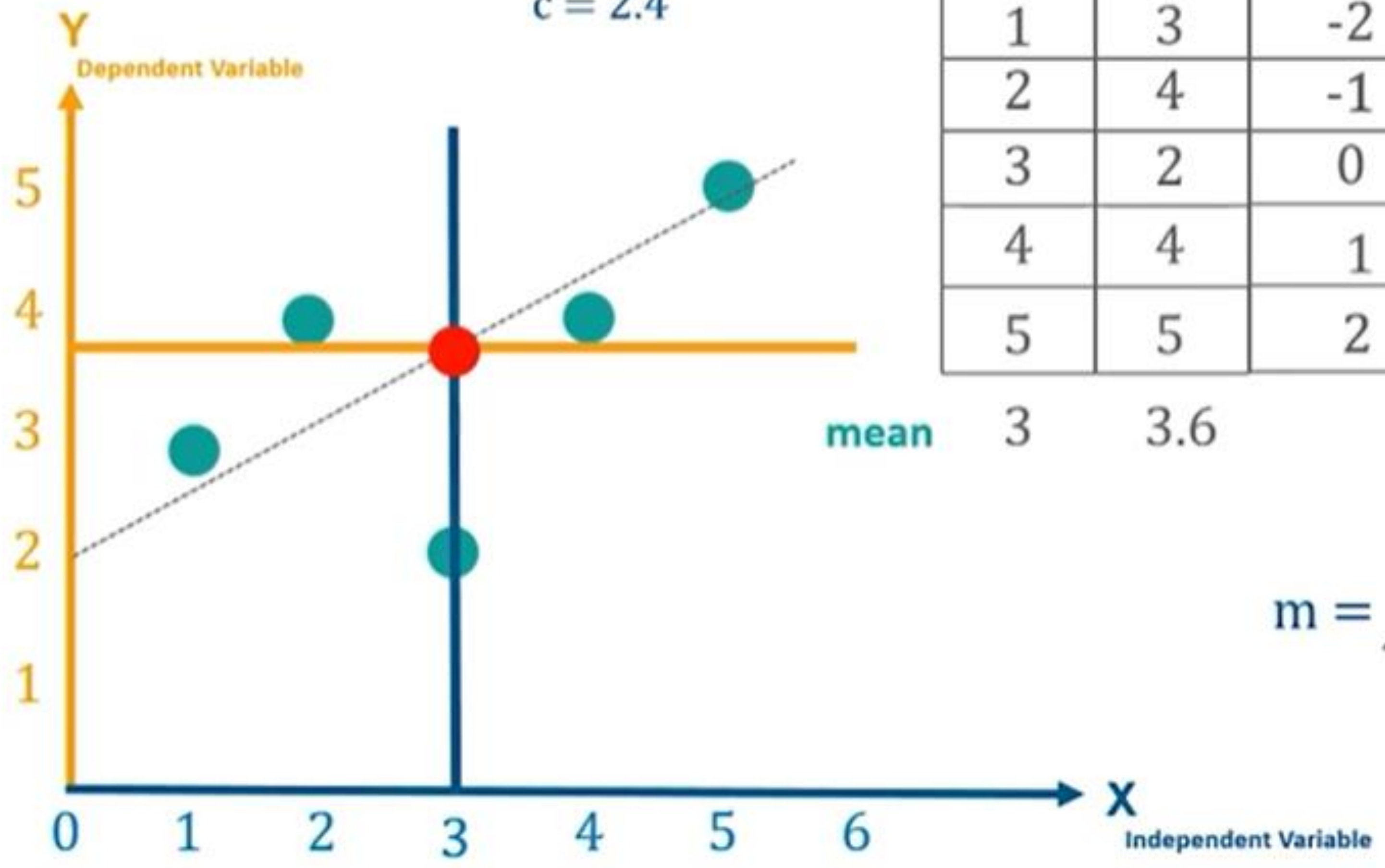
	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
	1	3	-2	-0.6	4	1.2
	2	4	-1	0.4	1	-0.4
	3	2	0	-1.6	0	0
	4	4	1	0.4	1	0.4
	5	5	2	1.4	4	2.8
mean	3	3.6			$\Sigma = 10$	$\Sigma = 4$

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

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			$\Sigma = 10$	$\Sigma = 4$

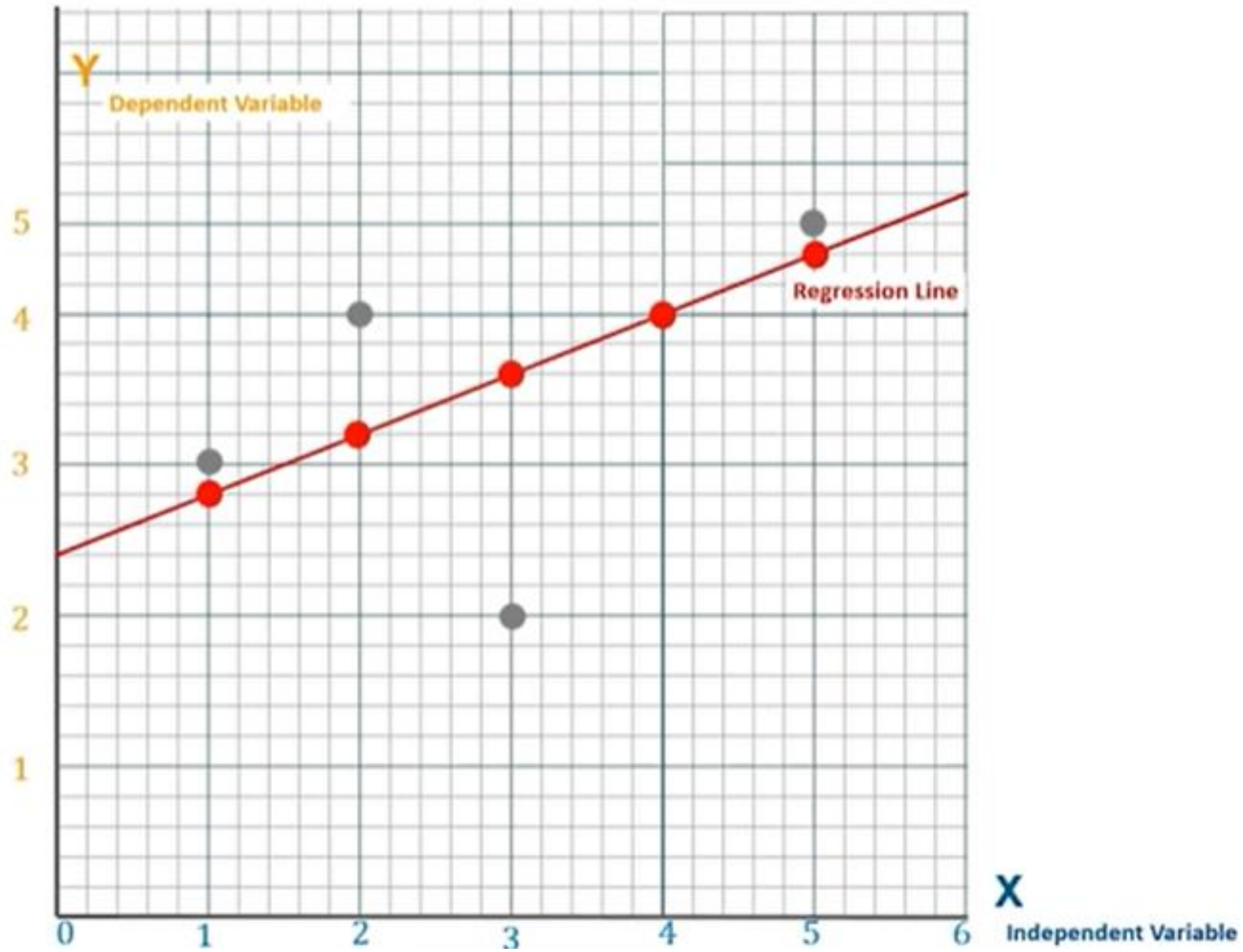
$$m = \sum \frac{(x - \bar{x})(y - \bar{y})}{(x - \bar{x})^2} = \frac{4}{10}$$

$m = 0.4$

$c = 2.4$

$y = 0.4x + 2.4$

Mean Square Error



$m = 0.4$

$c = 2.4$

$y = 0.4x + 2.4$

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

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

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

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

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

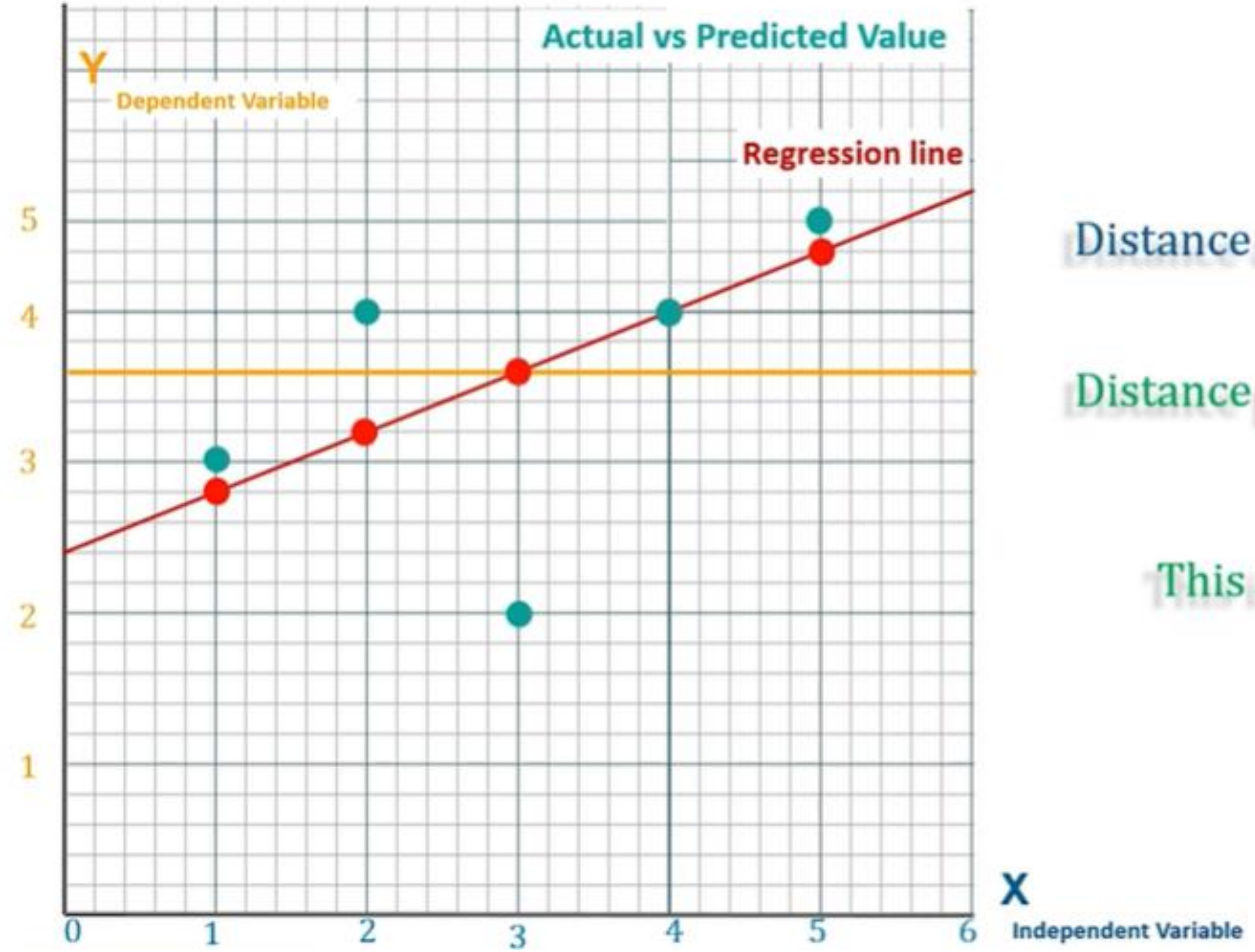
$y = 0.4 \times 5 + 2.4 = 4.4$

Let’s check the Goodness of fit

What is R-Square?

- **R-squared** value is a statistical measure of how close the data are to the fitted regression line
- It is also known as **coefficient of determination**, or the **coefficient of multiple determination**

Calculation of R^2



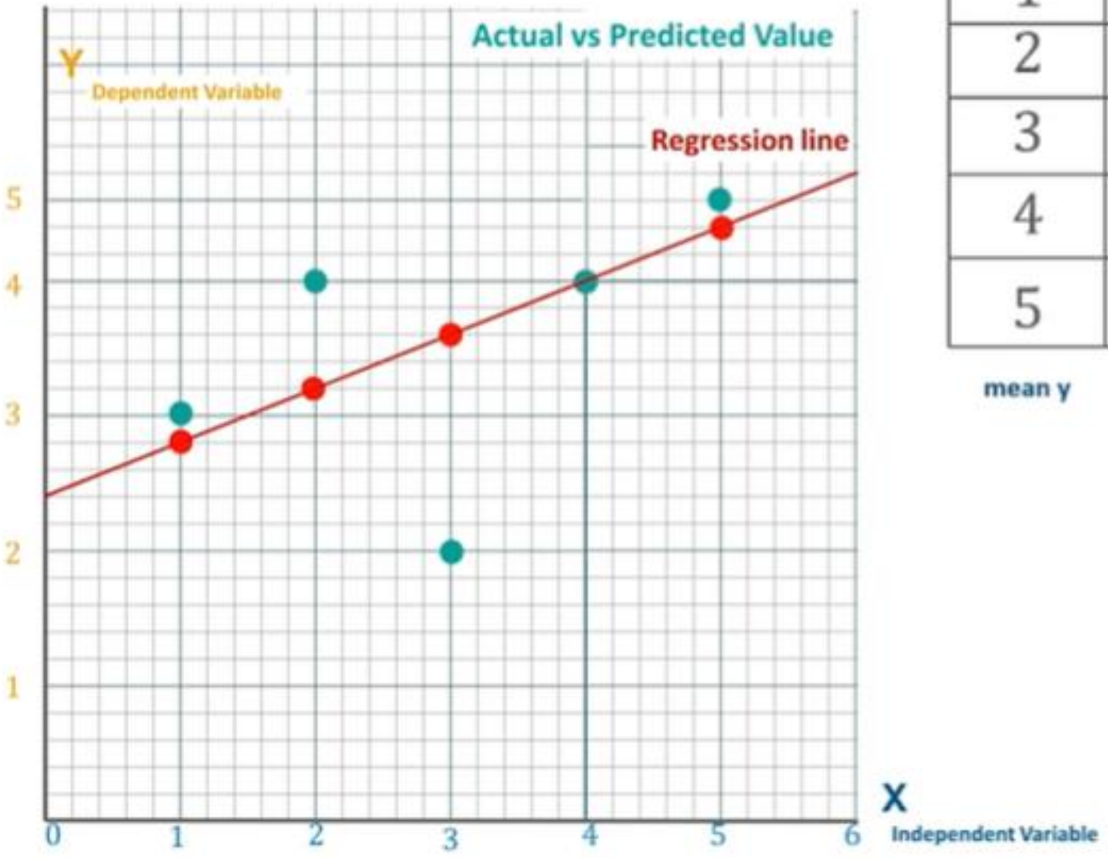
Distance actual - mean

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

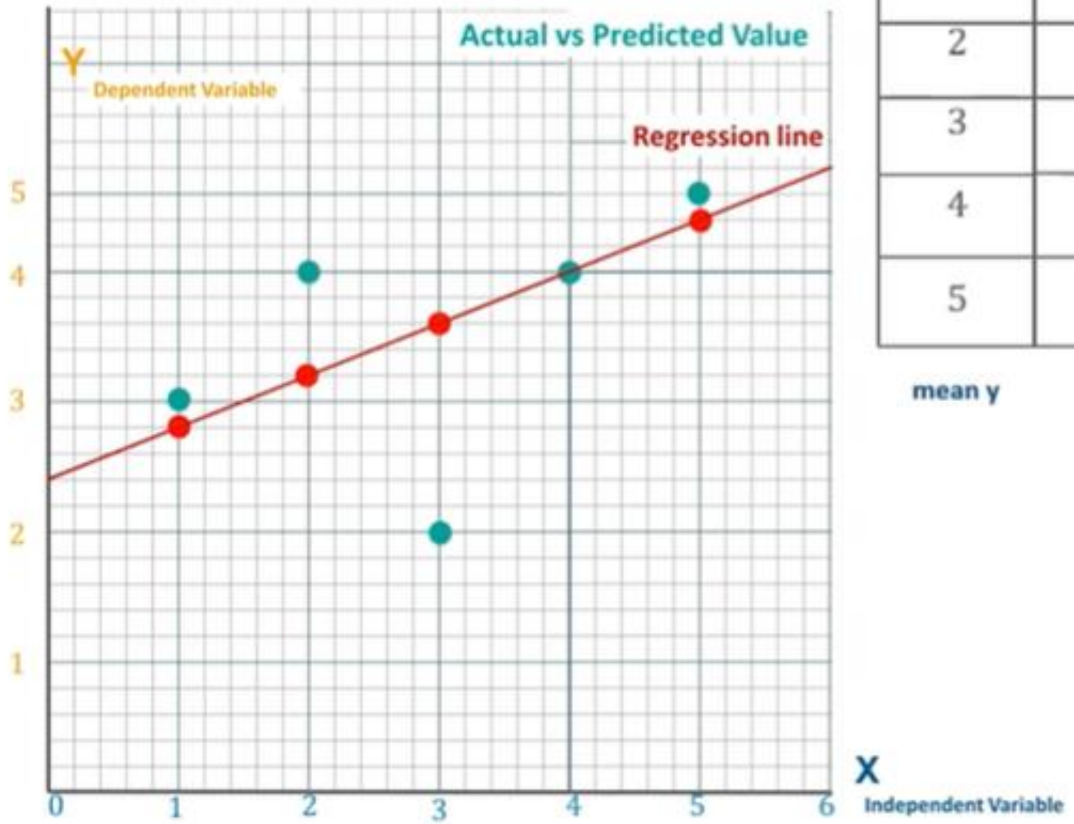


x	y	$y - \bar{y}$
1	3	– 0.6
2	4	0.4
3	2	–1.6
4	4	0.4
5	5	1.4

mean y 3.6

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

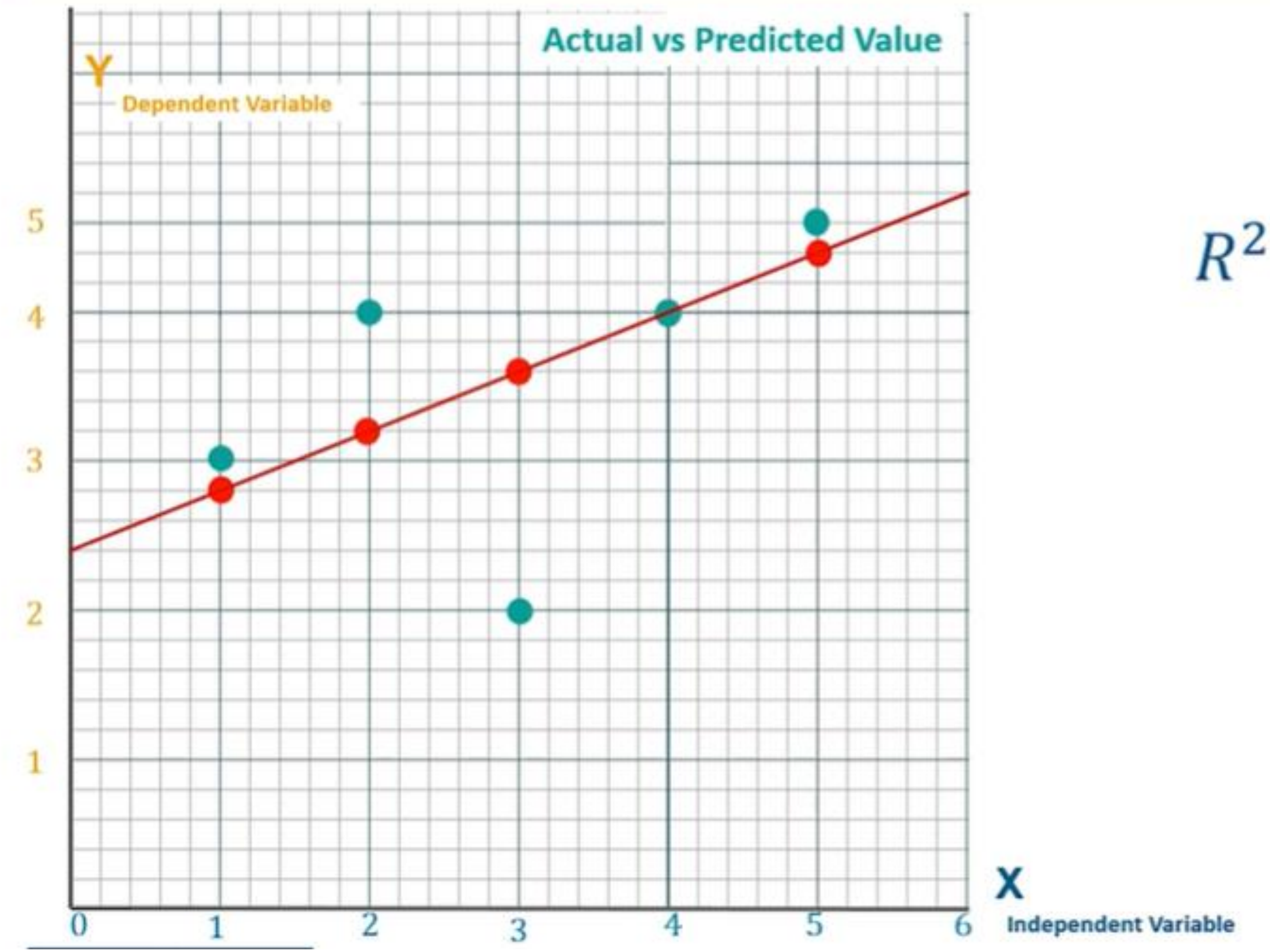
Calculation of R^2



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		Σ 5.2			Σ 1.6

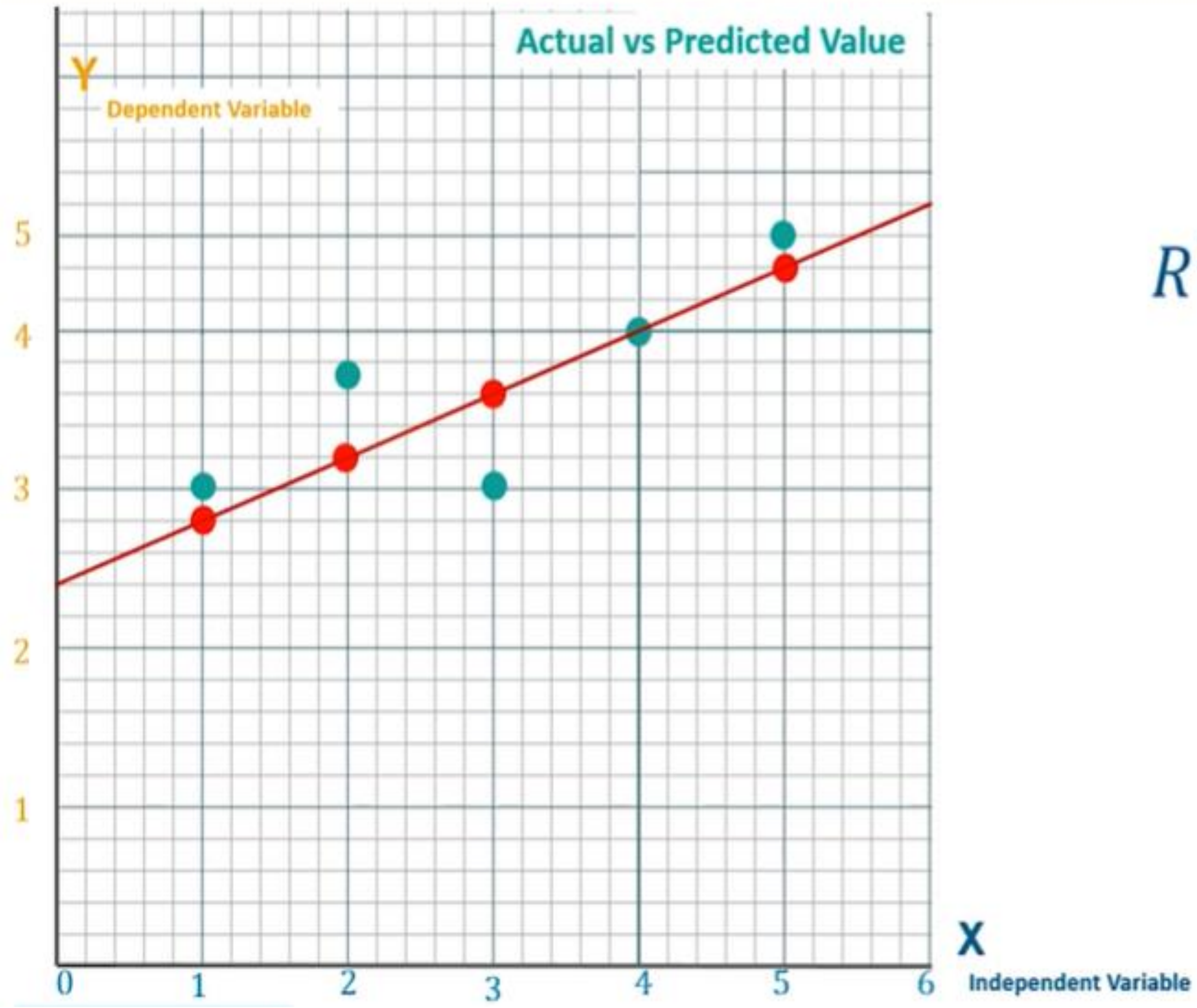
$$R^2 = \frac{1.6}{5.2} = \frac{\Sigma\left((y_p - \bar{y}\right)^2}{\Sigma (y - \bar{y})^2}$$

Calculation of R^2

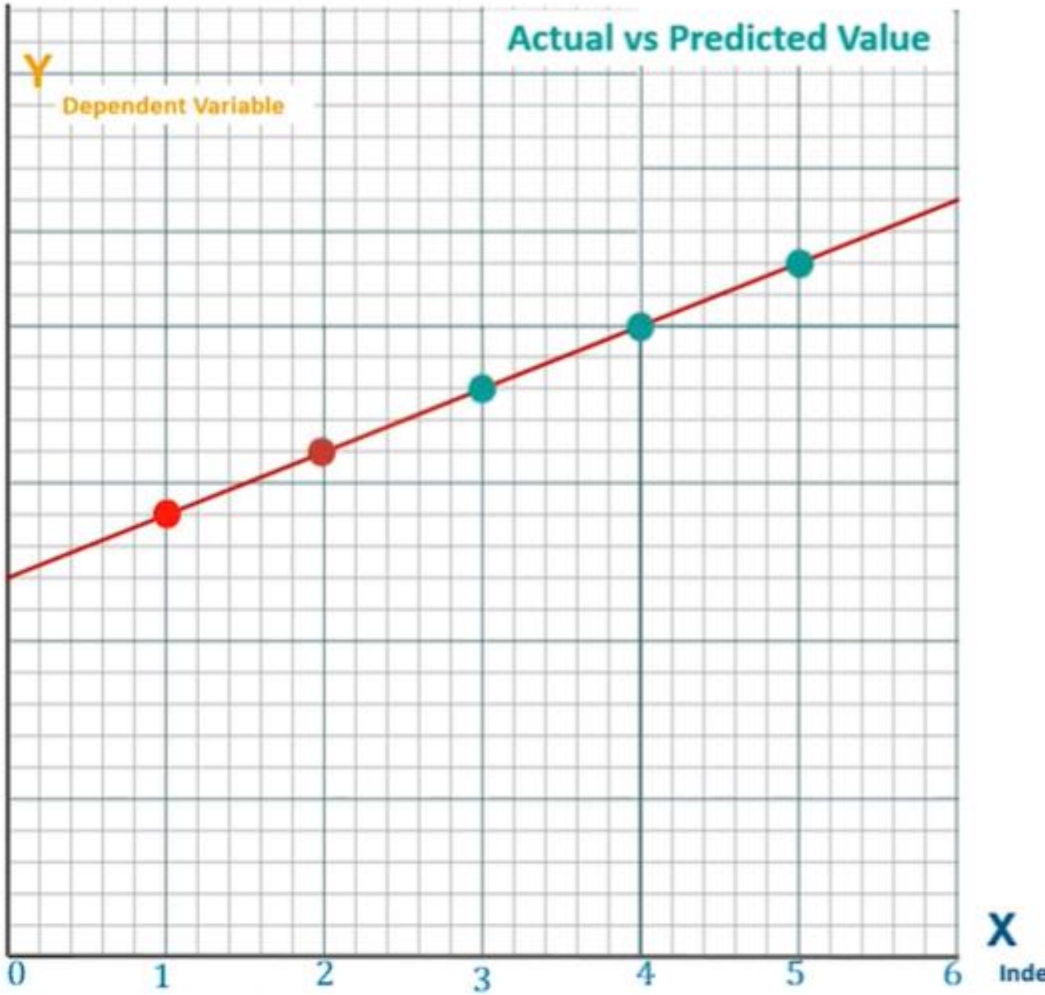


$$R^2 \approx 0.3$$

Calculation of R^2

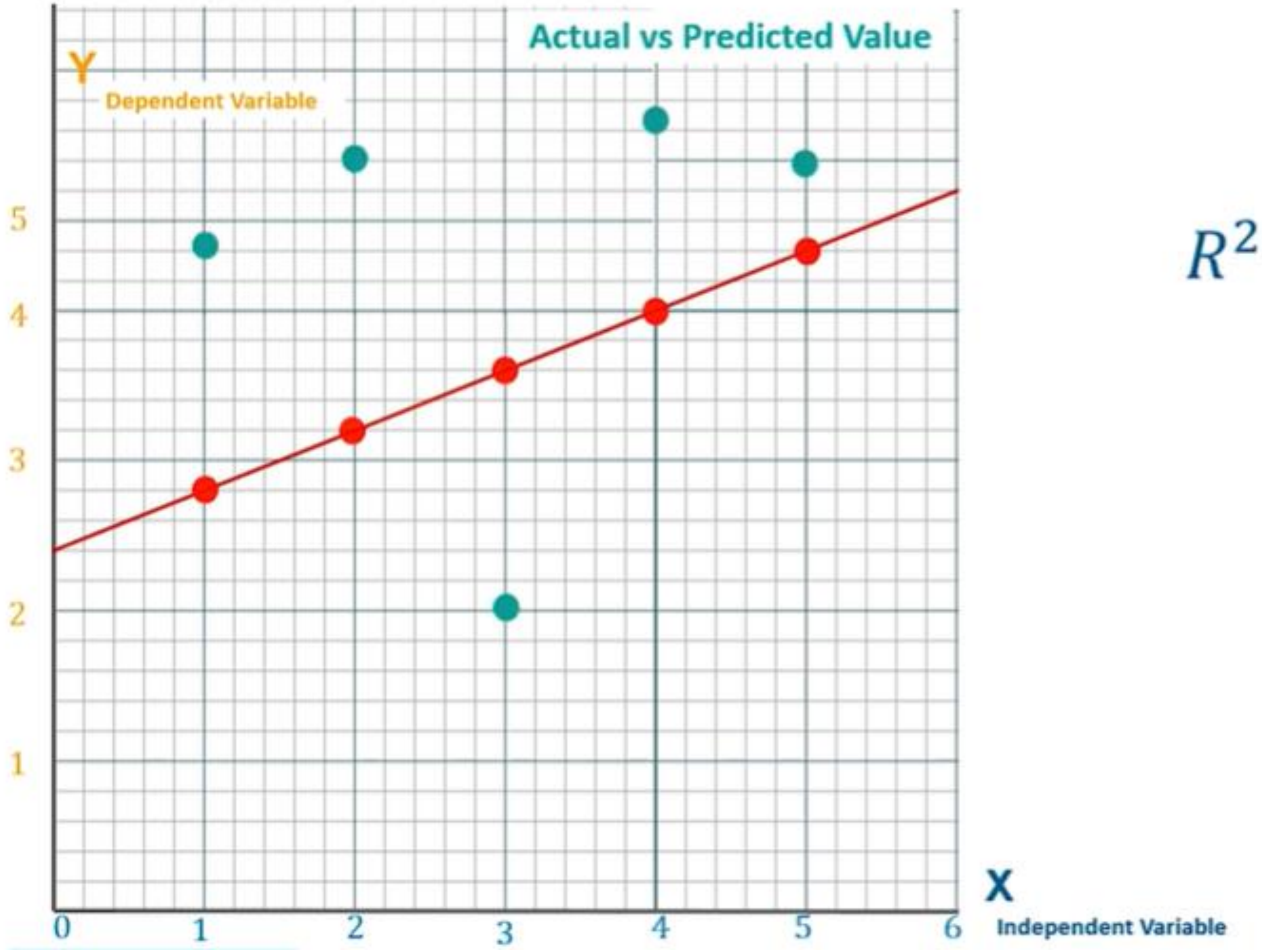


$R^2 \approx 0.7$



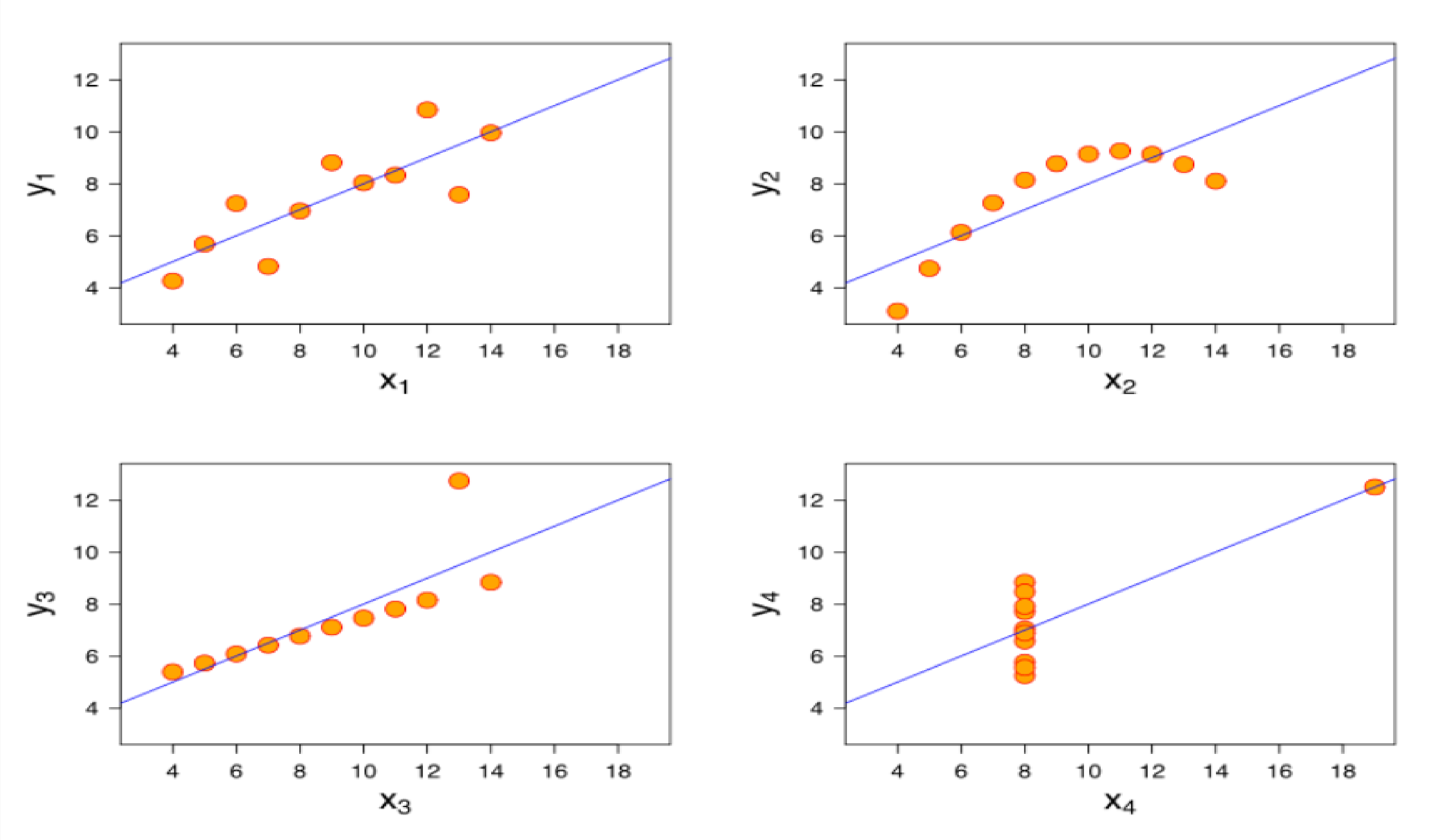
$R^2 \approx 1$

Calculation of R^2



$$R^2 \approx 0.02$$

EXAMPLES

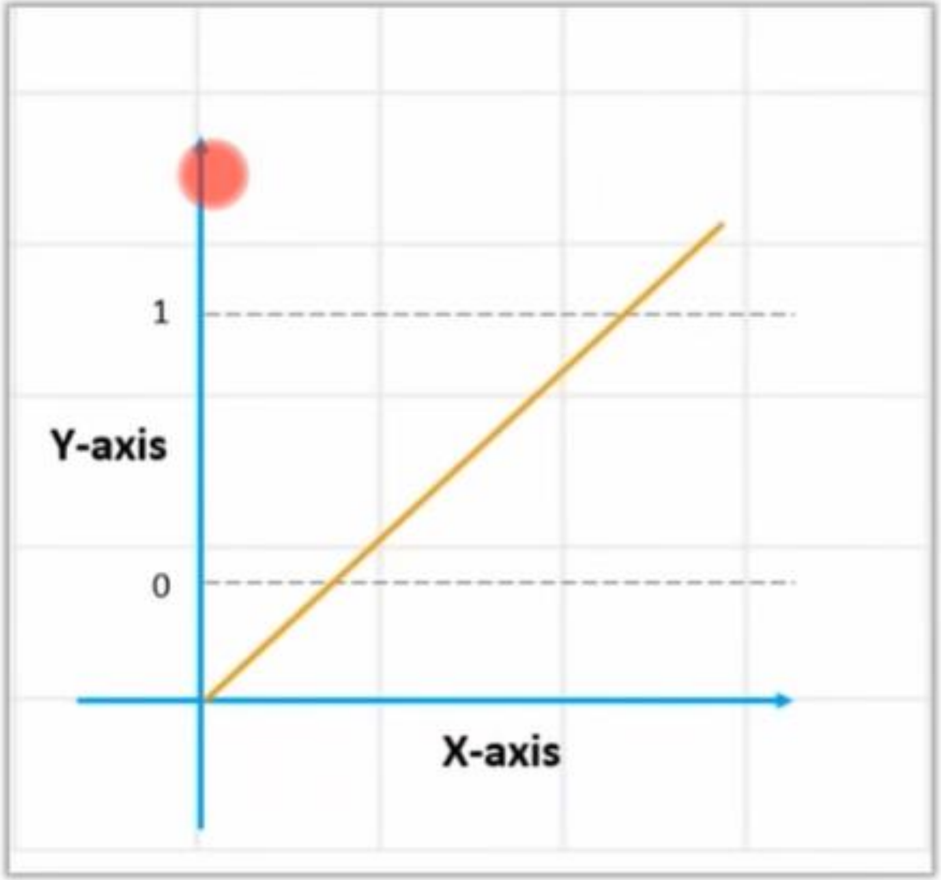


What lines "really" best fit each case? – different approaches

Logistic Regression: What And Why?

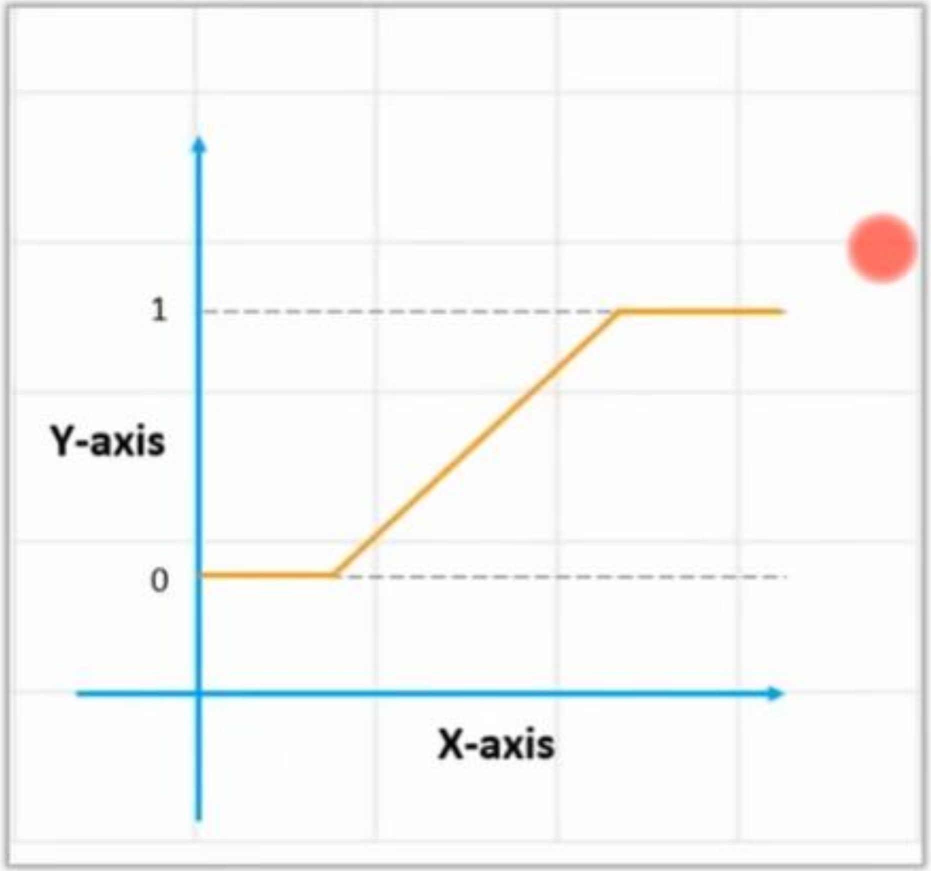
Logistic Regression produces results in a binary format which is used to predict the outcome of a categorical dependent variable. So the outcome should be discrete/ categorical such as:

Why Not Linear Regression?



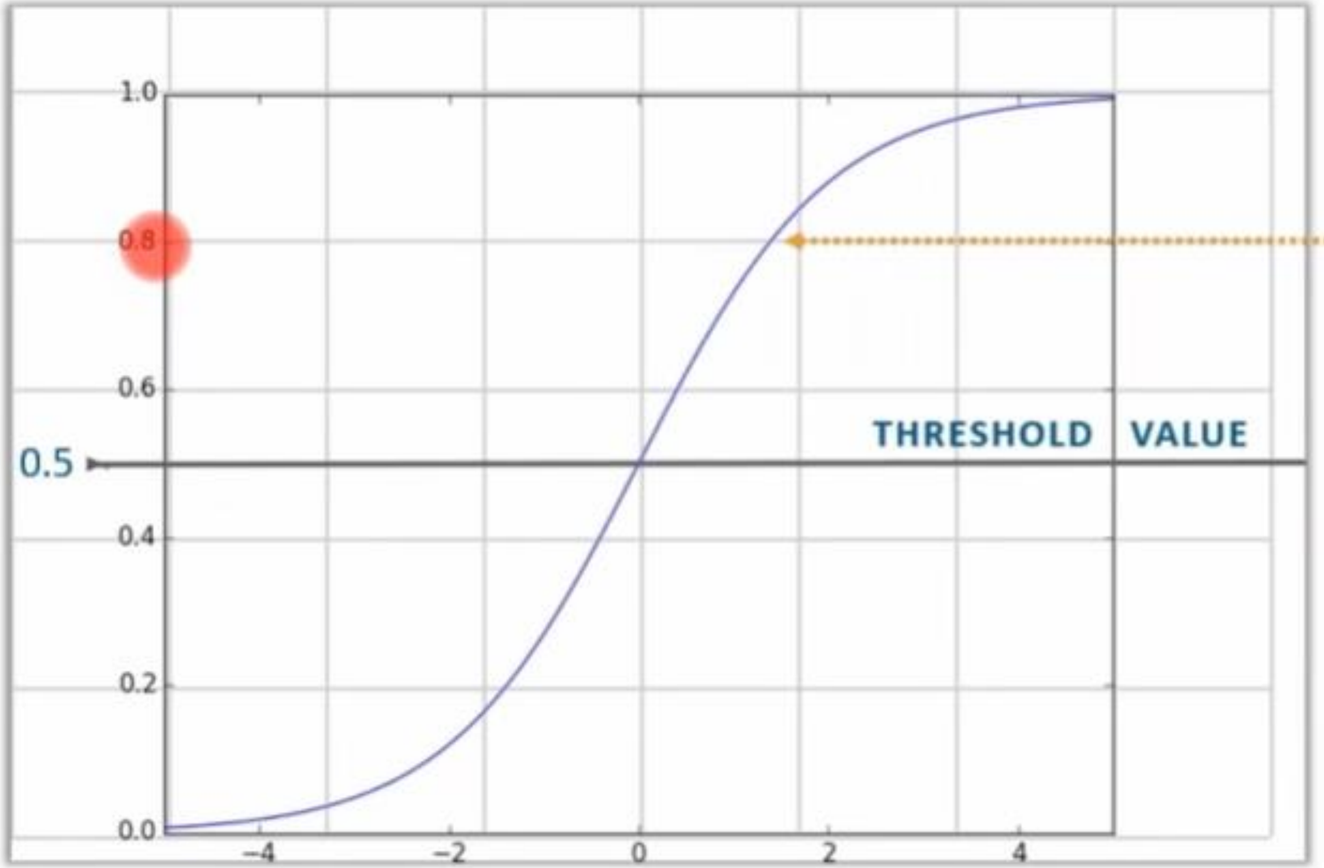
Since our value of Y will be between 0 and 1, the linear line has to be clipped at 0 and 1.

Why Not Linear Regression?



With this, our resulting curve cannot be formulated into a single formula. Hence we came up with **Logistic**!

Logistic Regression Curve



With this, the threshold value indicates the probability of winning or losing

Let us transform it further, to get range between –(infinity) and (infinity)

$\log \left[\frac{Y}{1-Y}\right] \Rightarrow Y = C + B1X1 + B2X2 + \ldots$ — Final Logistic Regression Equation

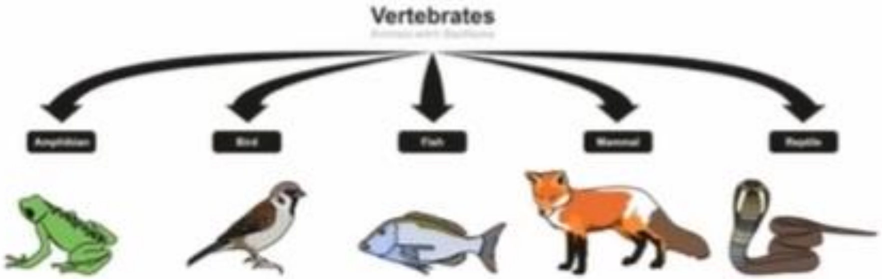
$$f(x) = \frac{1}{1 + e^{-x}}$$

Linear Vs Logistic Regression

Linear Regression	Logistic Regression
1 Continuous variables	1 Categorical variables
2 Solves Regression Problems	2 Solves Classification Problems
3 Straight line	3 S-Curve

Logistic Regression: Use - Cases

Determines Illness



Your best quote that reflects your approach… "It's one small step for man, one giant leap for mankind."

- NEIL ARMSTRONG