Module 3: Regression

What is Regression?

"Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable"

A regression analysis involves graphing a line over a set of data points that most closely fits the overall shape of the data.
A regression shows the changes in a dependent variable on the y-axis to the explanatory variable on the x-axis.

Uses of Regression

Three major uses for regression analysis are

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

3] Predicts trends and future values. Regression analysis can be used to get point estimates. Questions like what will be the price of bitcoin in the next 6 months.

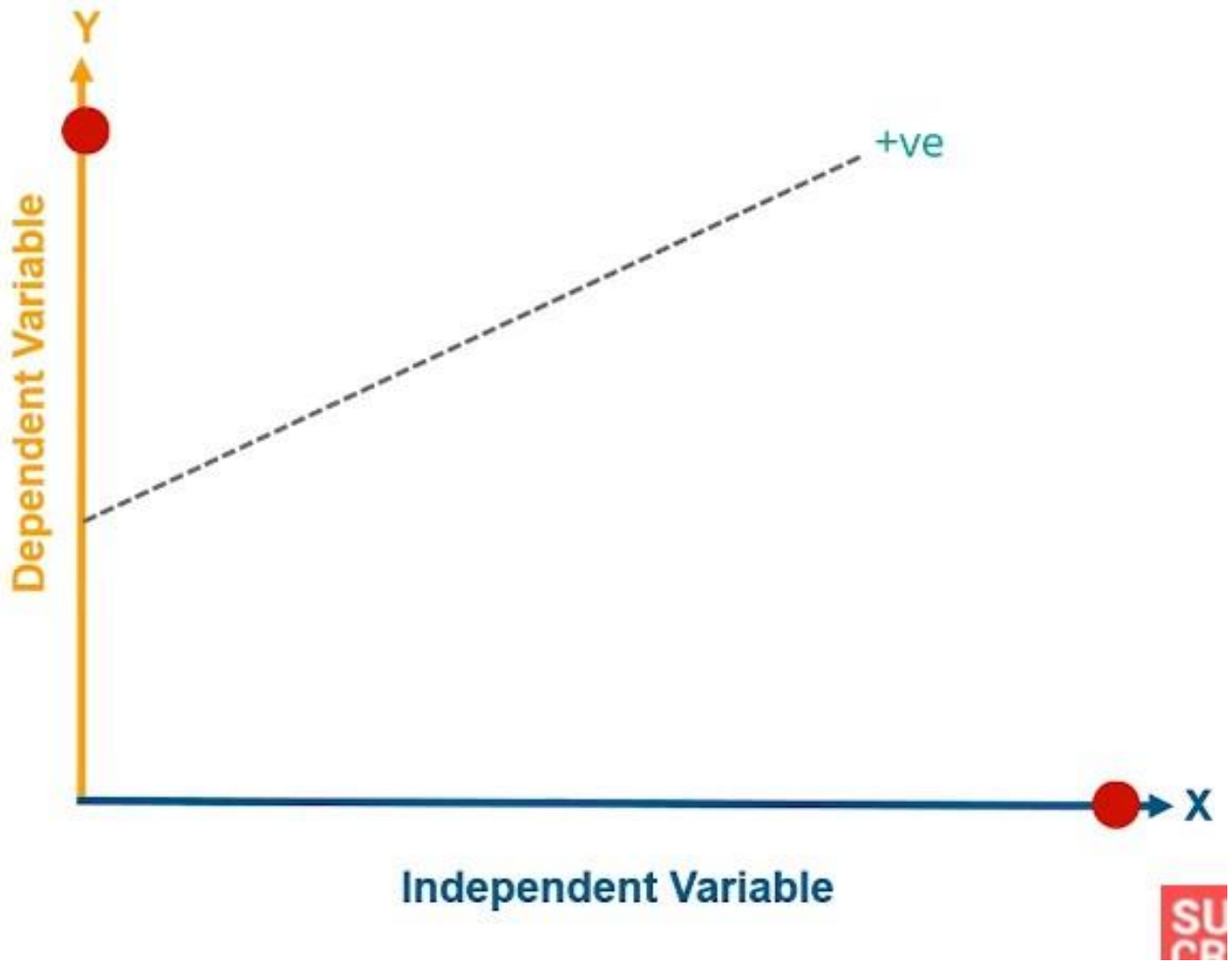
1] Regression might be used to identify the strength of the effect that the independent variables have on the dependent variables.
For e.g., what is the strength of the relationship between sales and marketing spending or what is the relationship between age and income.

2] Regression can be used to forecast effects or impact of changes.
In other words, we come to know how much the dependent variable changes with the change in one or more independent variables.
For e.g., how much additional sale income for each thousand dollars spent on marketing.

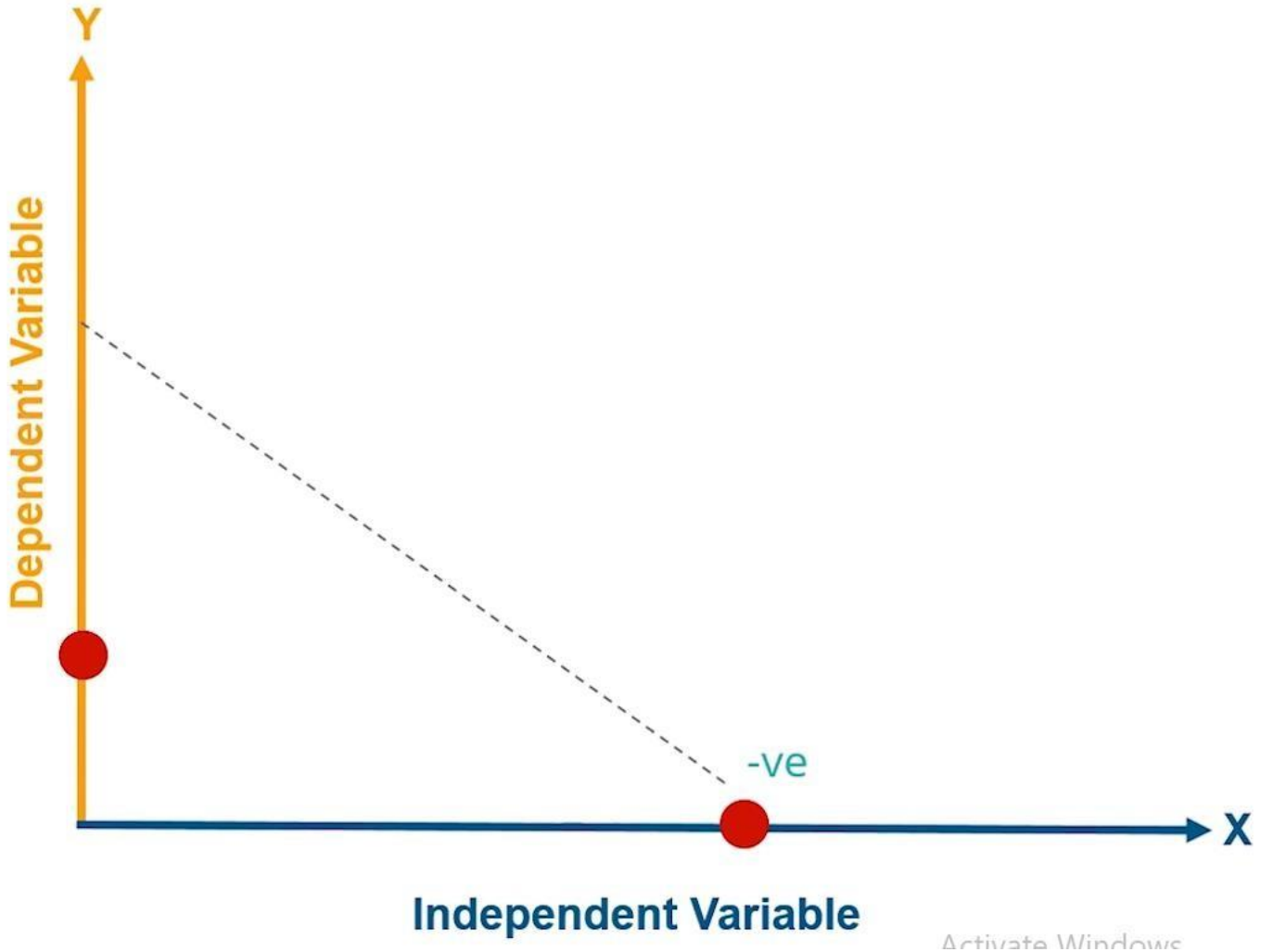
Where is Linear Regression used?

- Evaluating Trends and Sales Estimates
- Analyzing the Impact of Price Changes
- Assessment of risk in financial services and insurance domain

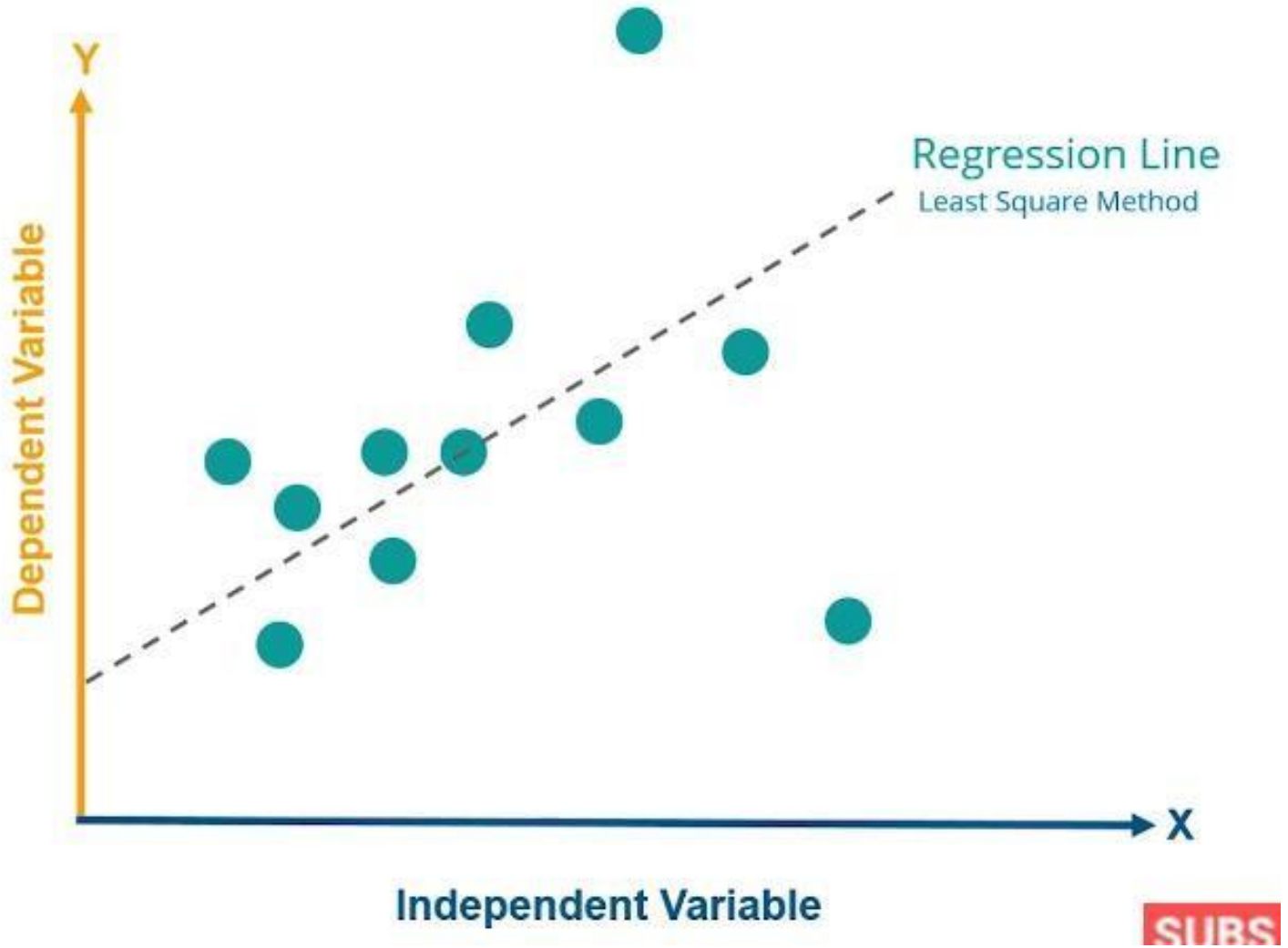
Understanding Linear Regression Algorithm



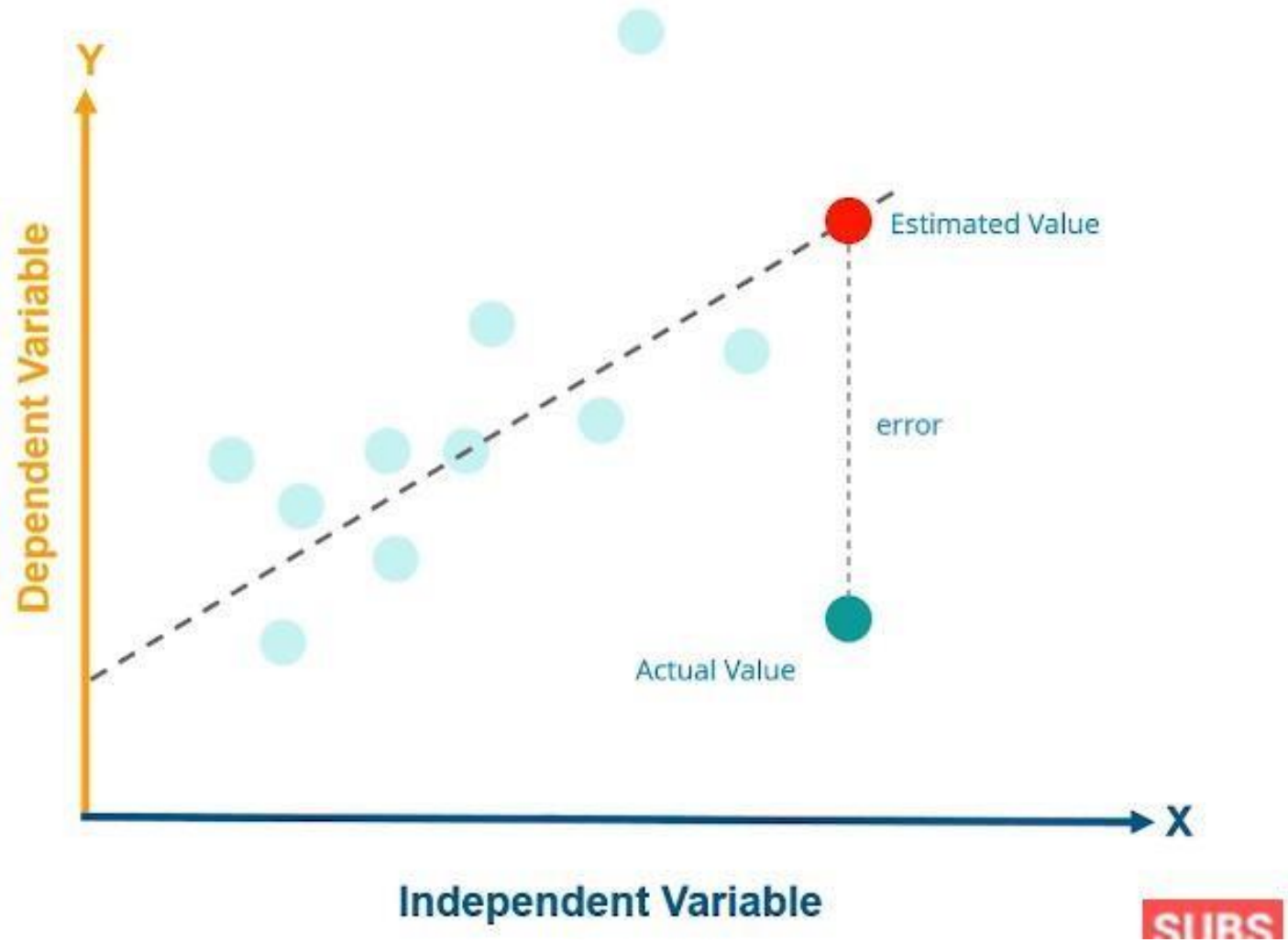
Understanding Linear Regression Algorithm



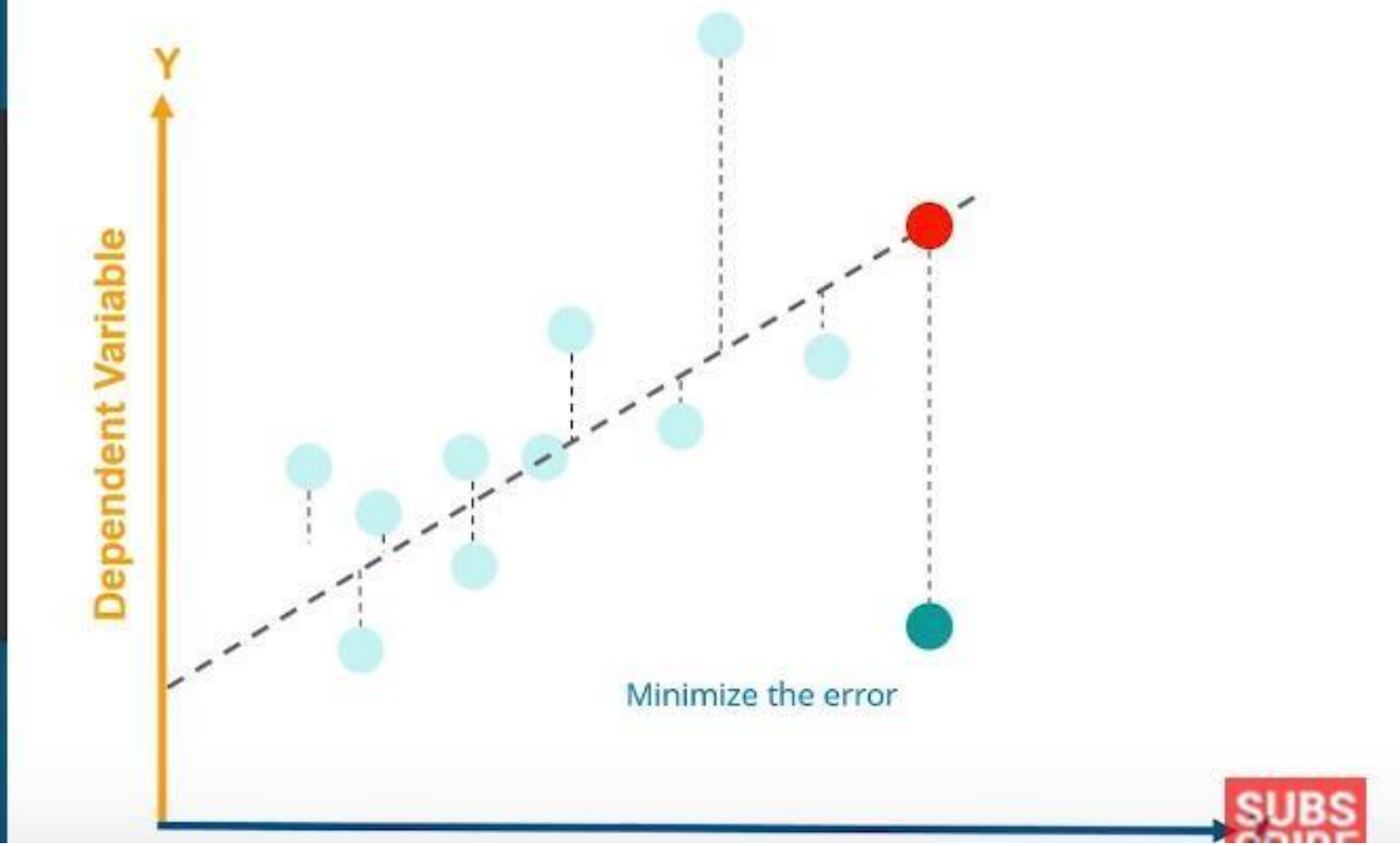
Understanding Linear Regression Algorithm



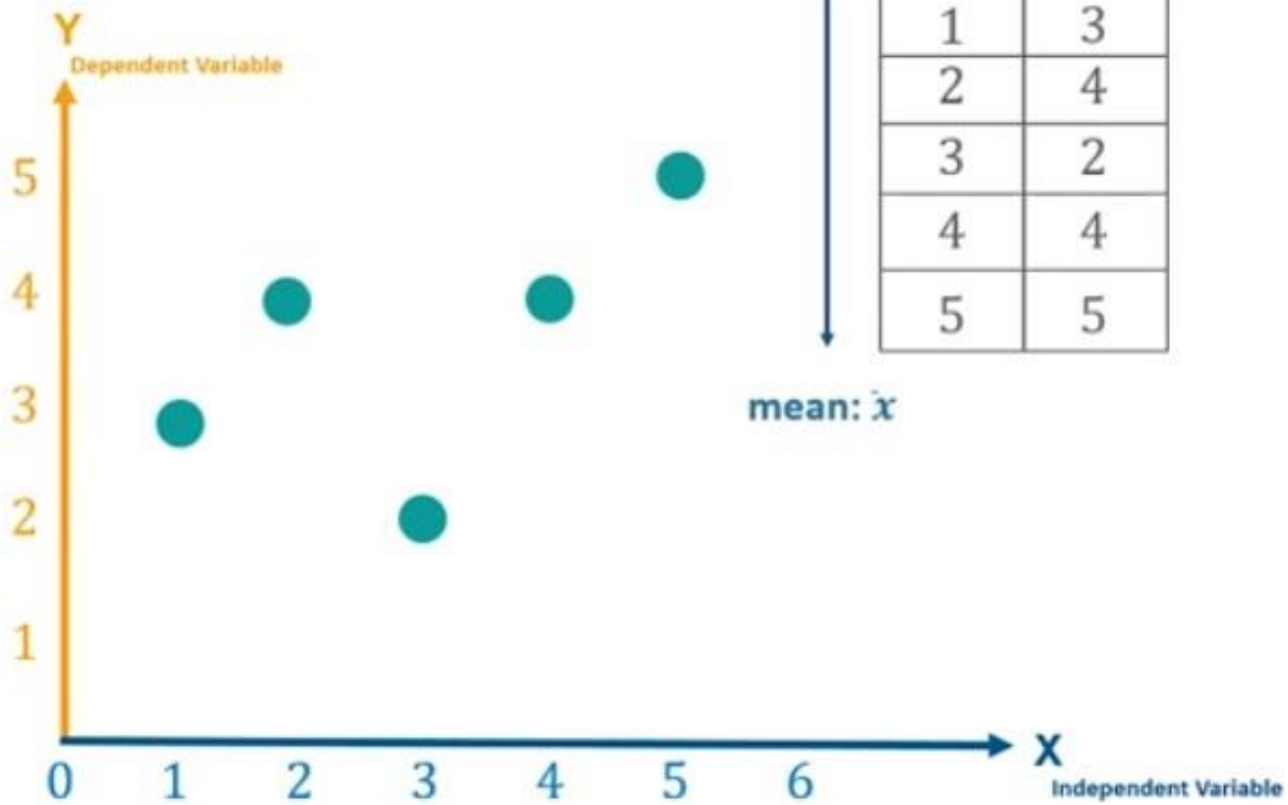
Understanding Linear Regression Algorithm



Understanding Linear Regression Algorithm



Understanding Linear Regression Algorithm

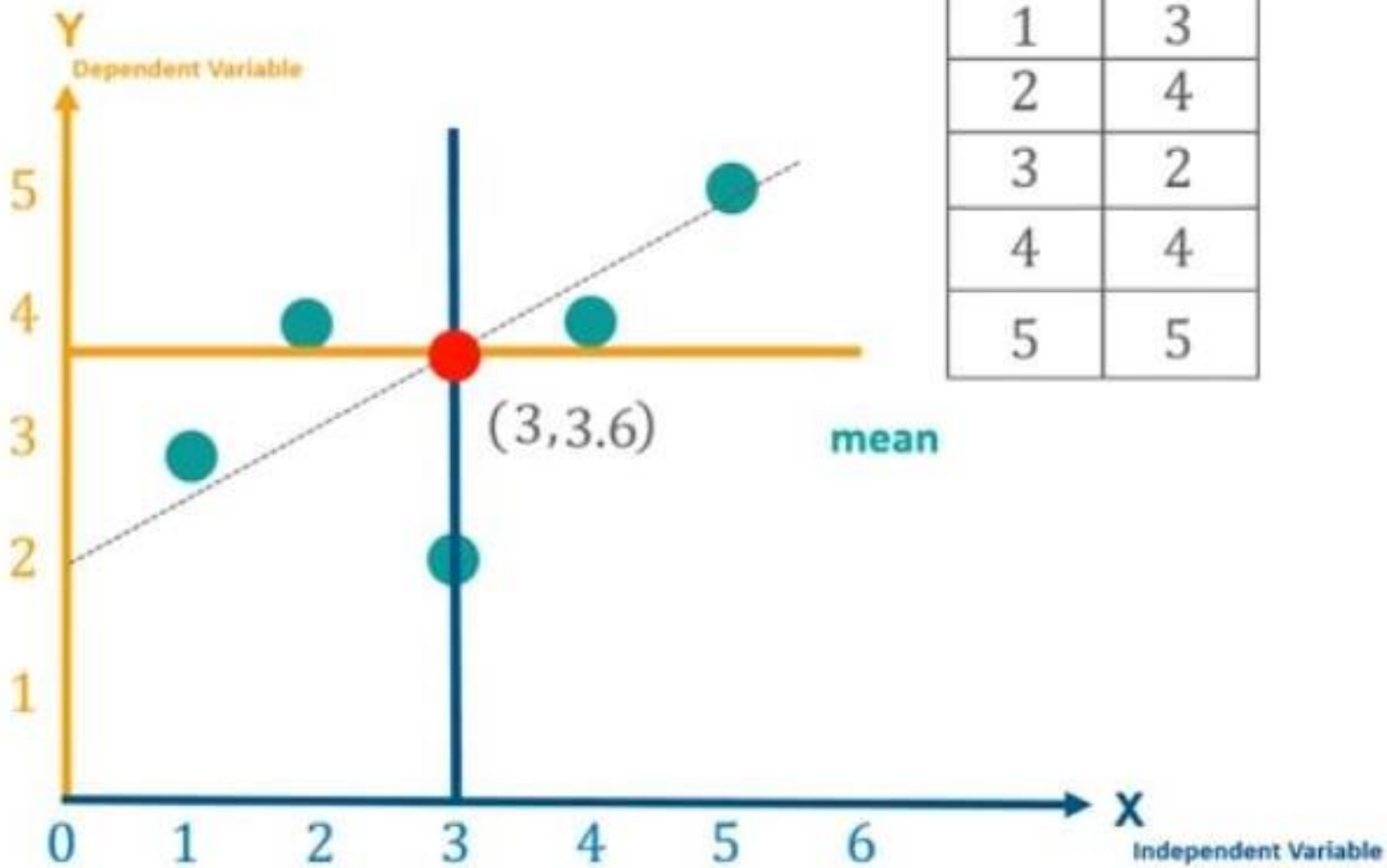


x	y
1	3
2	4
3	2
4	4
5	5

mean: $\bar{x}$

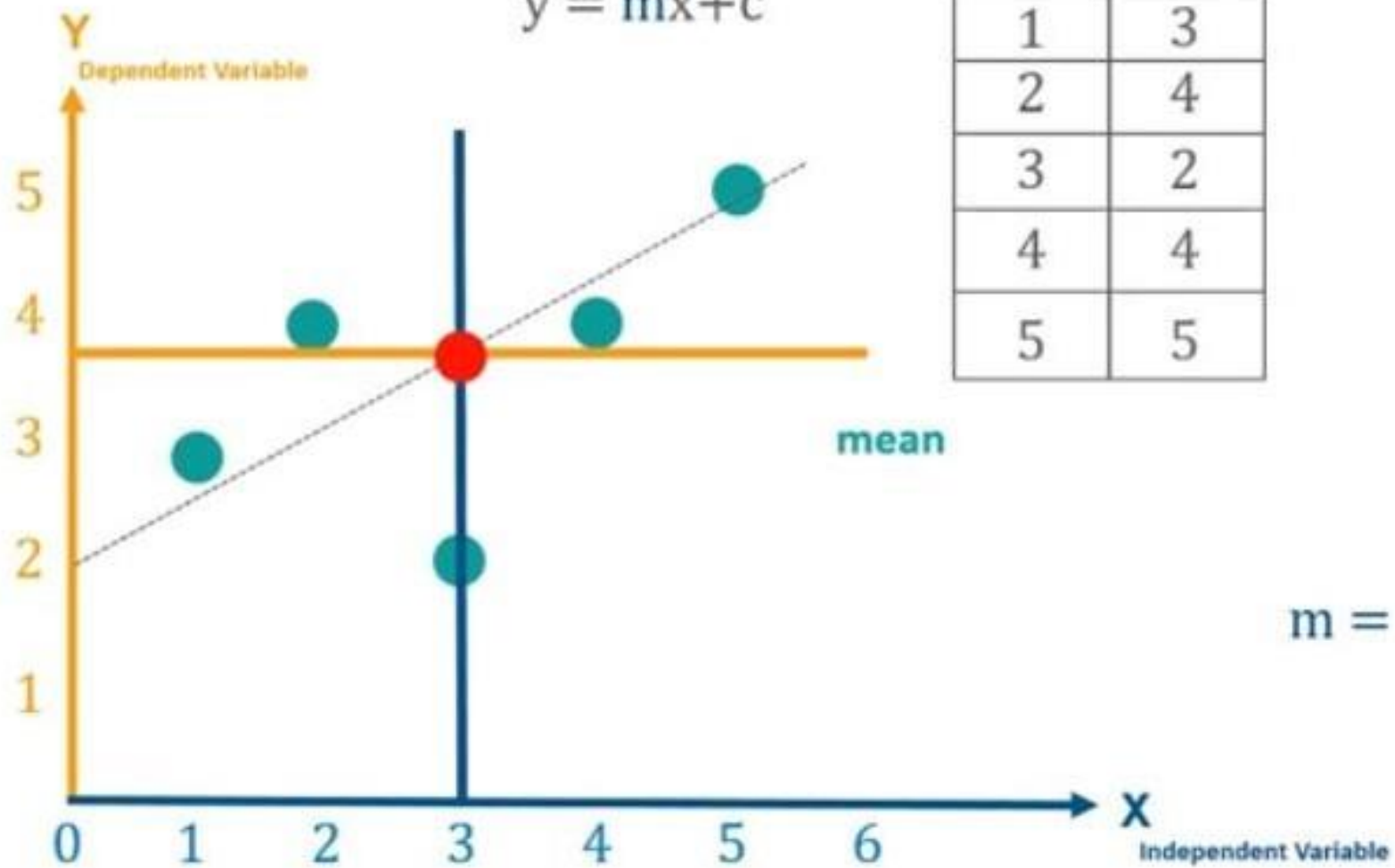
Understanding Linear Regression Algorithm

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



Understanding Linear Regression Algorithm

$$y = mx+c$$

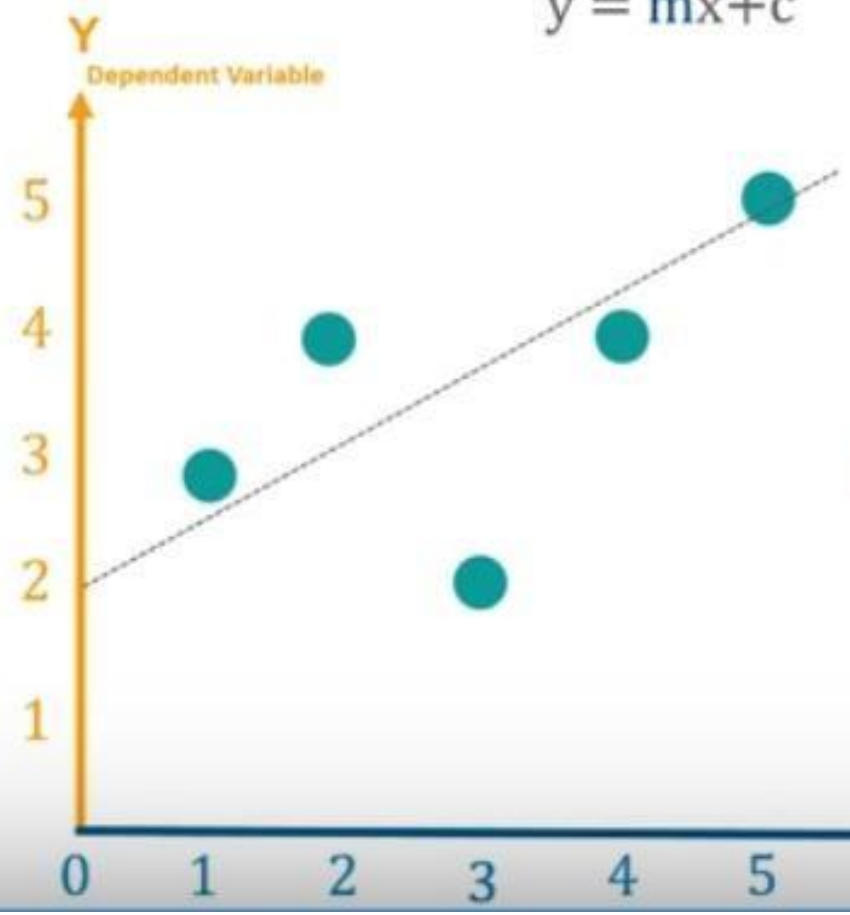


x	y
1	3
2	4
3	2
4	4
5	5

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

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	4				
3	2				
4	4				
5	5				

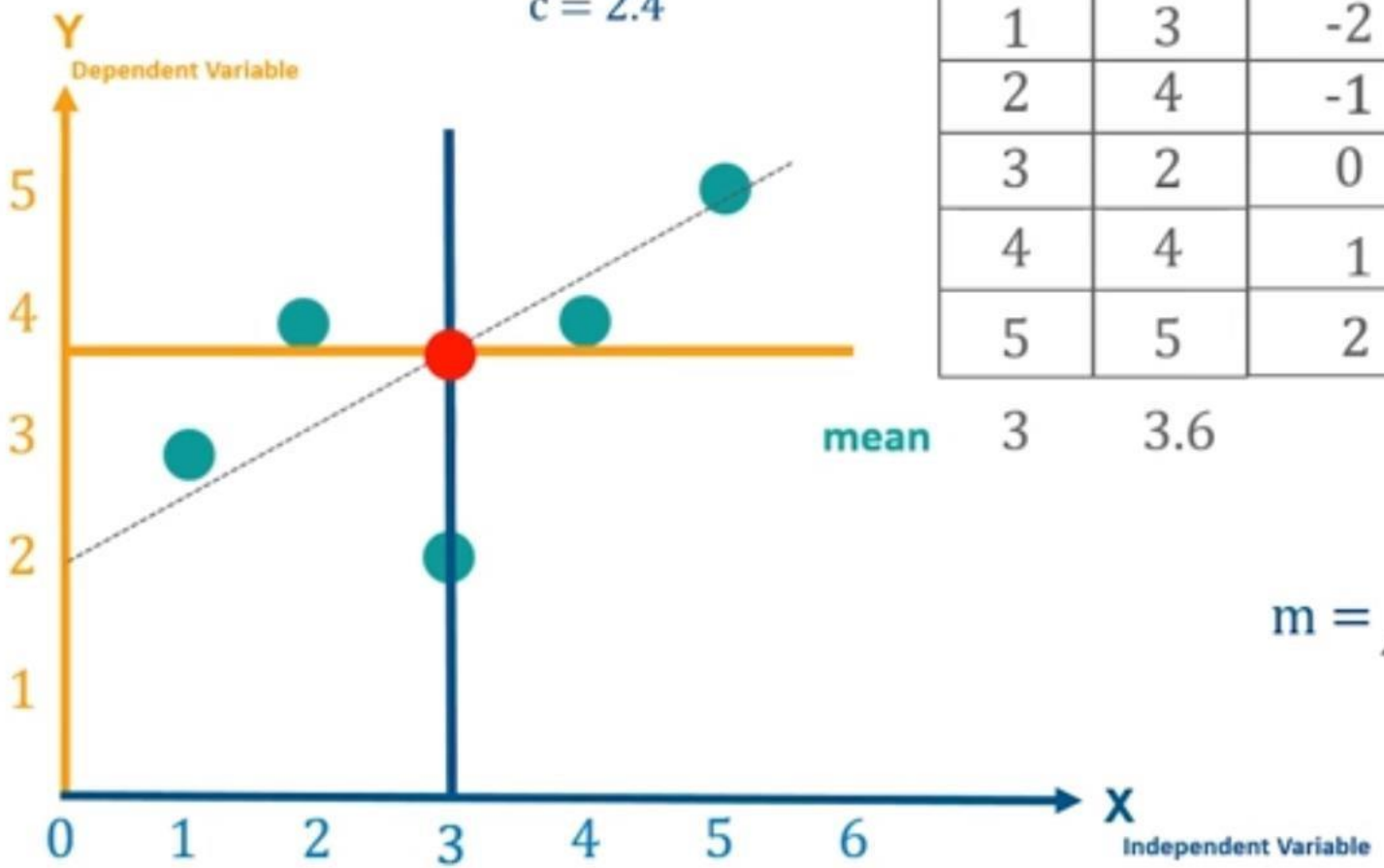
mean

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

Understanding Linear Regression Algorithm

$$y = mx+c$$

$$c = 2.4$$



x	y	$x-\dot{x}$	$y-\dot{y}$	$(x-\dot{x})^2$	$(x-\dot{x})(y-\dot{y})$
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		$\Sigma = 10$	$\Sigma = 4$

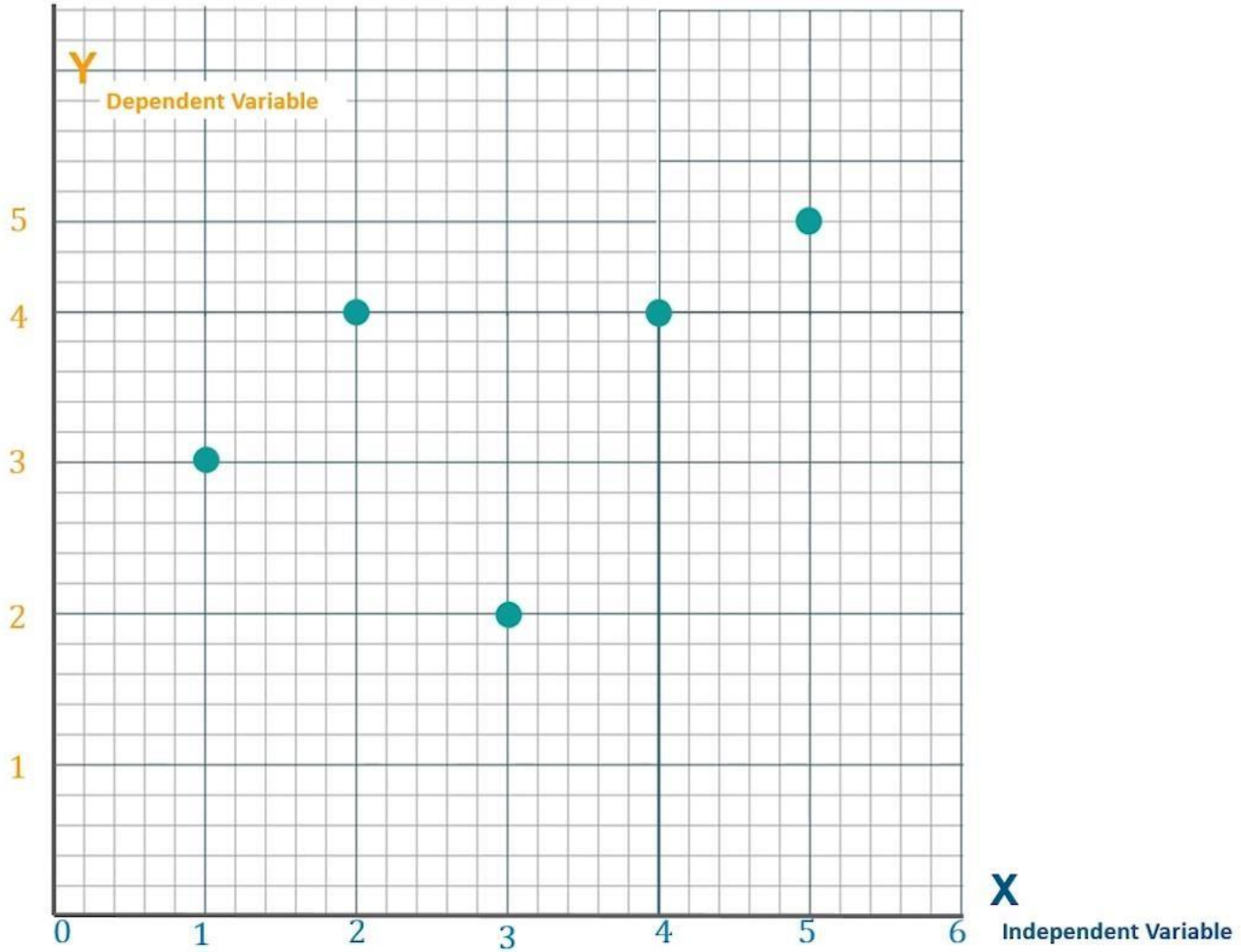
$$m = \sum \frac{(x-\dot{x})(y-\dot{y})}{(x-\dot{x})^2} = \frac{4}{10}$$

$$m = 0.4$$

$$c = 2.4$$

$$y = 0.4x + 2.4$$

Mean Square Error

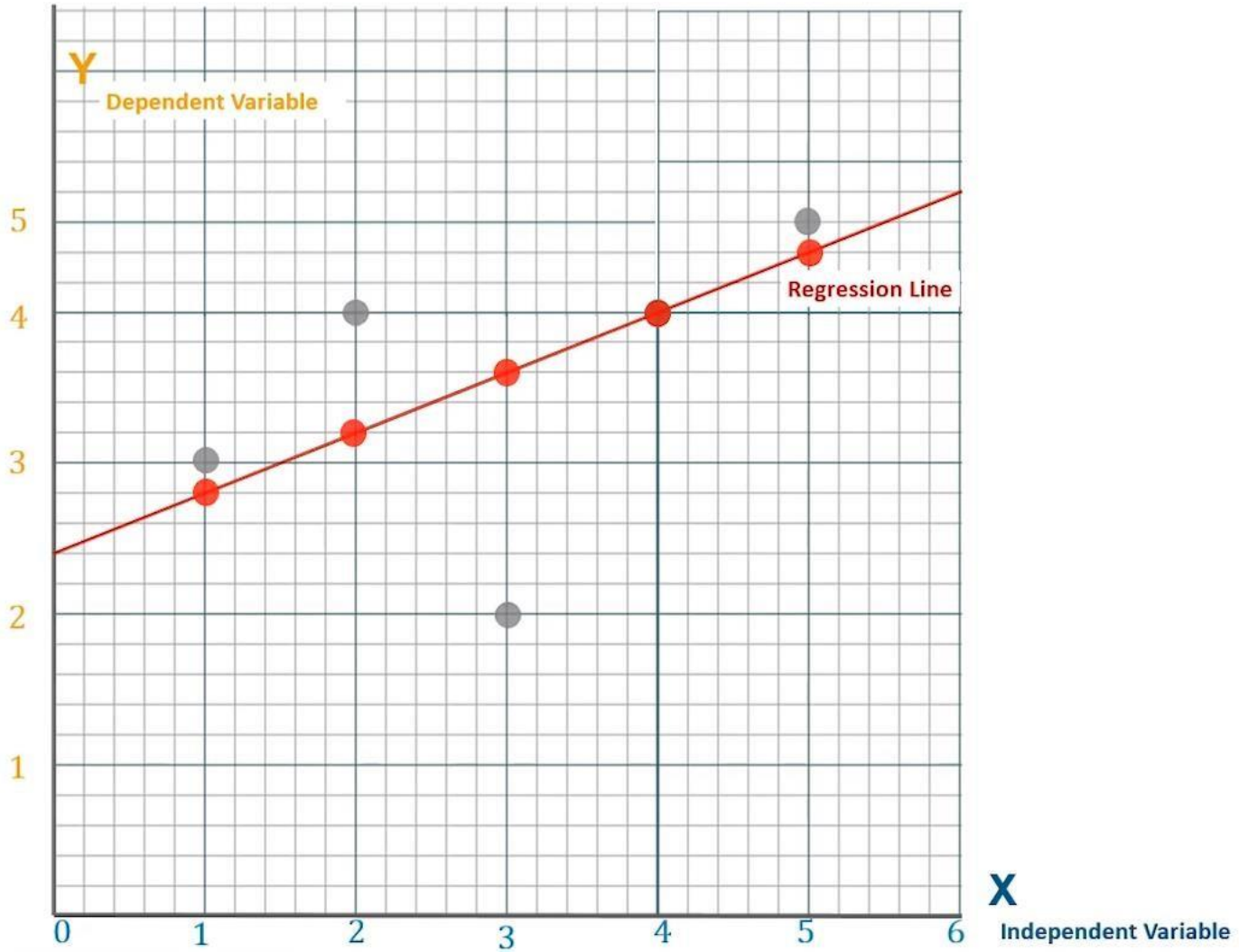


$$m = 0.4$$
$$c = 2.4$$
$$y = 0.4x + 2.4$$

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

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

Mean Square Error

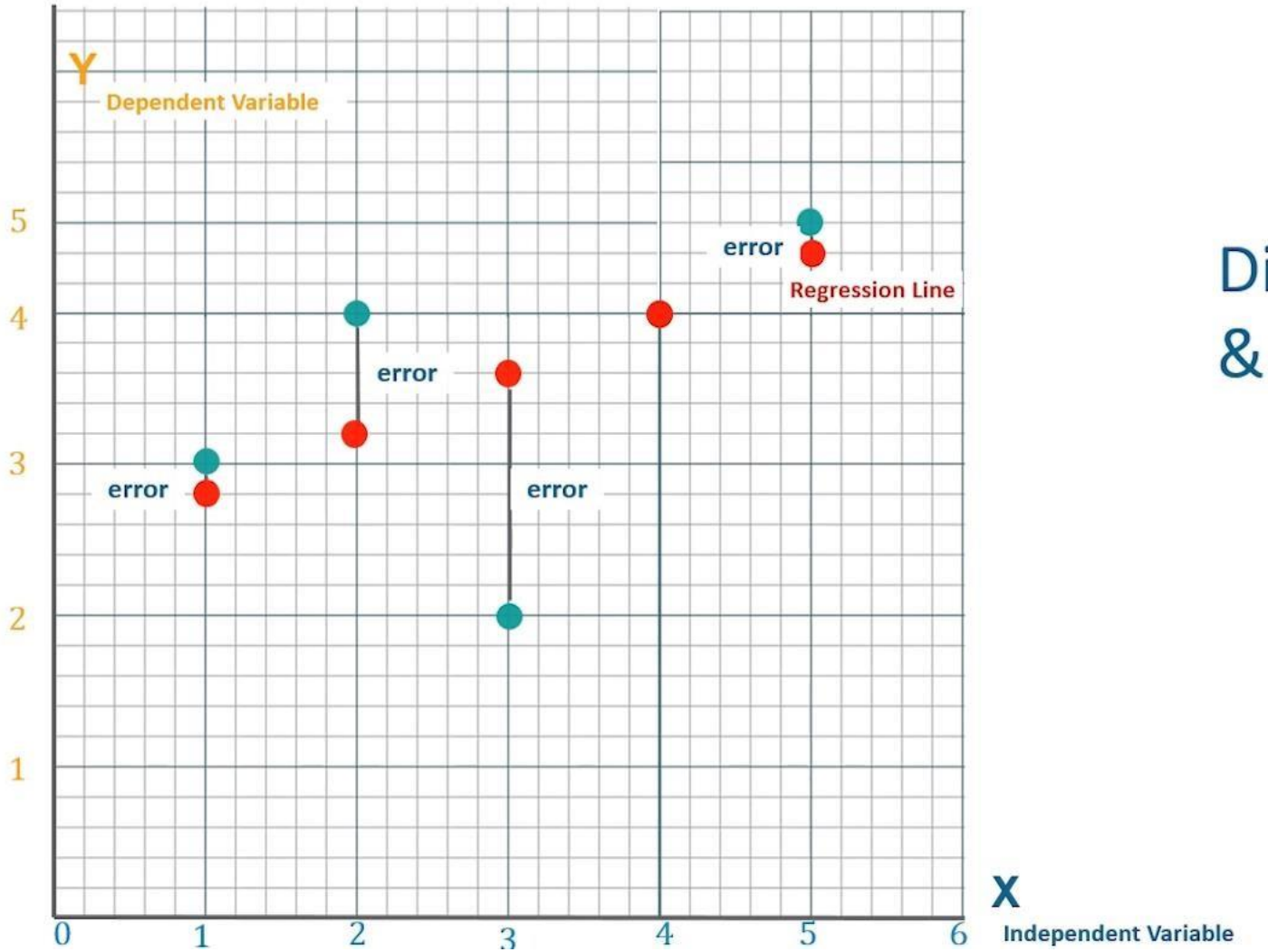


$$m = 0.4$$
$$c = 2.4$$
$$y = 0.4x + 2.4$$

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

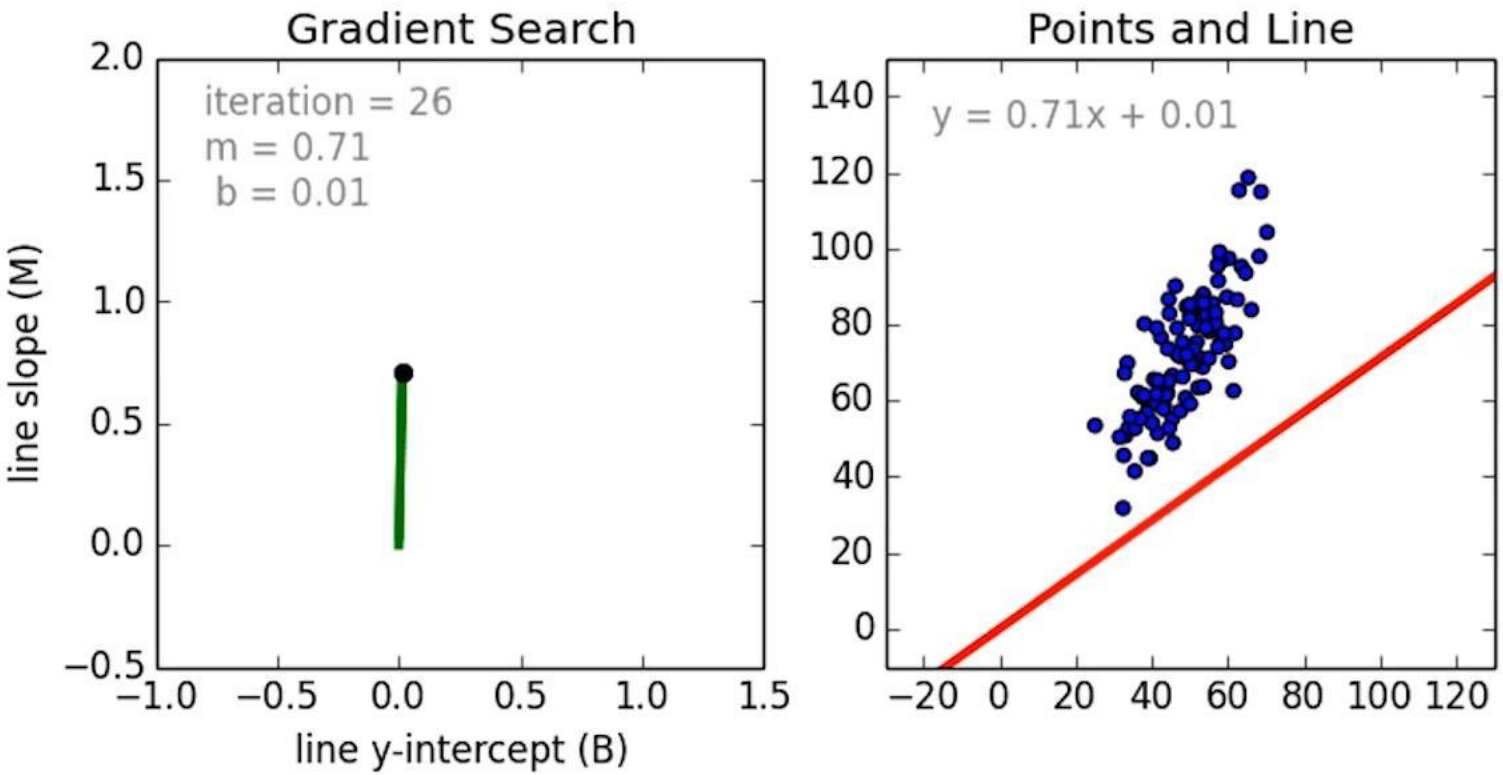
$$y = 0.4 \times 1 + 2.4 = 2.8$$
$$y = 0.4 \times 2 + 2.4 = 3.2$$
$$y = 0.4 \times 3 + 2.4 = 3.6$$
$$y = 0.4 \times 4 + 2.4 = 4.0$$
$$y = 0.4 \times 5 + 2.4 = 4.4$$

Mean Square Error



Distance between actual & predicted value

Finding the best fit line

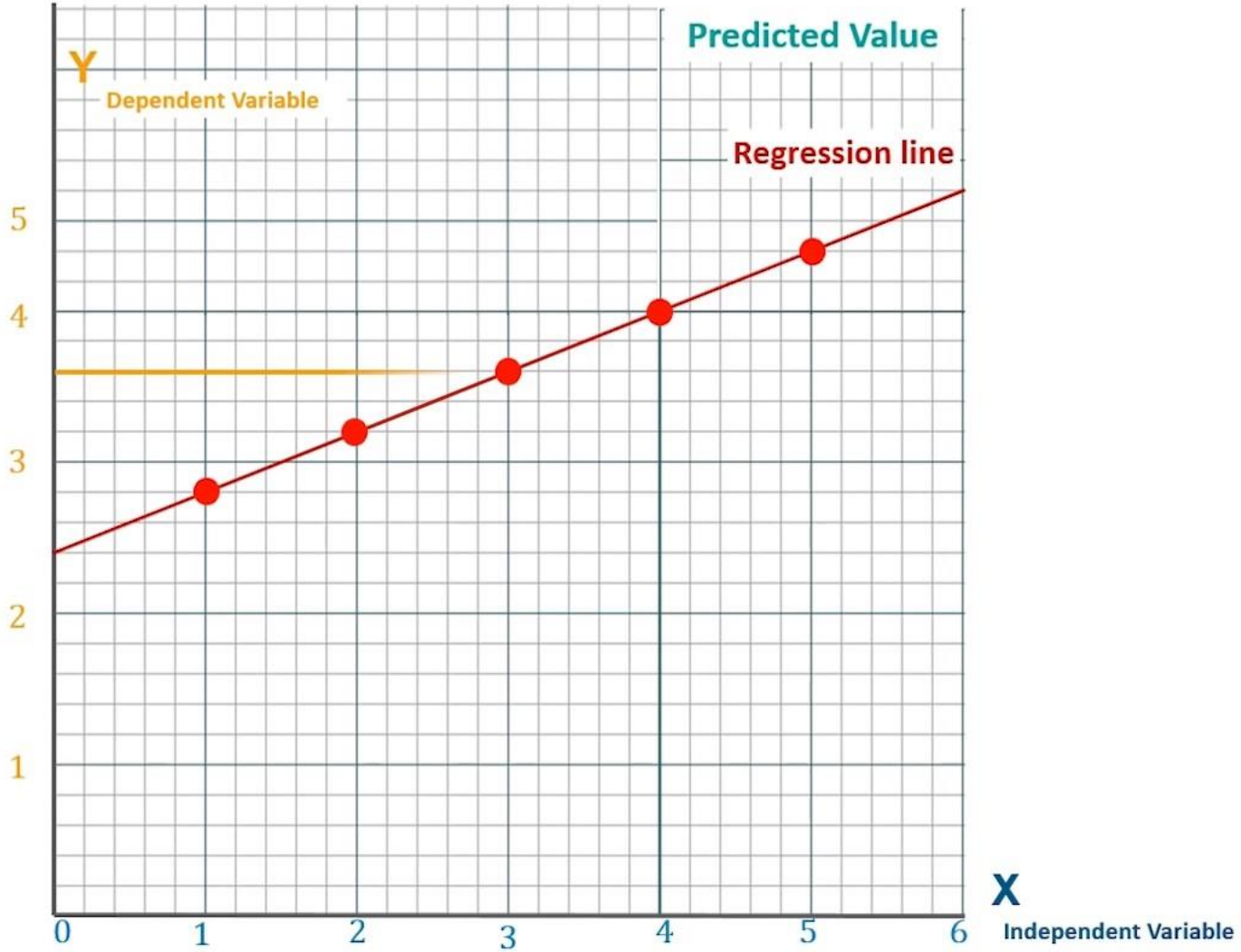


Let’s check the Goodness of fit

What is R-Square?

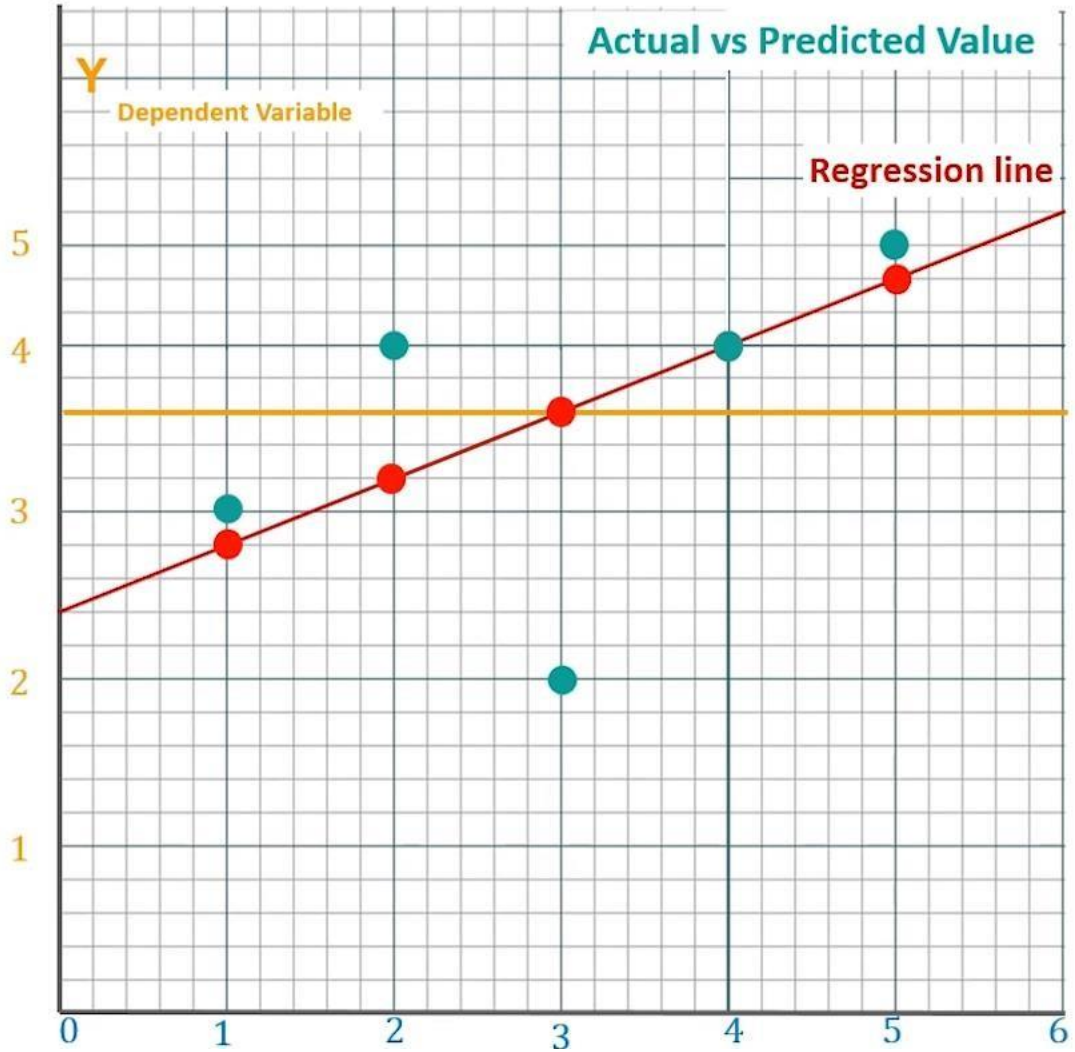
- **R-squared** value is a statistical measure of how close the data are to the fitted regression line
- It is also known as **coefficient of determination**, or the **coefficient of multiple determination**

Calculation of R^2



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

Calculation of R^2



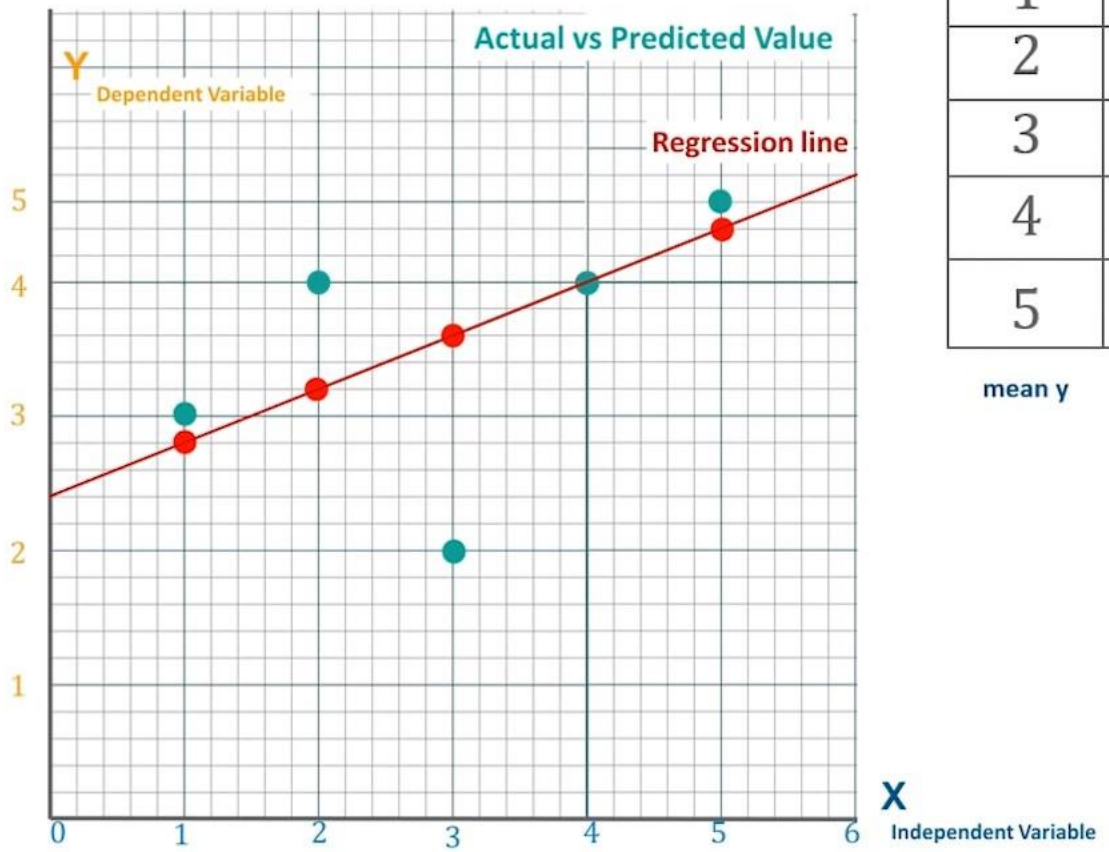
Distance actual - mean

vs

Distance predicted - mean

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

Calculation of R^2

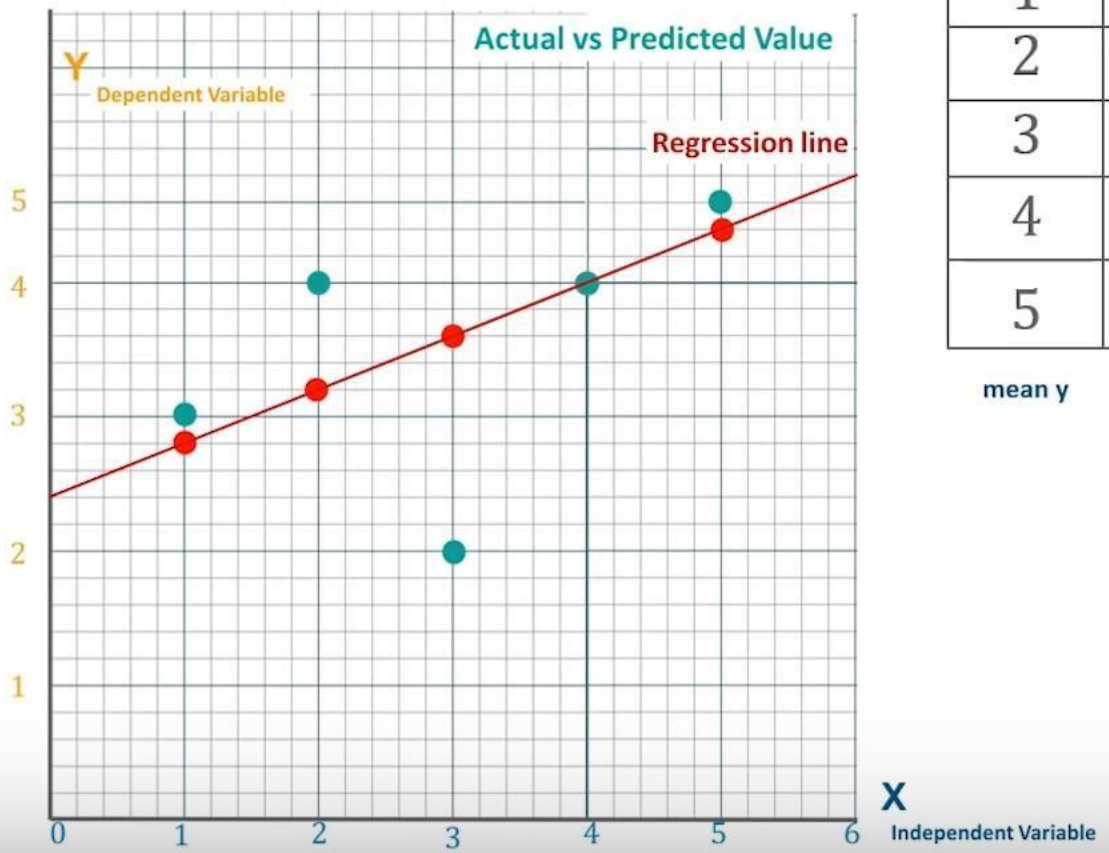


x	y	$y - \bar{y}$
1	3	– 0.6
2	4	0.4
3	2	–1.6
4	4	0.4
5	5	1.4

mean y 3.6

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

Calculation of R^2

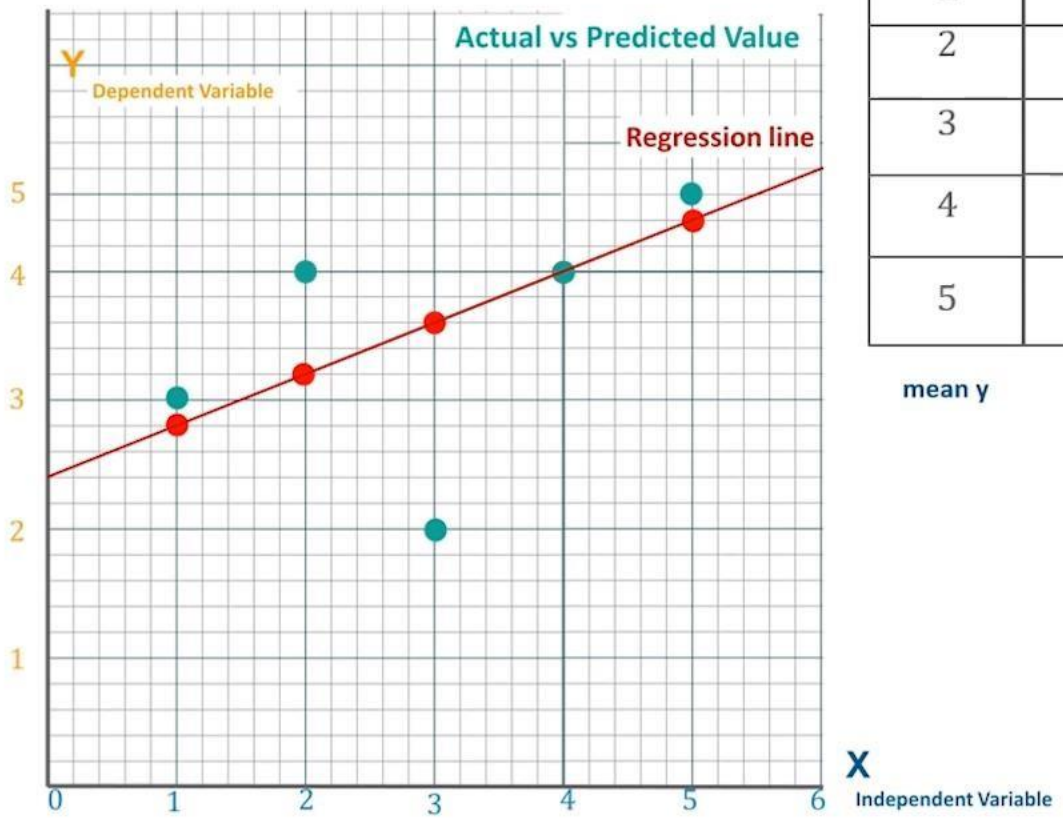


x	y	$y-\bar{y}$	$(y-\bar{y})^2$	y_p	$(y_p-\bar{y})$
1	3	– 0.6	0.36	2.8	-0.8
2	4	0.4	0.16	3.2	-0.4
3	2	–1.6	2.56	3.6	0
4	4	0.4	0.16	4.0	0.4
5	5	1.4	1.96	4.4	0.8

mean y 3.6

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

Calculation of R^2

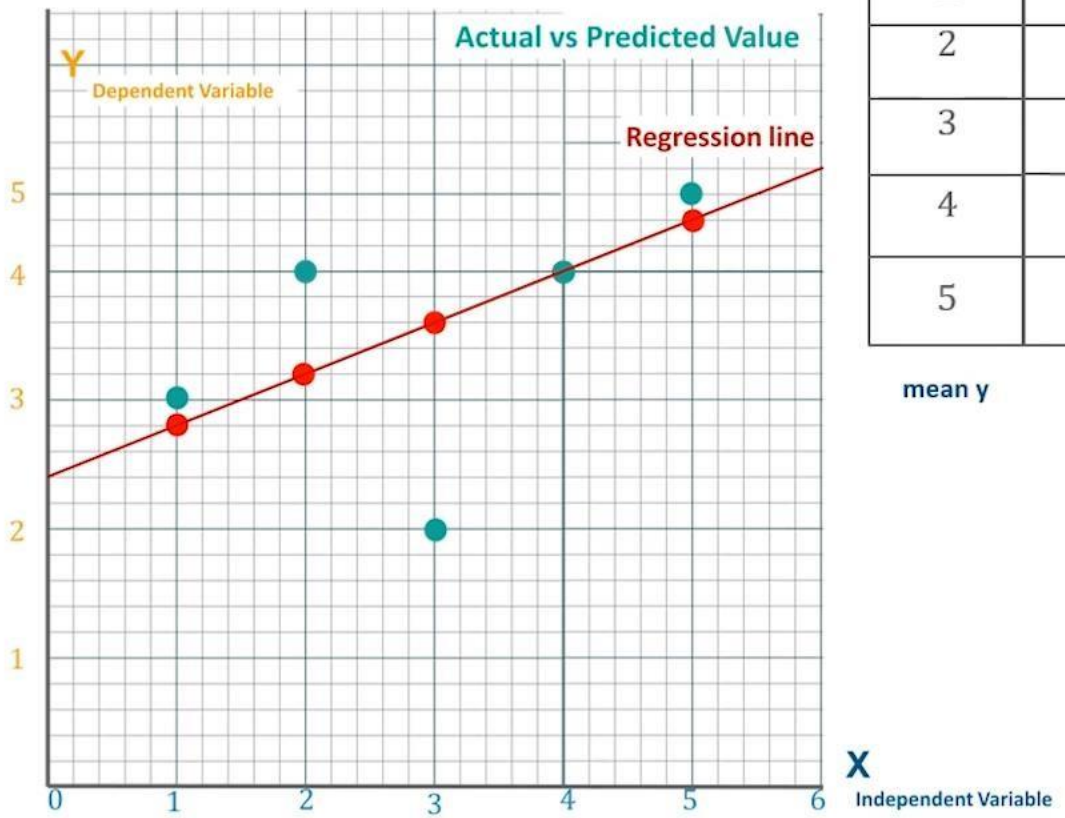


x	y	$y - \dot{y}$	$(y - \dot{y})^2$	y_p	$(y_p - \dot{y})$	$(y_p - \dot{y})^2$
1	3	– 0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	–1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64

mean y 3.6

$$\frac{\sum (y_p - \dot{y})^2}{\sum (y - \dot{y})^2}$$

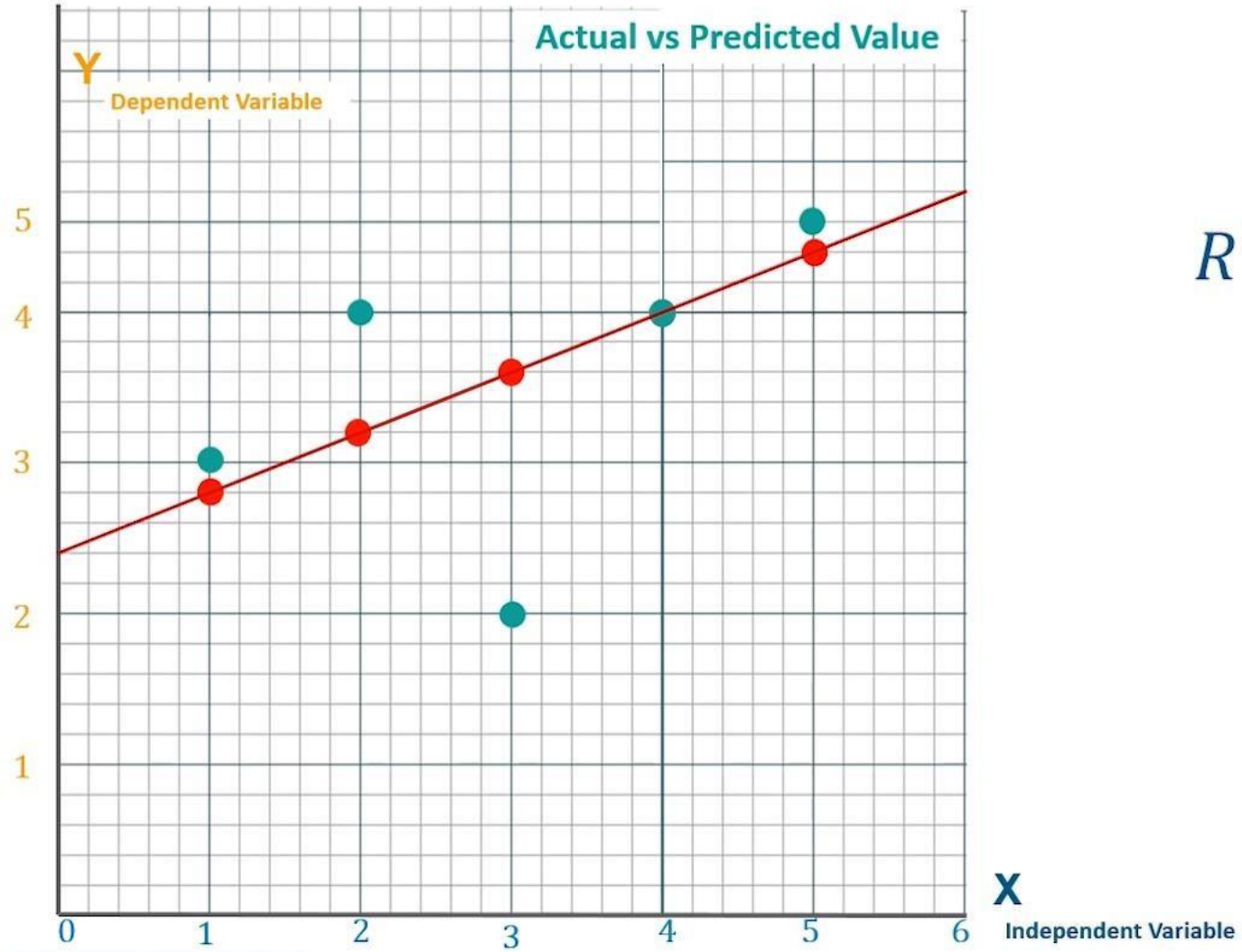
Calculation of R^2



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

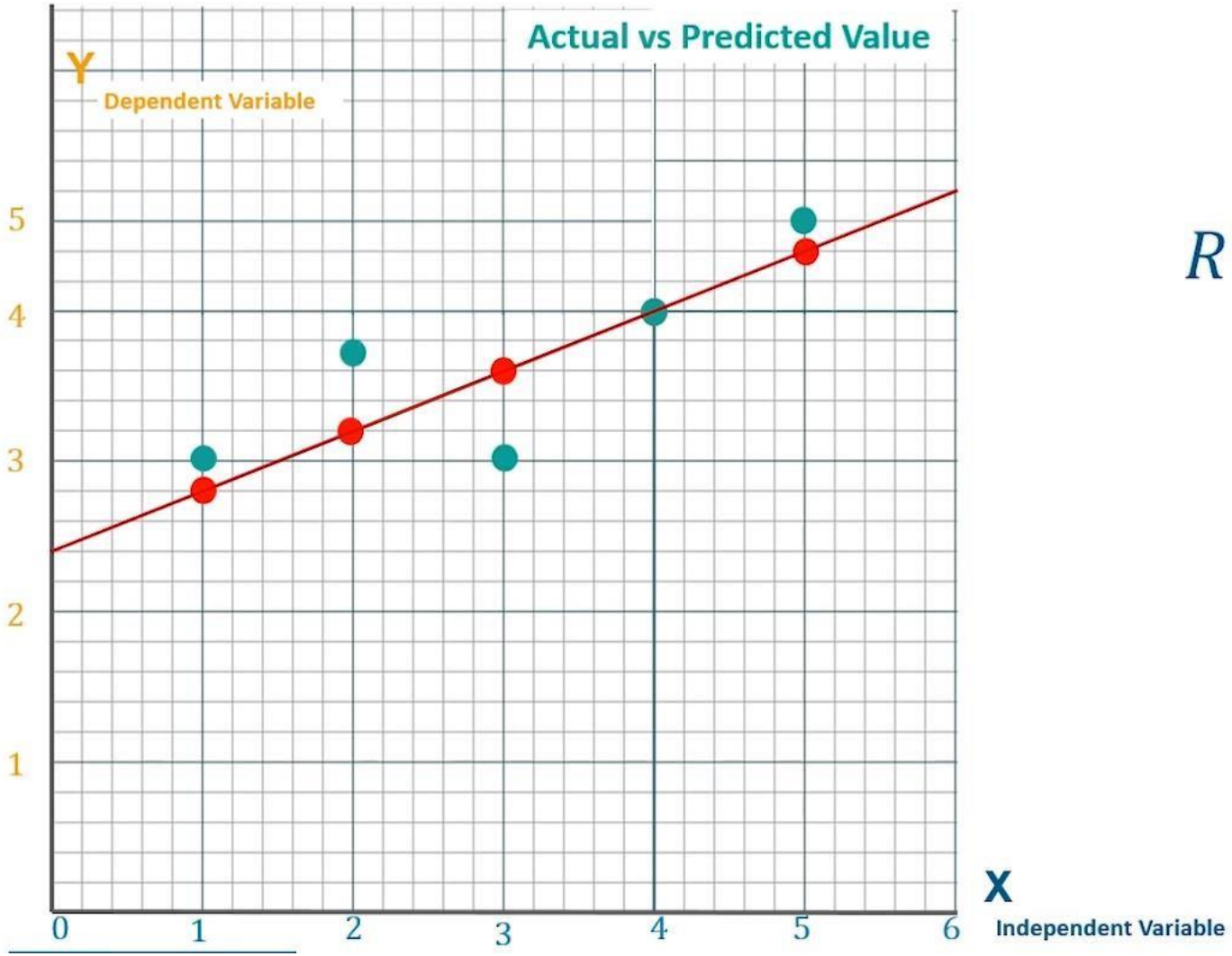
$$R^2 = \frac{1.6}{5.2} = \frac{\sum \left((y_p - \bar{y}\right)^2}{\sum (y-\bar{y})^2}$$

Calculation of R^2



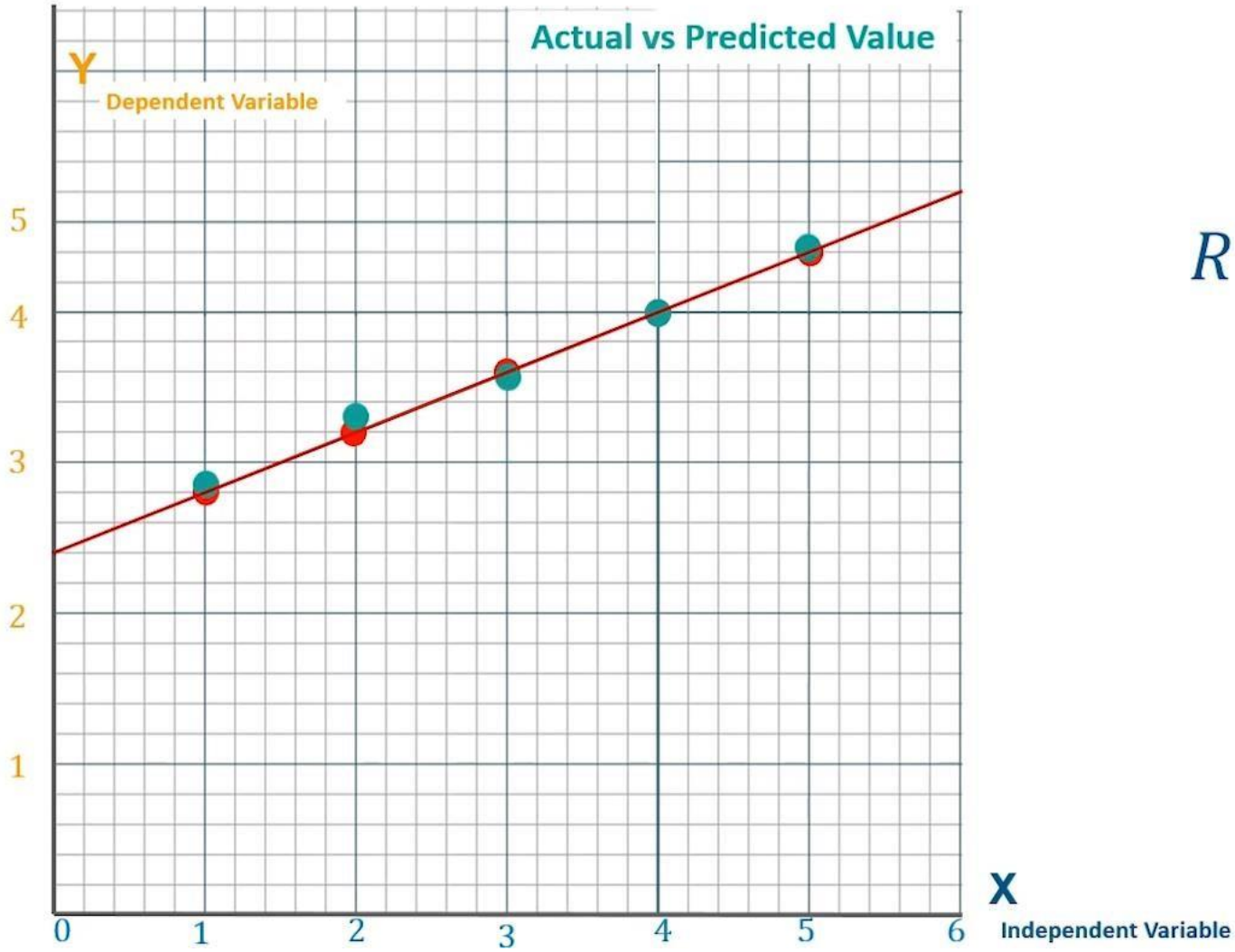
$$R^2 \approx 0.3$$

Calculation of R^2



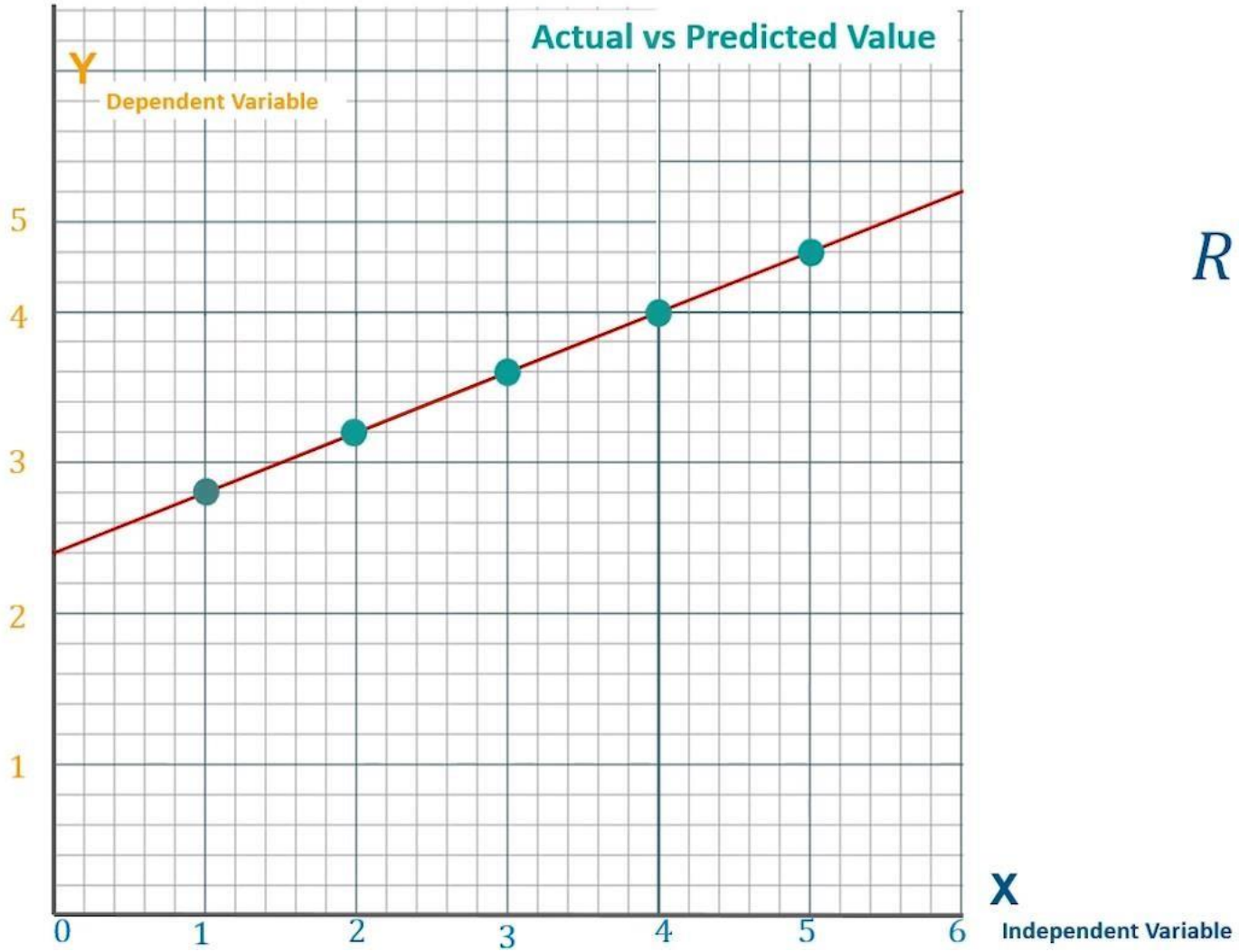
$$R^2 \approx 0.7$$

Calculation of R^2



$$R^2 \approx 0.9$$

Calculation of R^2



$$R^2 \approx 1$$

Karl Pearson’s coefficient of correlation

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

Standard Error of Estimate

- The standard error of the estimate is a measure of the accuracy of predictions. It is given by

$$\sigma_{est} = \sqrt{\Sigma(Y - Y_p)^2/\mathrm{N}-2}$$

x	y	y predict	y - y predict	(y - y predict)^2
1	3	2.8	0.2	0.04
2	4	3.2	0.8	0.64
3	2	3.6	-1.6	2.56
4	4	4	0	0
5	5	4.4	0.6	0.36
				Sum = 3.6

y = b0 + b1 x

b0 = 2.4

b1 = 0.4

y = 2.4 + 0.4x

$$\sigma_{est} = \sqrt{\Sigma(Y - Y_p)^2/(N-2)} = 1.095$$

5 **Interpret your result.** The Standard Error of the Estimate is a statistical figure that tells you how well your measured data relates to a theoretical straight line, the line of regression. A score of 0 would mean a perfect match, that every measured data point fell directly on the line. Widely scattered data will have a much higher score.[6]

Standard Error of the Estimate

X	Y
3.25	18.71
3.96	18.15
4.35	19.72
4.40	23.02
4.42	22.26
4.51	19.61
4.87	27.74
5.65	24.89
5.68	27.83
5.71	23.09
6.28	24.25
6.52	31.55

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.788214
R Square	**0.621281**
Adjusted R Square	0.583409
Standard Error	**2.660008**
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**1**	116.0743271	116.0743	16.40477	0.00232253
Residual	10	70.75643953	7.075644		
Total	11	186.8307667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	**7.152578**	4.084668789	1.751079	0.110488	-1.948630984	16.25379	-1.94863	16.25379
X Variable 1	**3.271629**	0.807753571	4.050281	0.002323	1.471841428	5.071416	1.471841	5.071416

Standard Error of the Estimate

x	y	y'	$y-y'$	$(y-y')^2$
1	2			
2	4			
3	5			
4	4			
5	5			

Summary

- Least squares method helps in computing the minimum of the squares of errors between actual and predicted values.
- Accuracy of the model is predicted by
  - Karl Pearson's coefficient whose value ranges from -1 to +1, where 1 is the total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.
  - R^2 (R-square) variable also helps to determine the correlation between input and output variables. High value of R^2 indicates a strong linear relationship.
  - Standard error of the estimate is a measure of the accuracy of predictions.