Regression Analysis

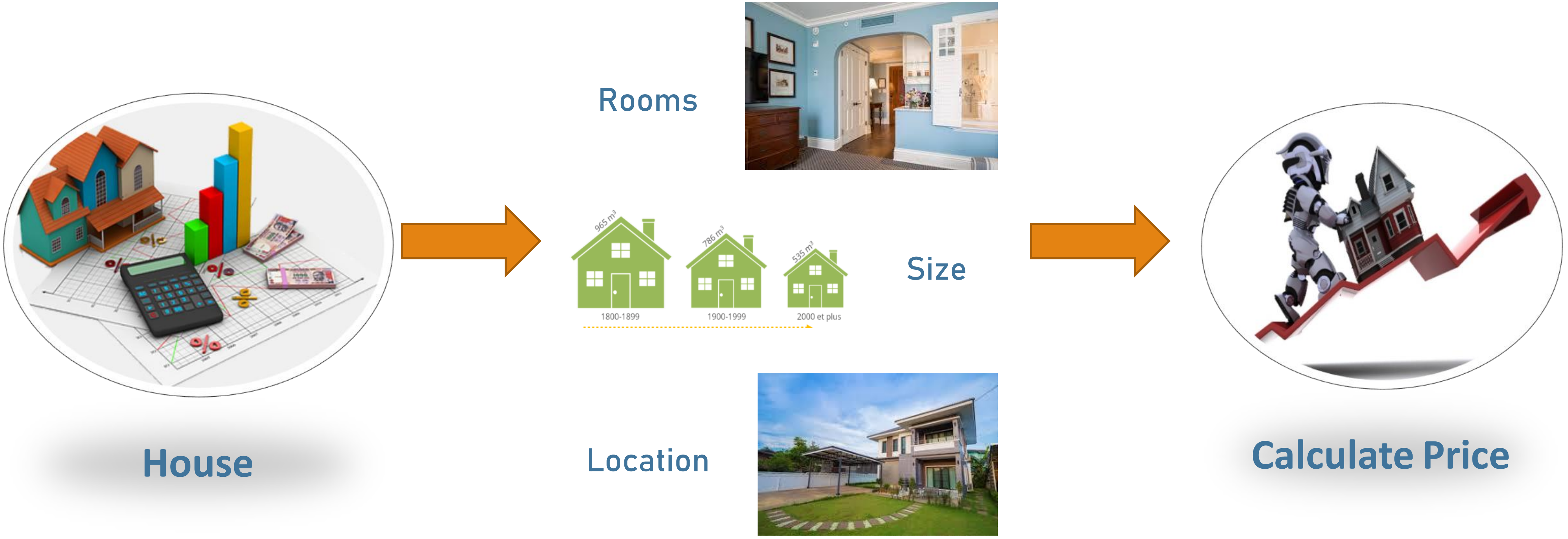
ANINDITA KUNDU

LECTURER

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MILITARY INSTITUTE OF SCIENCE AND TECHNOLOGY

House Price Prediction



House

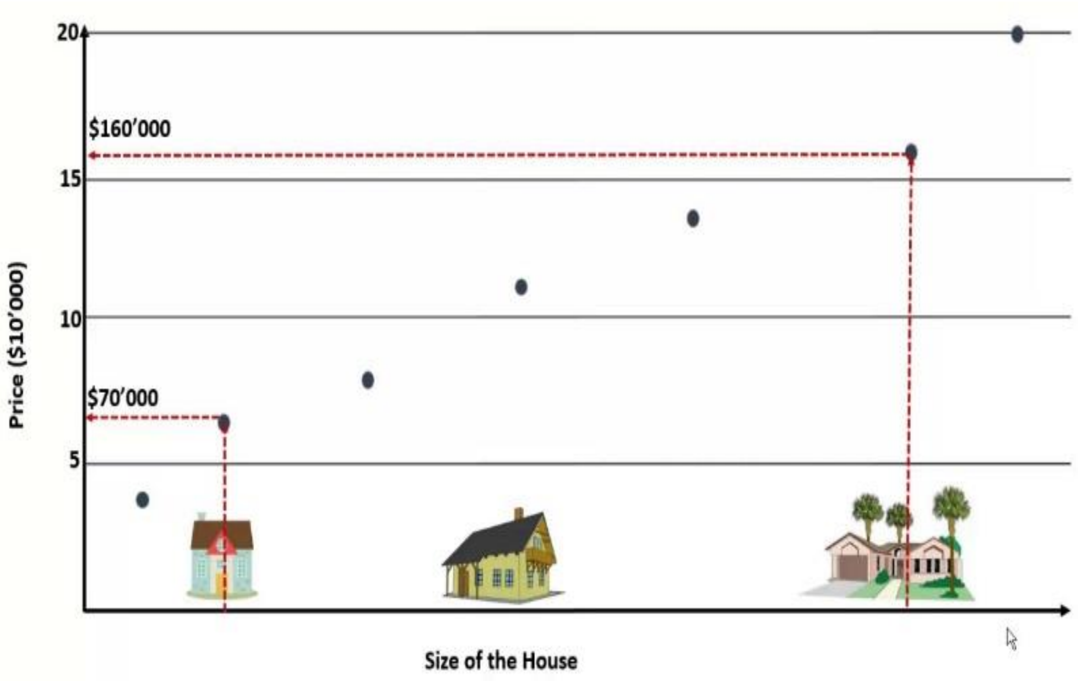
Rooms

Size

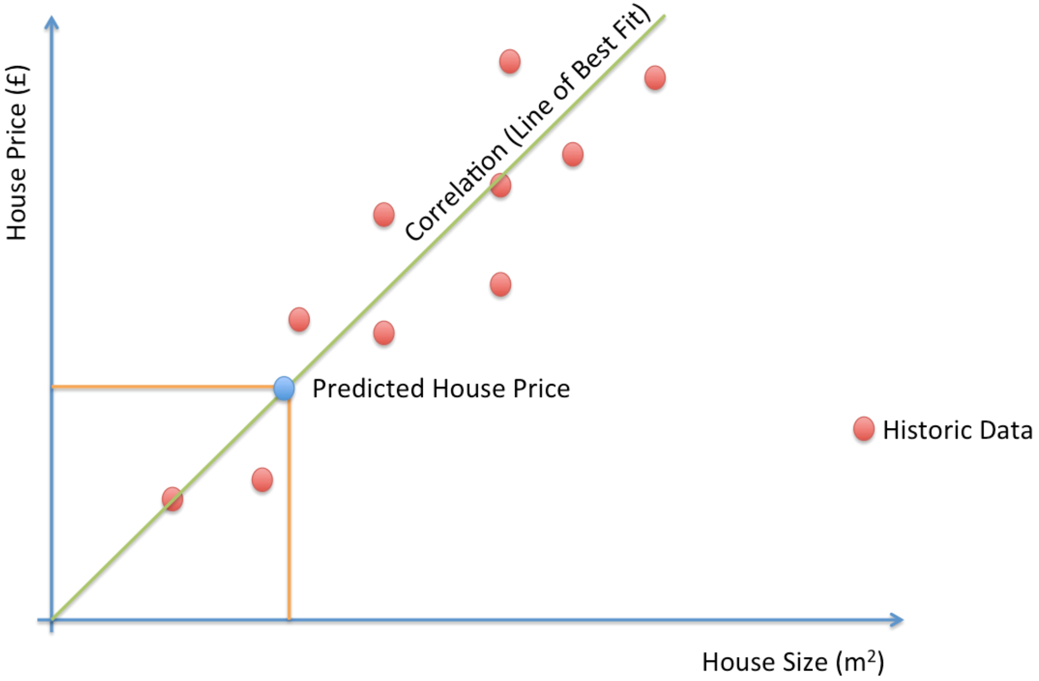
Location

Calculate Price

House Price Prediction



Plotting price based on Size of the house

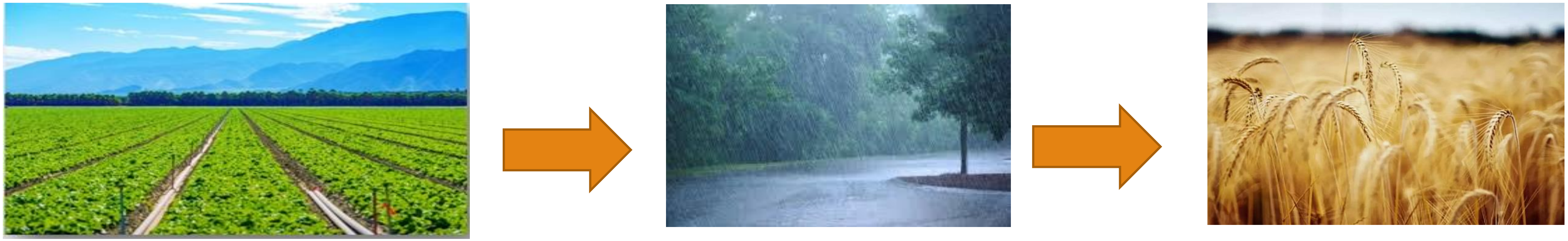


Prediction line to estimate price

Machine Learning Process

Machine learning (**ML**) is the study of computer algorithms that improve automatically through experience. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.

Based on the amount of rainfall, how much would be the crop yield?

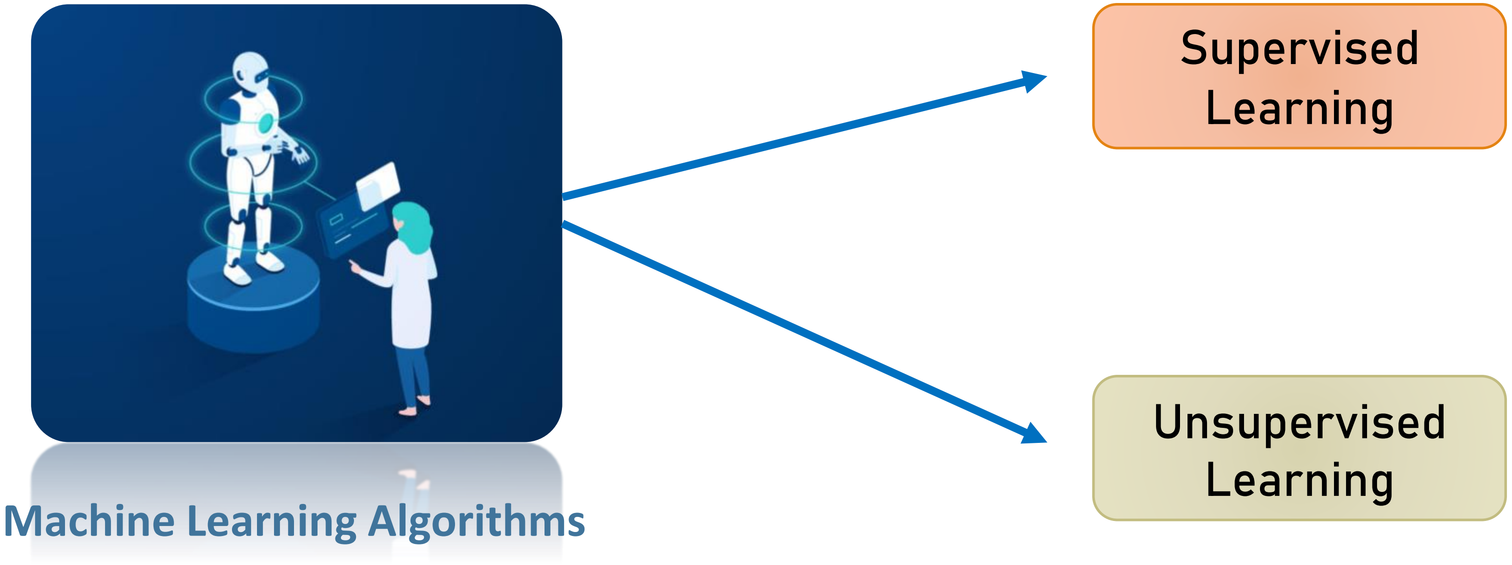


Crop Field

Rainfall

Crop Yield

Machine Learning Algorithms



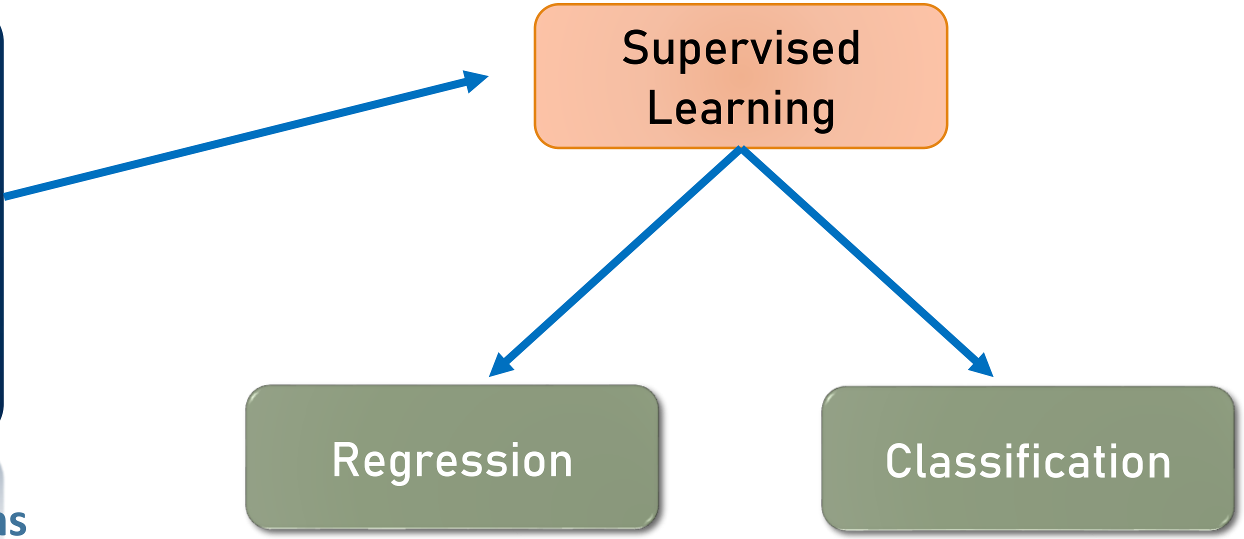
Machine Learning Algorithms

Supervised Learning

Unsupervised Learning

Machine Learning Algorithms

Machine Learning Algorithms



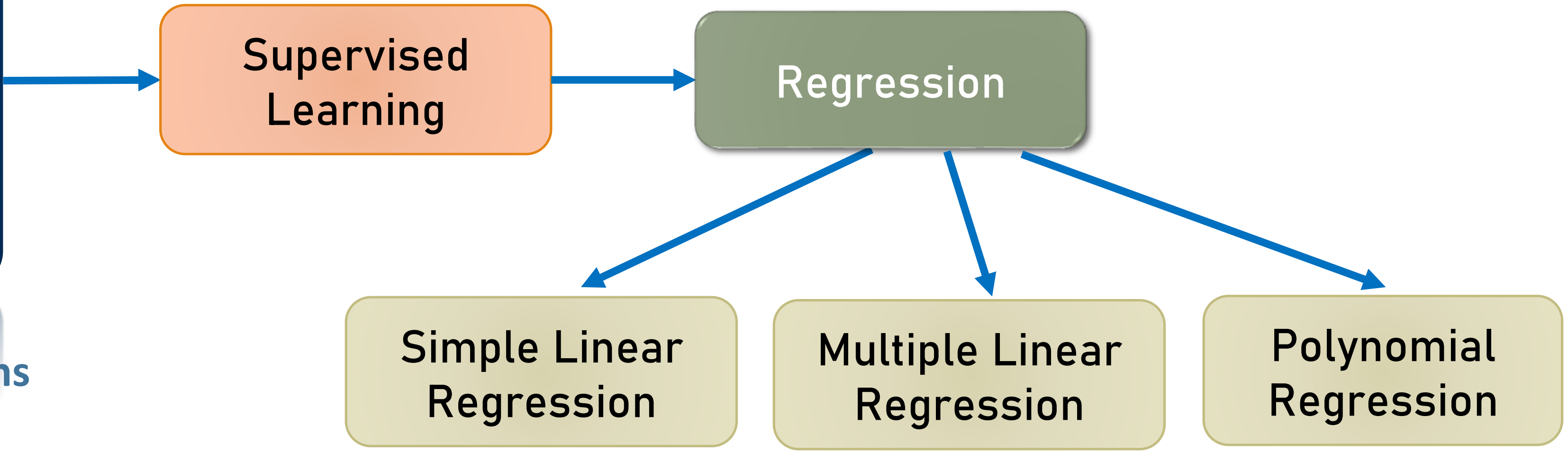
Supervised Learning

Regression

Classification

Machine Learning Algorithms

Machine Learning Algorithms



Supervised Learning

Regression

Simple Linear Regression

Multiple Linear Regression

Polynomial Regression

Usage of Regression

Three major uses for regression analysis are

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

Application of Linear Regression

- ❑ Evaluating Trends and Sales Estimates
- ❑ Analyzing the Impact of Price Changes
- ❑ Assessment of risk in financial services and insurance domain
- ❑ Determining the economic growth of a country in the coming quarter

Understanding Linear Regression

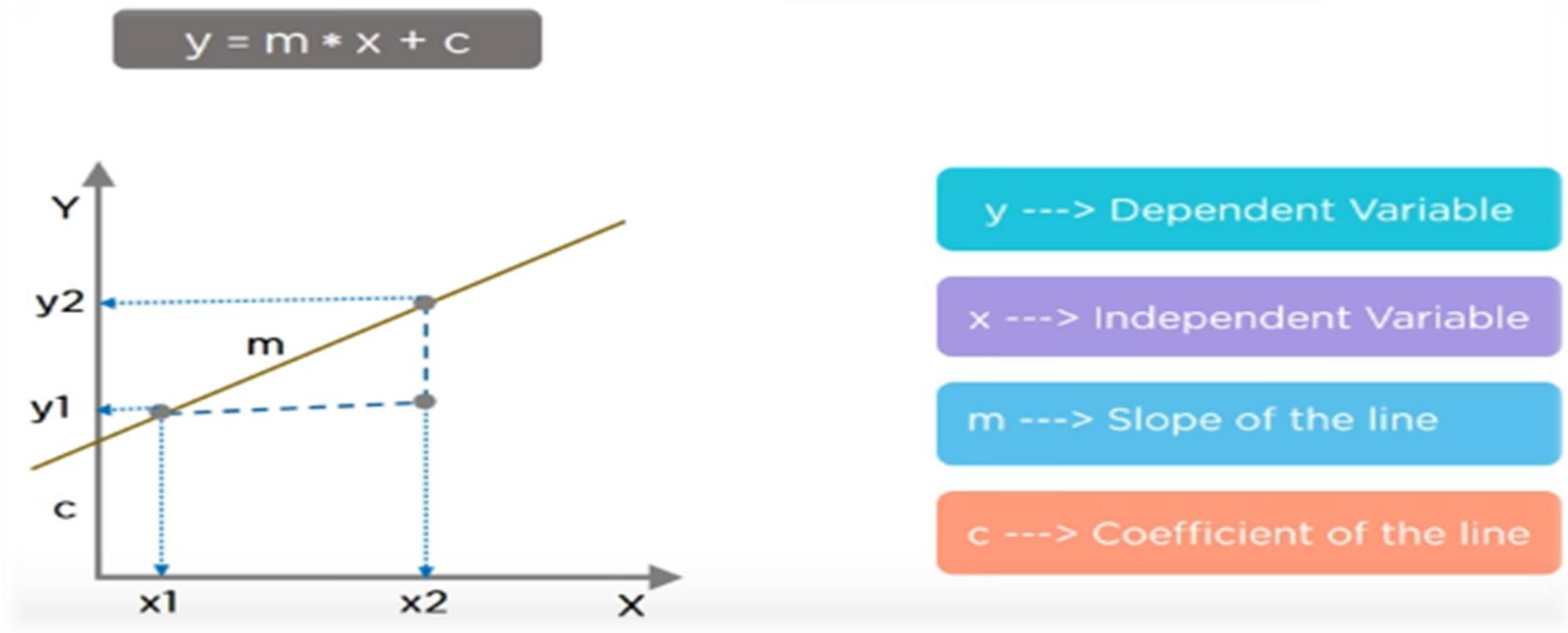
Linear Regression is a statistical model used to predict the relationship between independent and dependent variables. When implementing linear regression of some dependent variable $\boldsymbol{y}$ on the set of independent variables $\mathbf{x} = (x_1, \ldots, x_r)$, where r is the number of predictors, you assume a linear relationship between y and x:

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon$. This equation is the **regression equation**.

$\beta\beta_0, \beta_1, \ldots, \beta_r$ are the **regression coefficients**, and ε is the **random error**.

Regression Equation

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



y = m * x + c

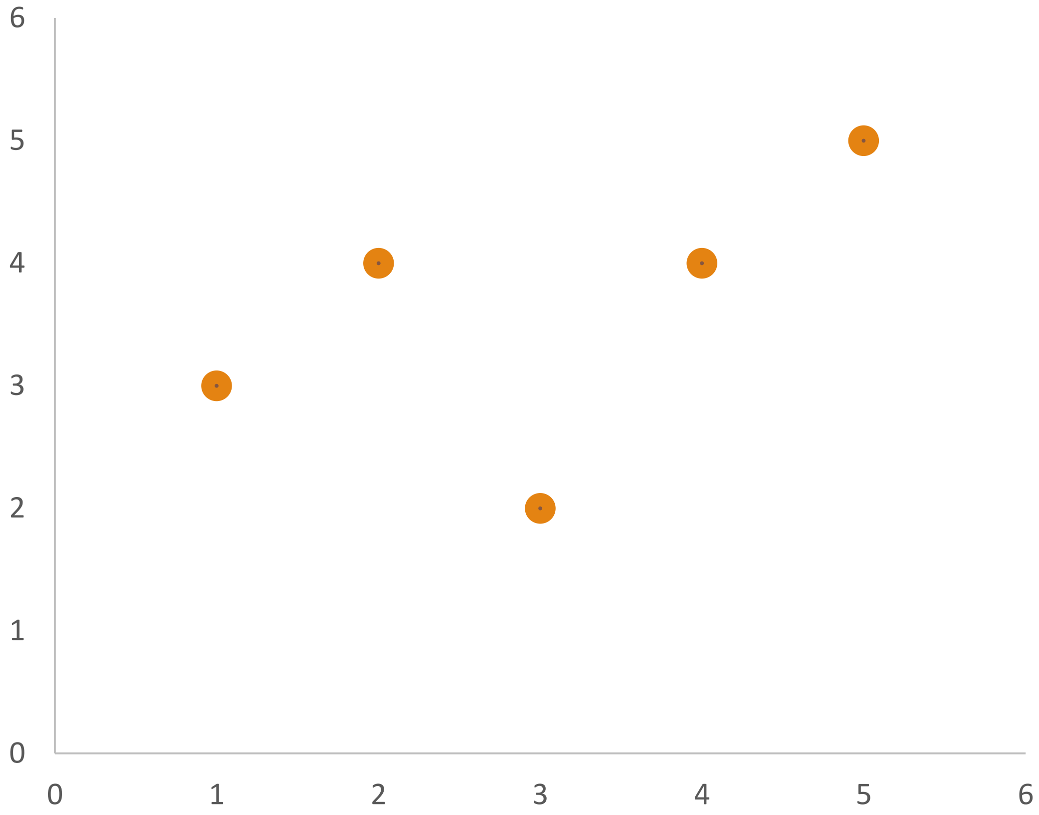
y ---> Dependent Variable

x ---> Independent Variable

m ---> Slope of the line

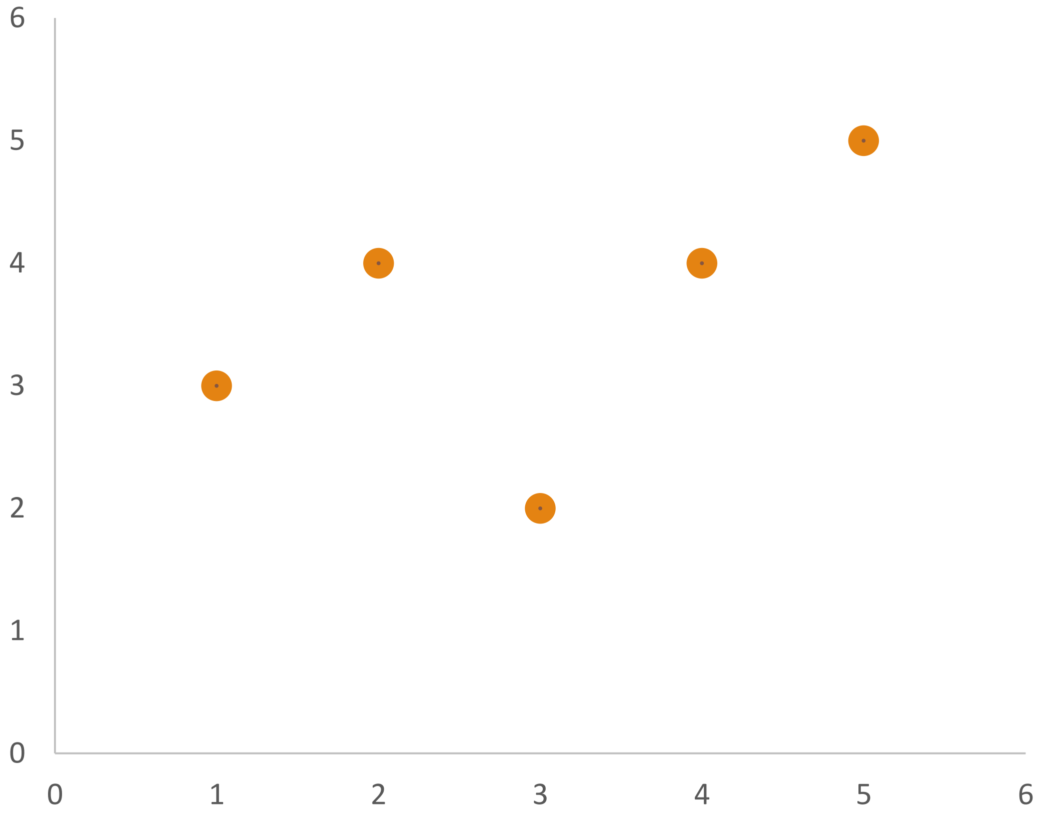
c ---> Coefficient of the line

Understanding Linear Regression Algorithm



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

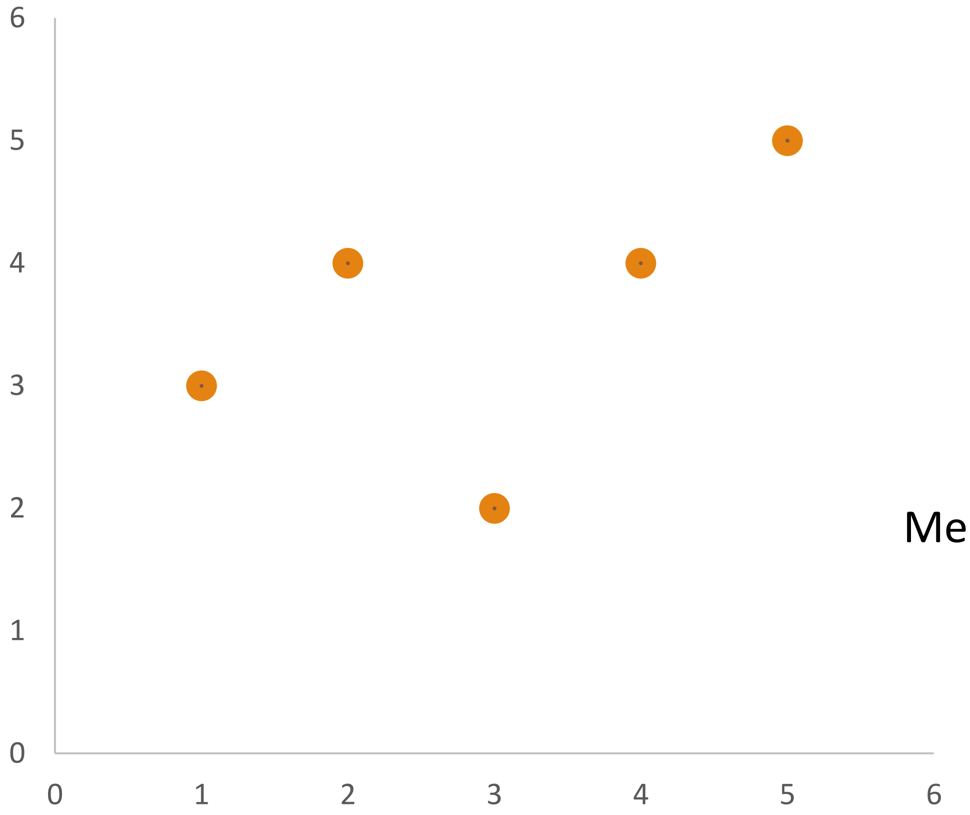
Understanding Linear Regression Algorithm



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

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

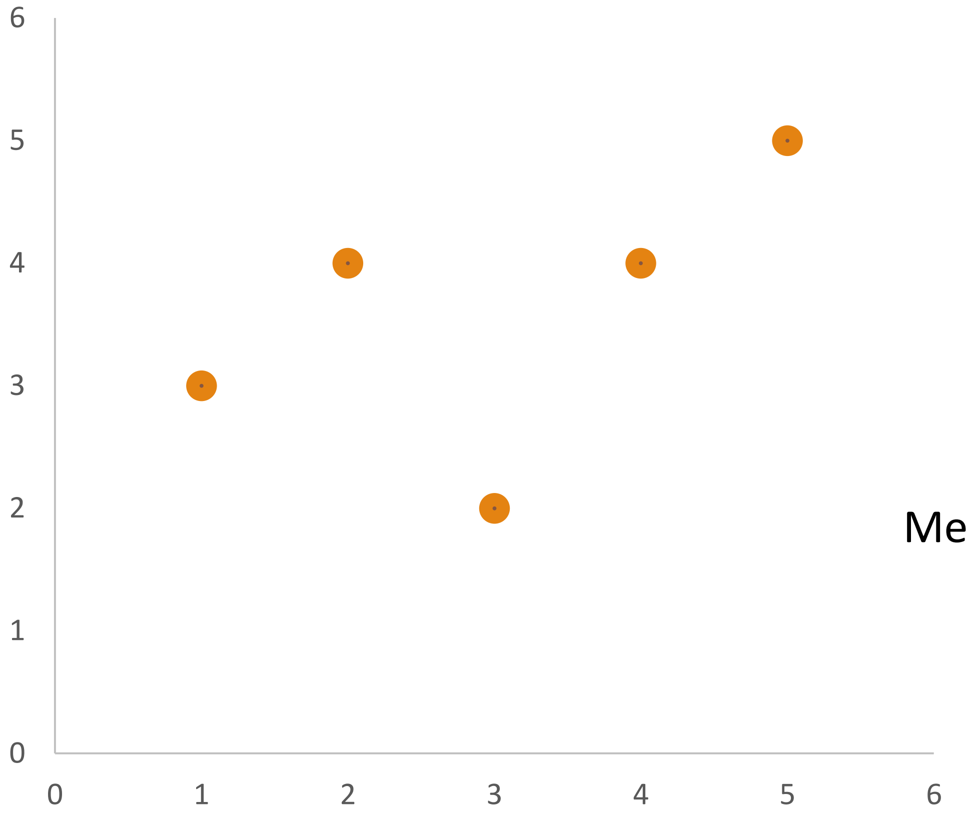
Understanding Linear Regression Algorithm



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

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

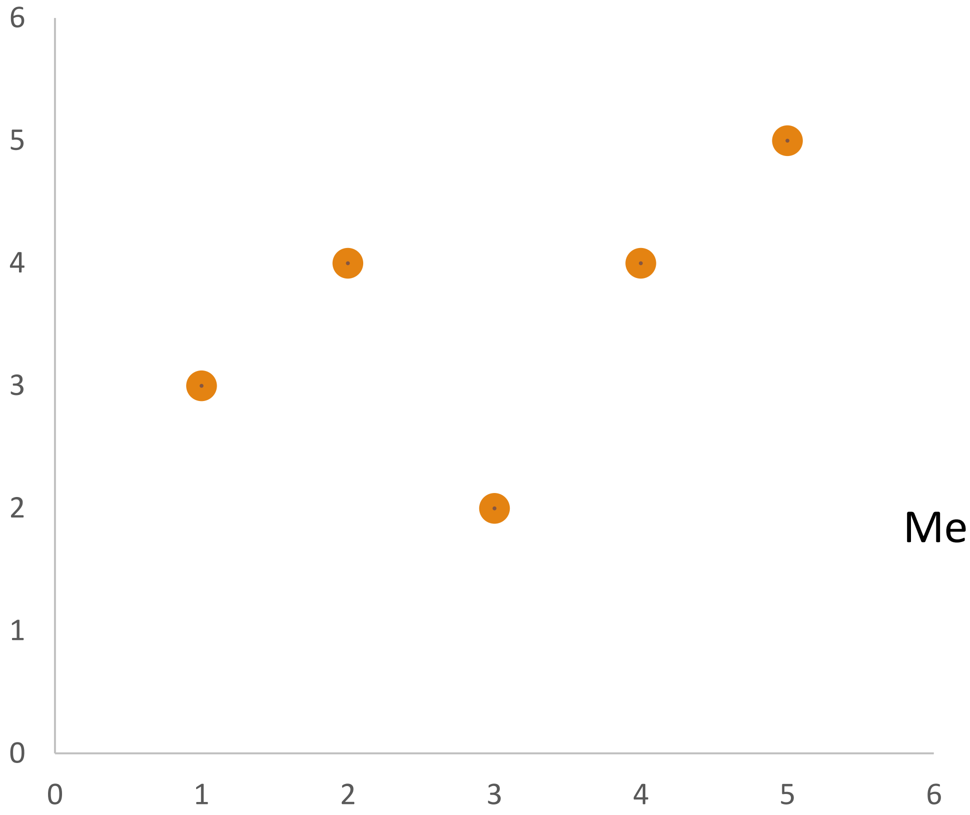
Understanding Linear Regression Algorithm



	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

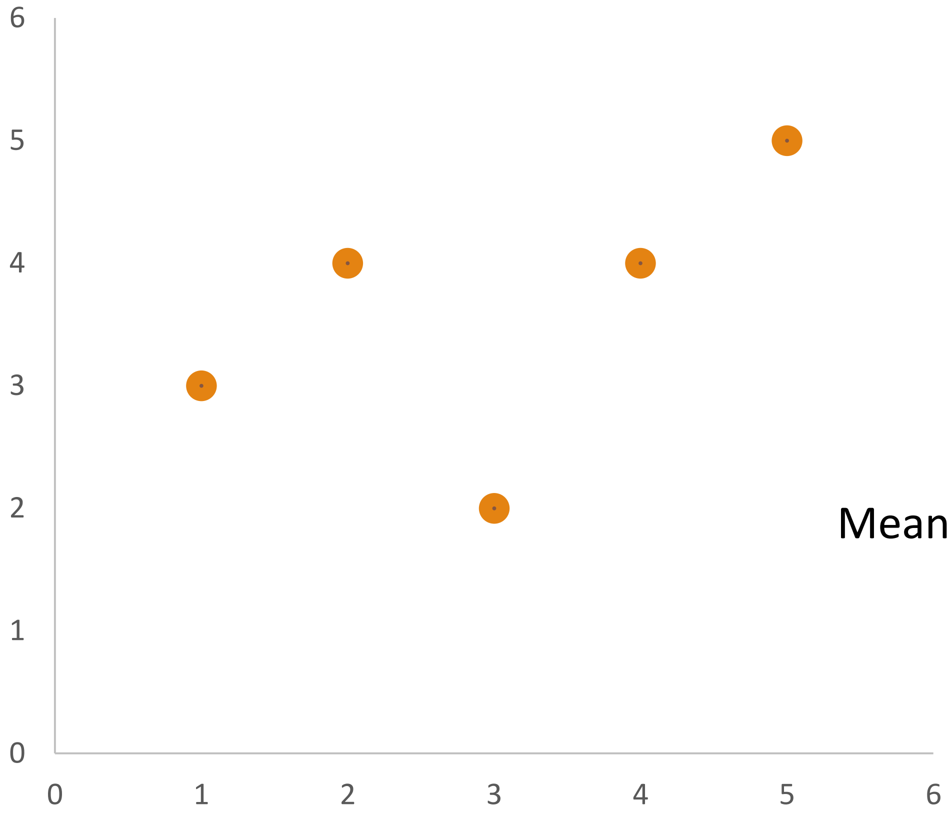


x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$
1	3	-2	-0.6	4
2	4	-1	0.4	1
3	2	0	-1.6	0
4	4	1	0.4	1
5	5	2	1.4	4

Mean 3 3.6

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

Understanding Linear Regression Algorithm

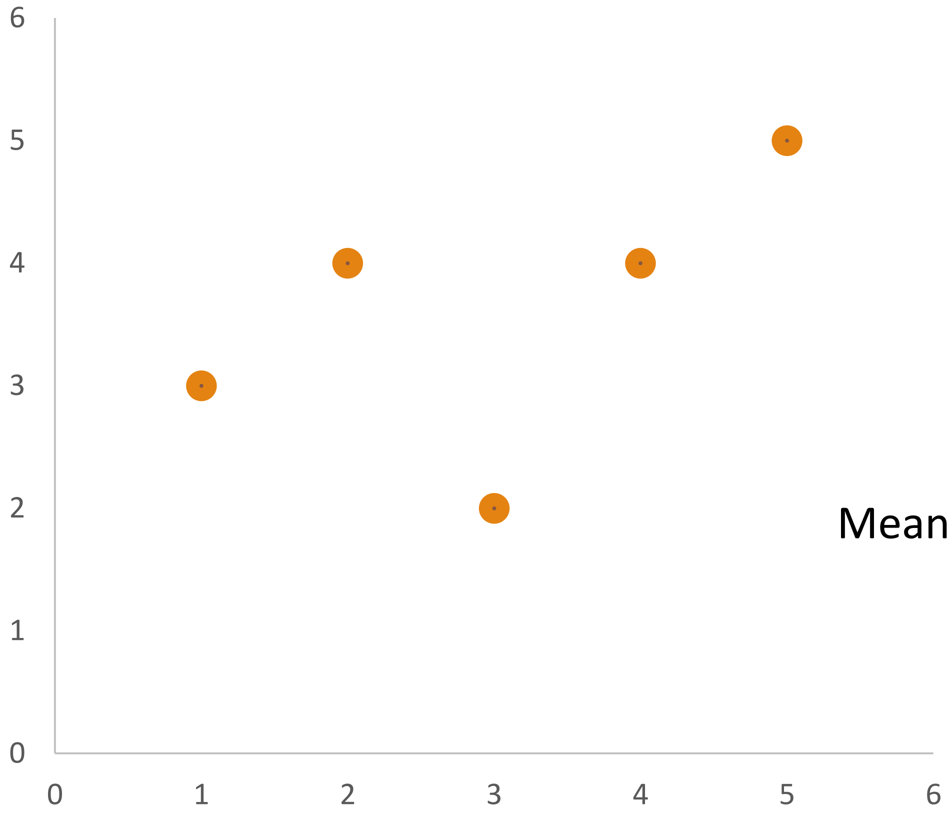


x	y	$\mathbf{x}-\bar{\boldsymbol{x}}$	$y\ -\bar{\boldsymbol{y}}$	$(\boldsymbol{x}-\bar{\boldsymbol{x}})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8

Mean 3 3.6

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

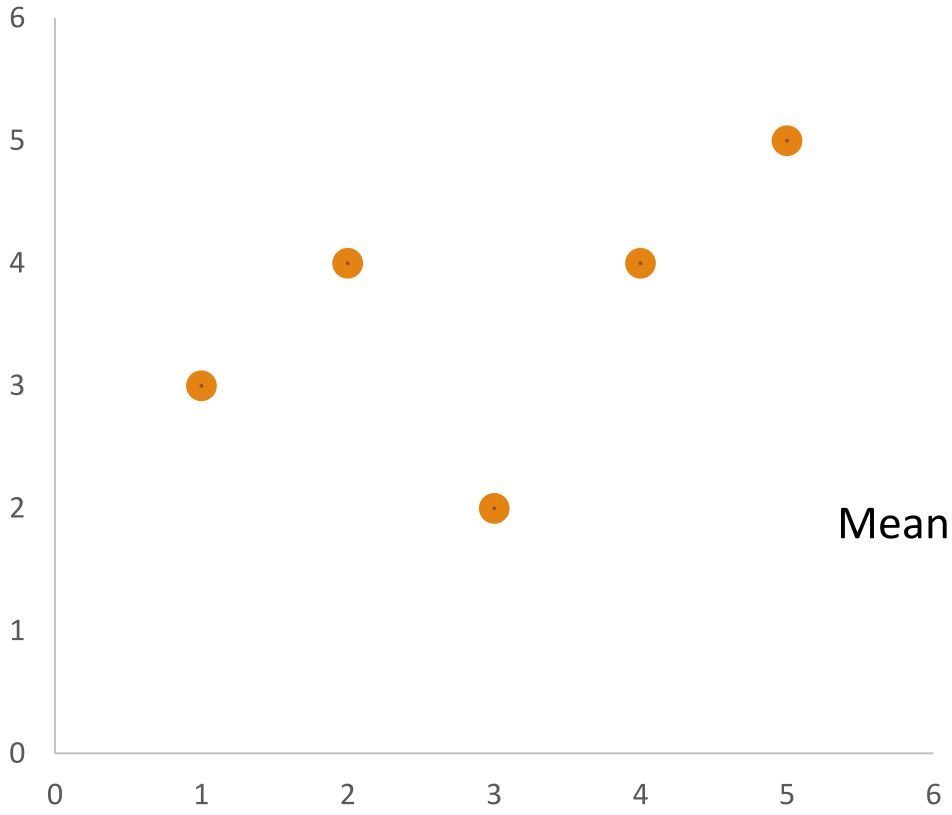
Understanding Linear Regression Algorithm



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

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

Understanding Linear Regression Algorithm



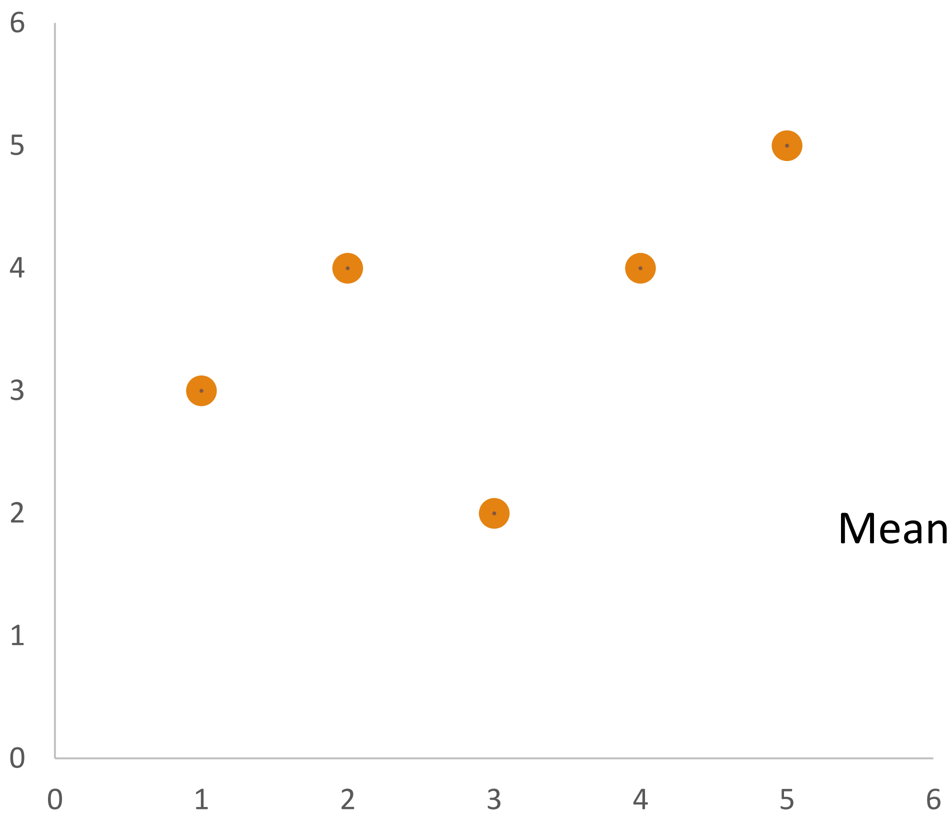
X	Y	X - $\bar{X}$	y - $\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8

Mean 3 3.6

$\sum$ = 10 $\sum$ = 4

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

Understanding Linear Regression Algorithm



X	Y	X - $\overline{X}$	y - $\overline{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
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

$\sum$ = 10 $\sum$ = 4

$$y = mx + c$$

$$3.6 = 0.4*3 + c$$

$$c = 2.4$$

Understanding Linear Regression Algorithm

$m = 0.4$

$c = 2.4$

$y = 0.4x + 2.4$

For given values of m and c, lets predict values for y for x = {1, 2, 3, 4, 5}

$y = 0.4 * 1 + 2.4 = 2.8$

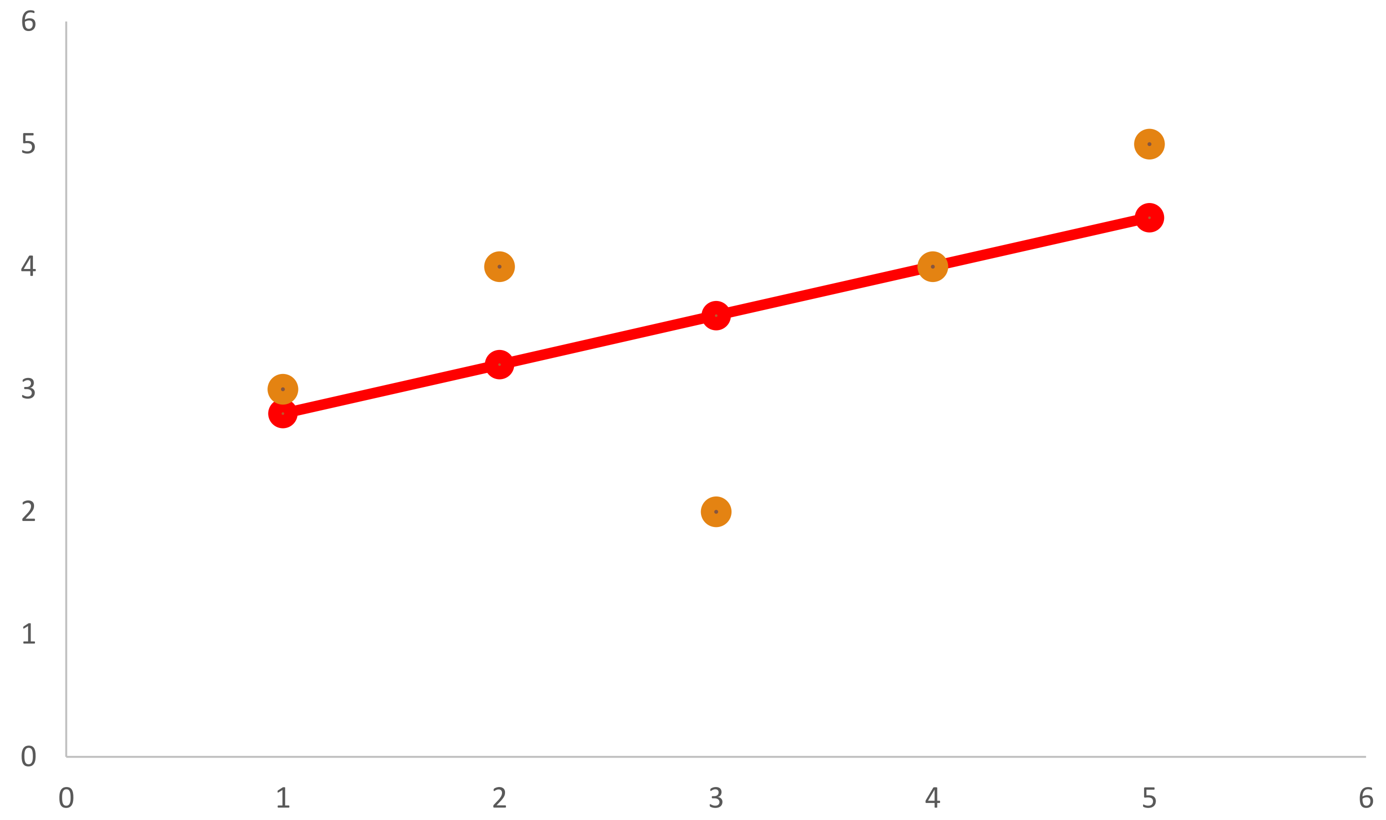
$y = 0.4 * 2 + 2.4 = 3.2$

$y = 0.4 * 3 + 2.4 = 3.6$

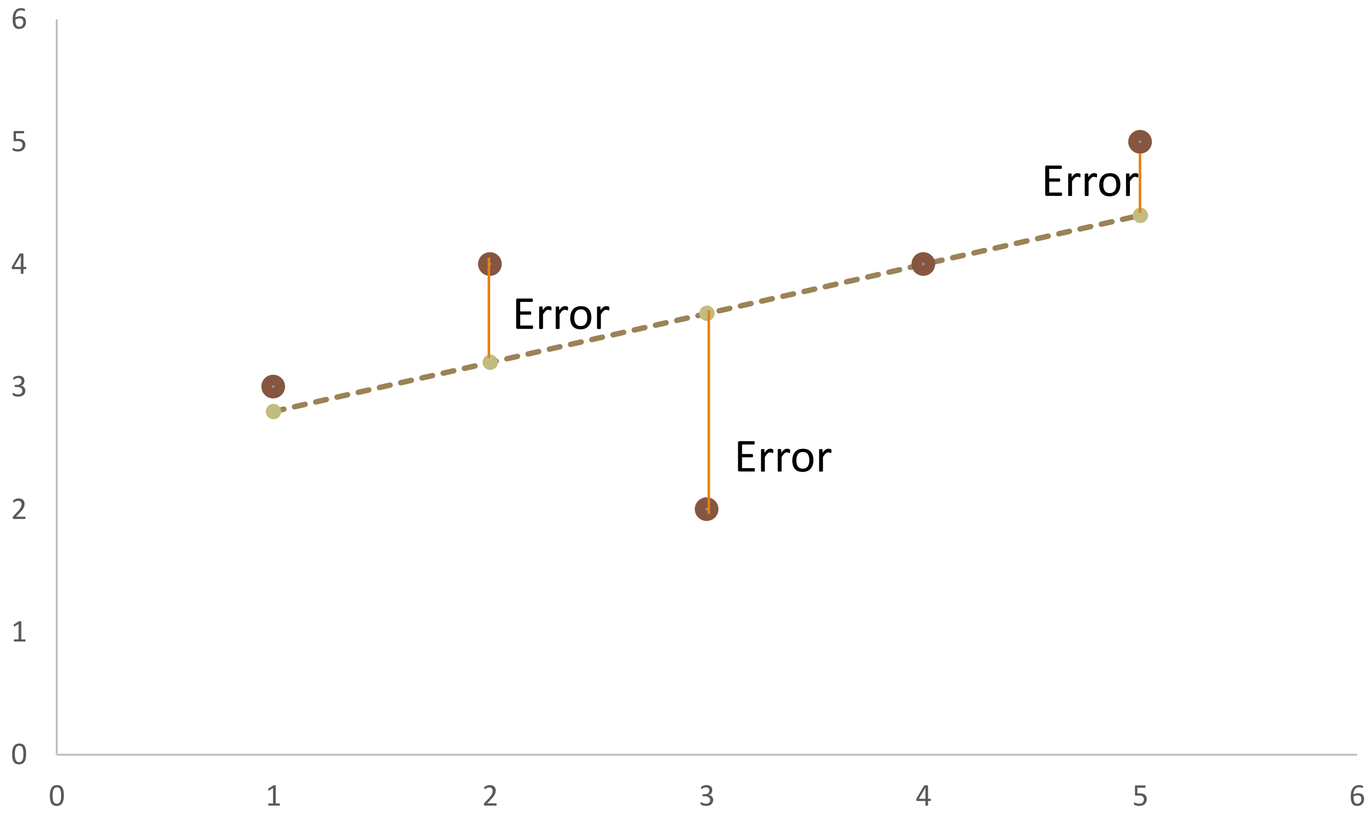
$y = 0.4 \ * 4 + 2.4 = 4.0$

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

Regression Line



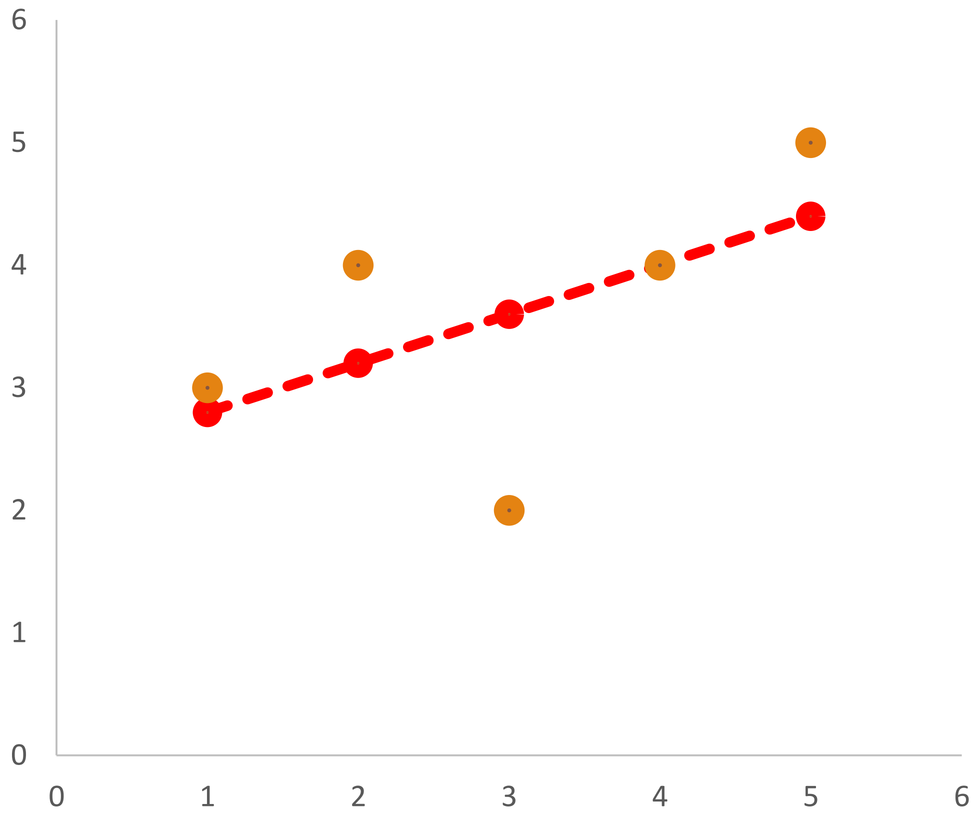
Regression Line



Goodness of Fit: 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

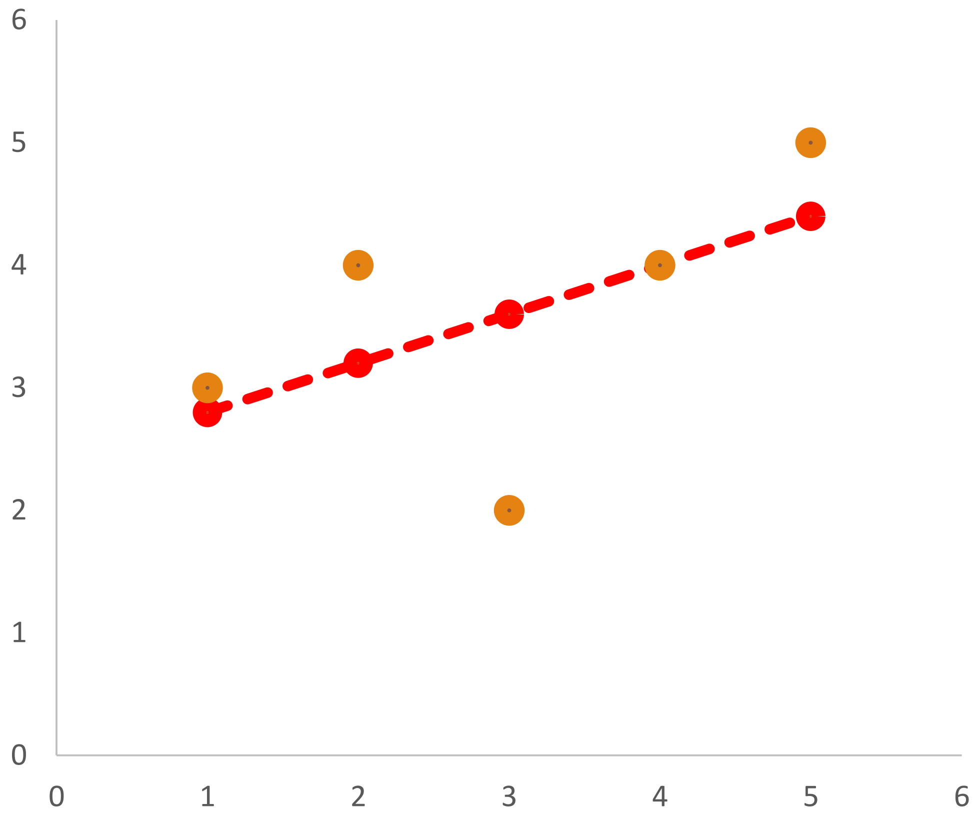
Distance actual – mean

VS

Distance predicted - mean

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

Calculation of R^2

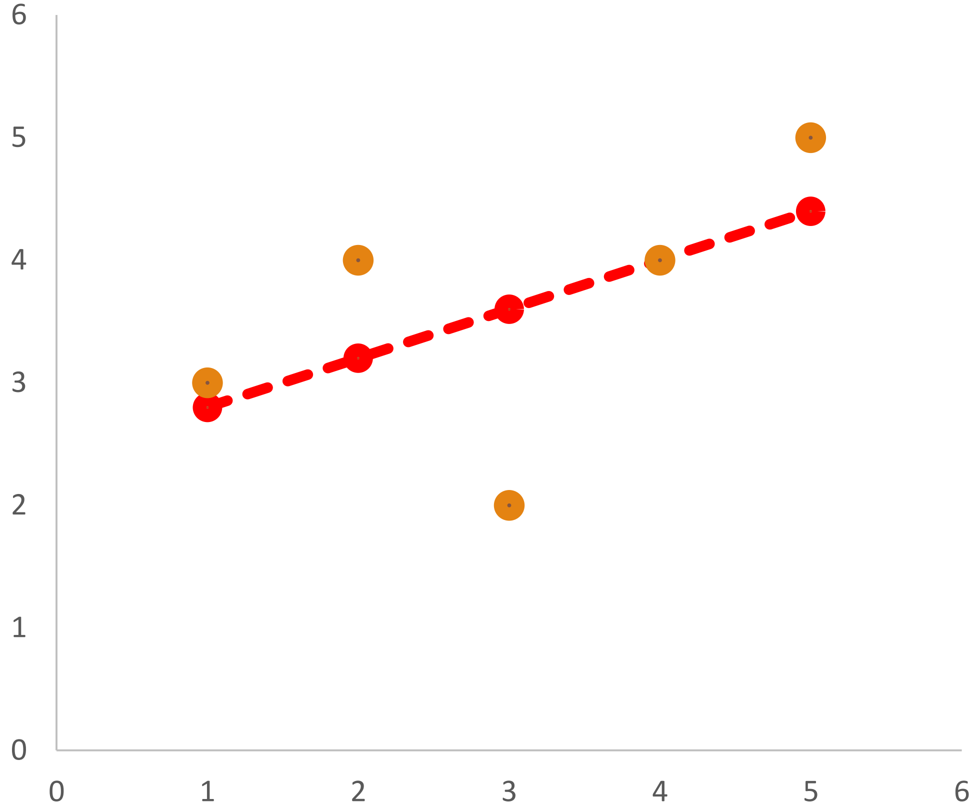


X	Y	$y - \bar{y}$
1	3	-0.6
2	4	0.4
3	2	-1.6
4	4	0.4
5	5	1.4

$\bar{y} = 3.6$

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

Calculation of R^2

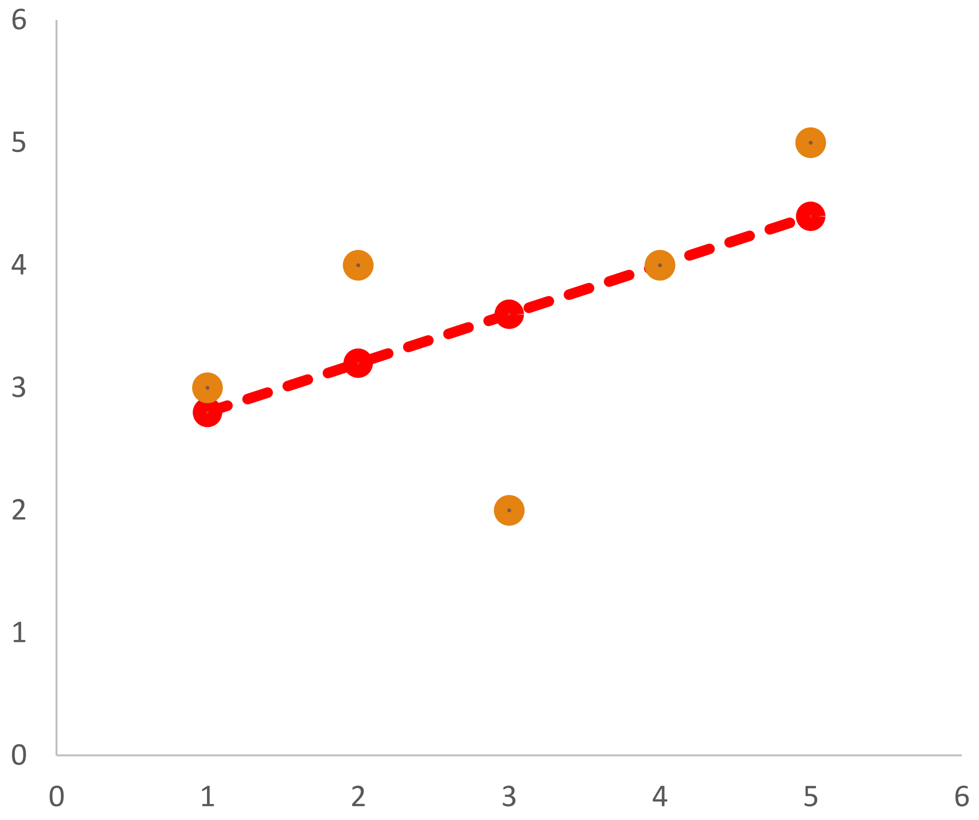


X	Y	$y - \bar{y}$	$(y - \bar{y})^2$
1	3	-0.6	0.36
2	4	0.4	0.16
3	2	-1.6	2.56
4	4	0.4	0.16
5	5	1.4	1.96

$\bar{y} = 3.6$

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

Calculation of R^2

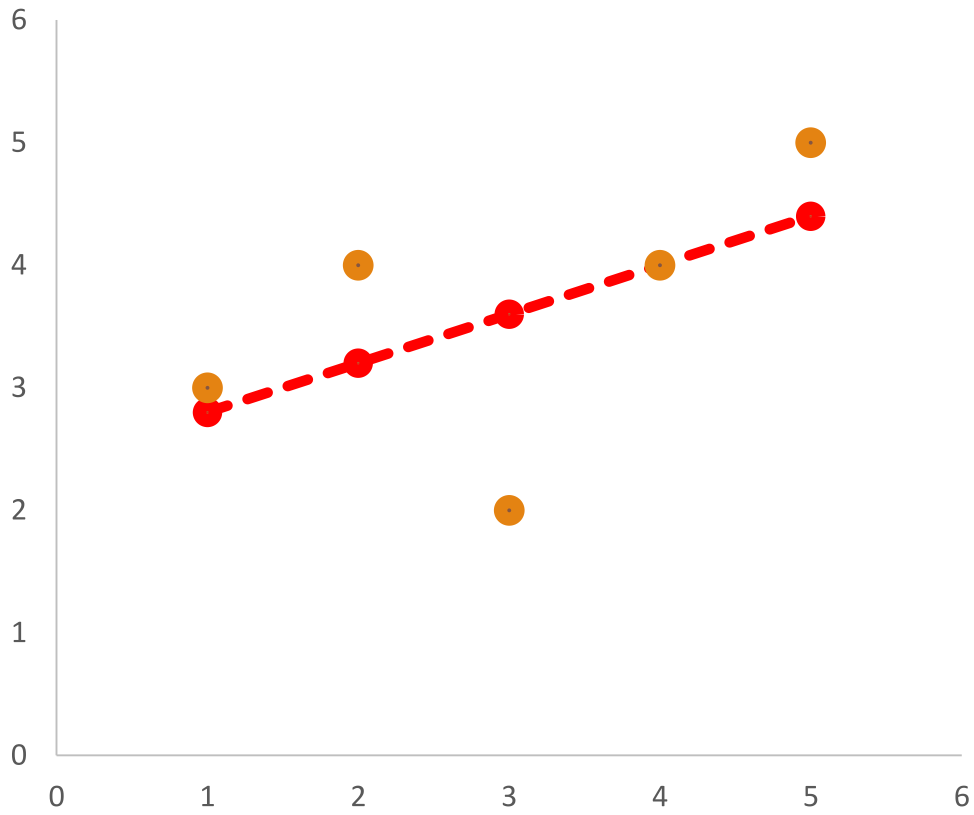


X	Y	$y-\bar{y}$	$(y-\bar{y})^2$	Y_p
1	3	-0.6	0.36	2.8
2	4	0.4	0.16	3.2
3	2	-1.6	2.56	3.6
4	4	0.4	0.16	4.0
5	5	1.4	1.96	4.4

$\bar{y} = 3.6$

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

Calculation of R^2

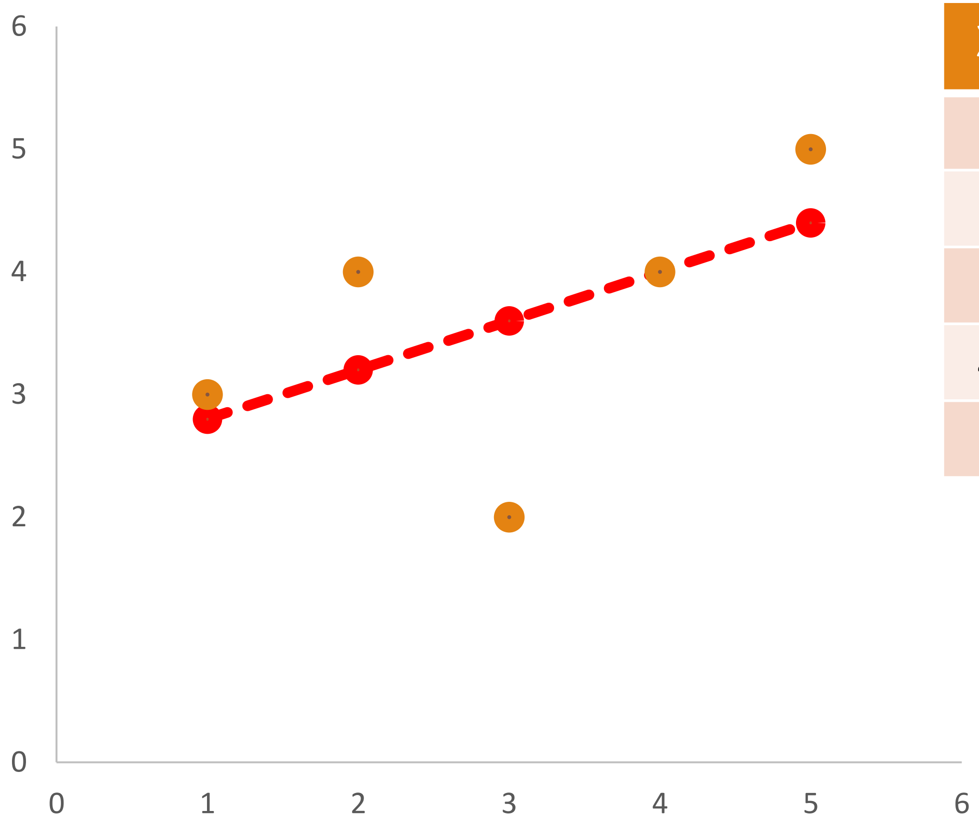


X	Y	$y - \bar{y}$	$(y - \bar{y})^2$	Y_p	$y_p - \bar{y}$
1	3	-0.6	0.36	2.8	-0.8
2	4	0.4	0.16	3.2	-0.4
3	2	-1.6	2.56	3.6	0
4	4	0.4	0.16	4.0	0.4
5	5	1.4	1.96	4.4	0.8

$\bar{y} = 3.6$

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

Calculation of R^2

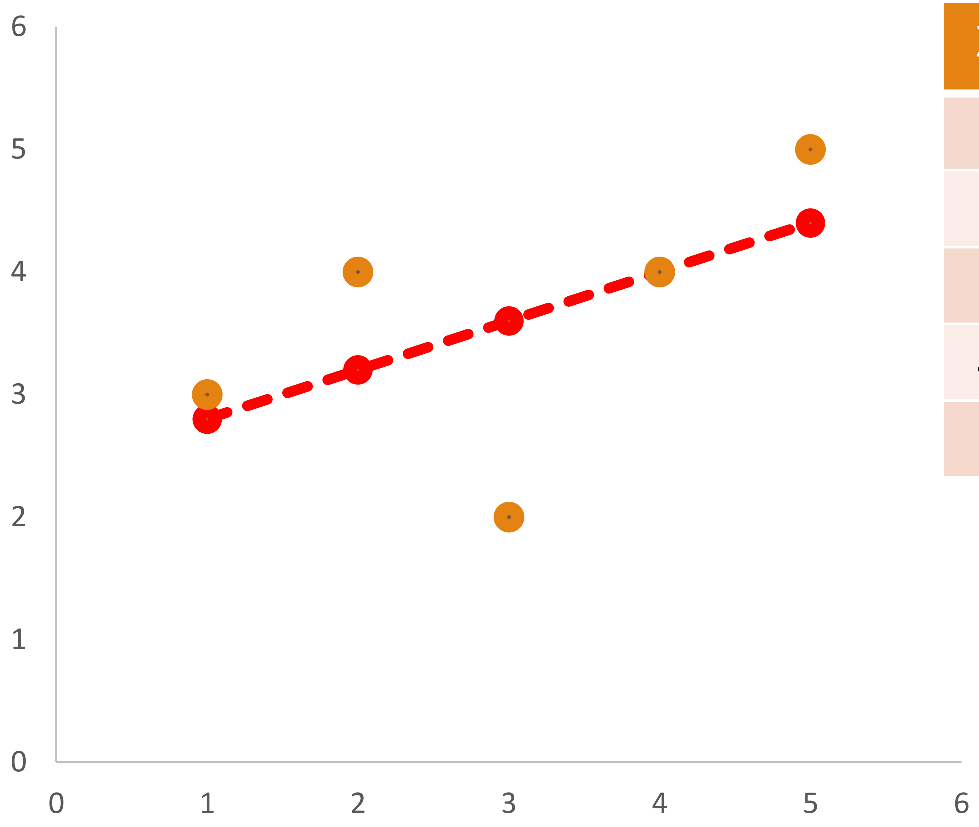


X	Y	$y - \bar{y}$	$(y - \bar{y})^2$	$\boldsymbol{Y_p}$	$y_p - \bar{y}$	$(y_p - \bar{y})^2$
1	3	-0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	-1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64

$\bar{y} = 3.6$

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

Calculation of R^2

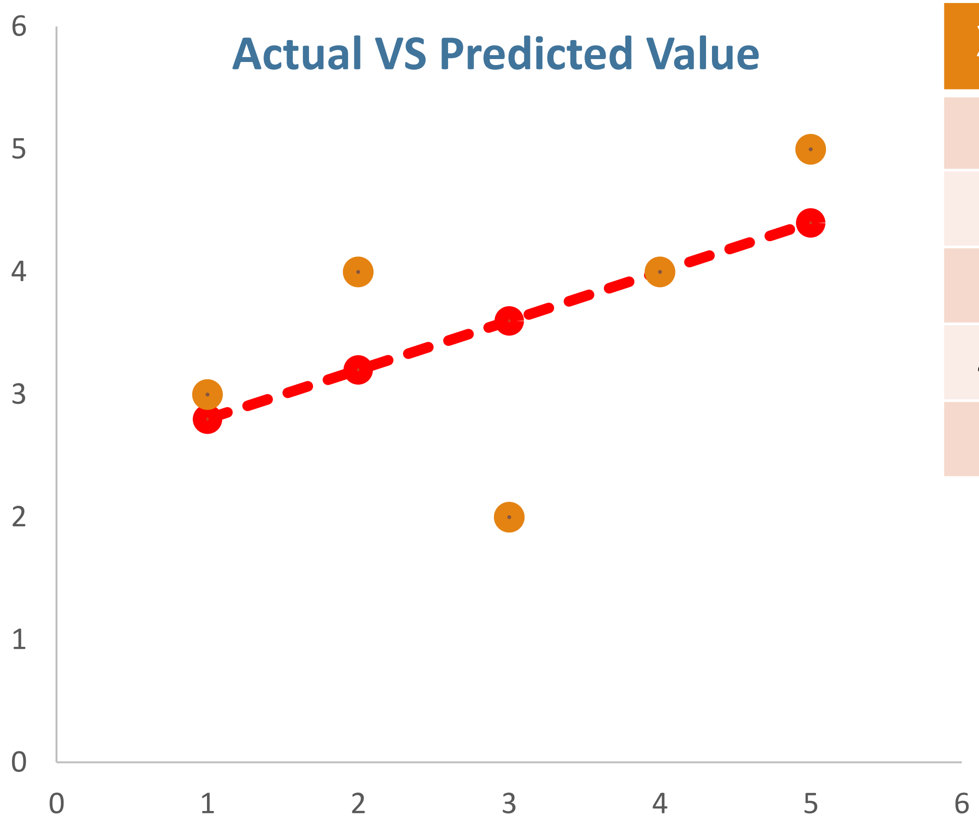


X	Y	$y - \bar{y}$	$(y - \bar{y})^2$	Y_p	$y_p - \bar{y}$	$(y_p - \bar{y})^2$
1	3	-0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	-1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64

$\bar{y} = 3.6$ $\sum = 5.2$ $\sum = 1.6$

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

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

$\bar{y} = 3.6$ $\sum = 5.2$ $\sum = 1.6$

$$R^2 = \frac{\sum(y_p - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{1.6}{5.2} \approx \mathbf{0.3}$$

Multiple Linear Regression

Multiple or multivariate linear regression is a case of linear regression with two or more independent variables.

Multiple Linear Regression

If there are just two independent variables, the estimated regression function is:

$$f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$$

It represents a regression plane in a three-dimensional space. The goal of regression is to determine the values of the weights b_0, b_1, and b_2 such that this plane is as close as possible to the actual responses and yield the minimal sum of squared residuals (SSR).

Multiple Linear Regression

The case of more than two independent variables is similar, but more general. The estimated regression function is:

$$f(x_1, \ldots, x_n) = b_0 + b_1x_1 + \cdots + b_nx_n,$$

There are n + 1 weights to be determined when the number of inputs is n.

Multiple Linear Regression Steps

1. Load the dataset
2. Split dataset into training set and test set
3. Fit regression model to training set
4. Predict the test set
5. Evaluate the goodness of fit