Linear Regression

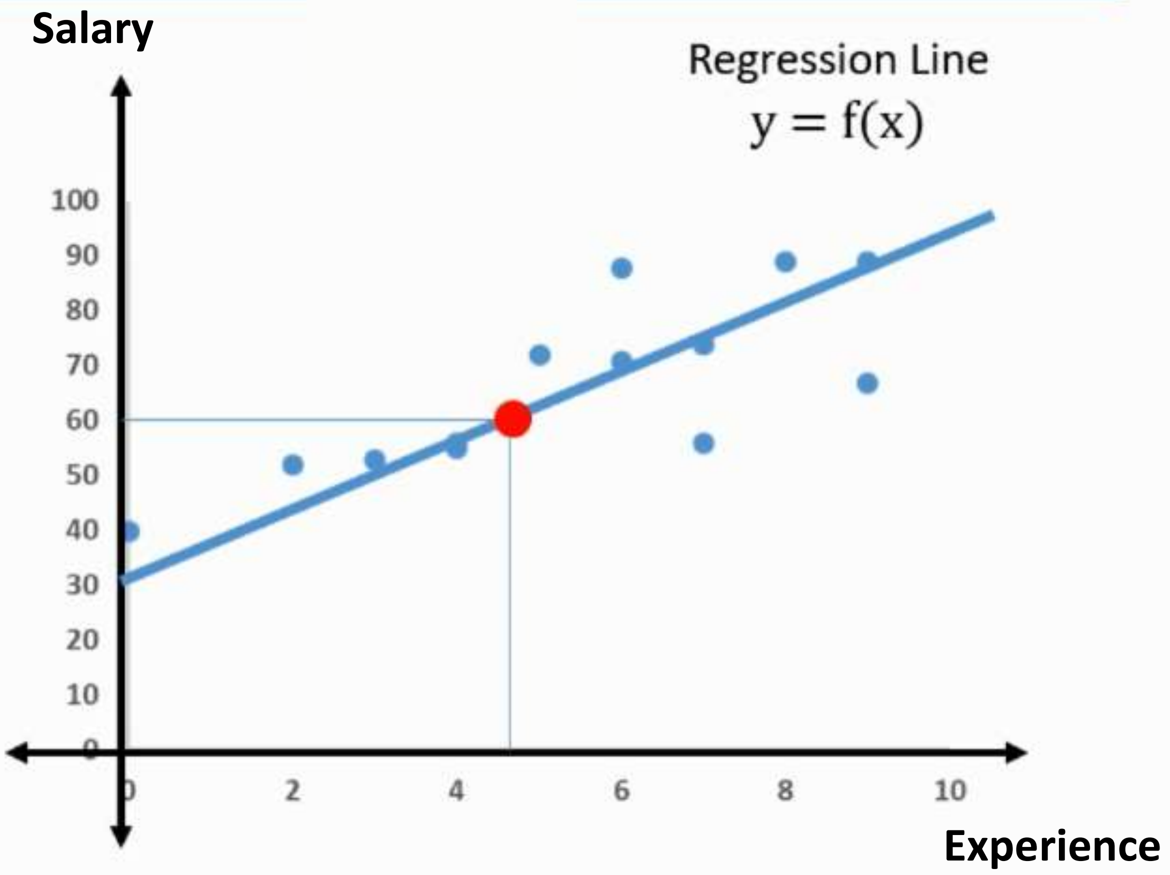
By Eng. Mohammed Marwan Shahin

Outlines

- Regression Algorithms Theories:
  - Simple Linear Regression
  - Multiple Linear Regression
  - Polynomial Regression
  - Decision Tree Regression
  - Random Forest Regression
- Building Regression models using (Scikit-learn) Library.
- Selecting best Model
- Creating Model Templates for each Regression algorithm

Regression Analysis

- Statistical process for estimating the relationships among variables
- The predictor is a continuous variable
- Relationship between a dependent variable and one or more independent variables (or 'predictors')
- Can also be used to infer causal relationships between dependent and independent variables.



Linear Regression

Univariate Linear Regression

$$y = m_1 x_1 + c$$

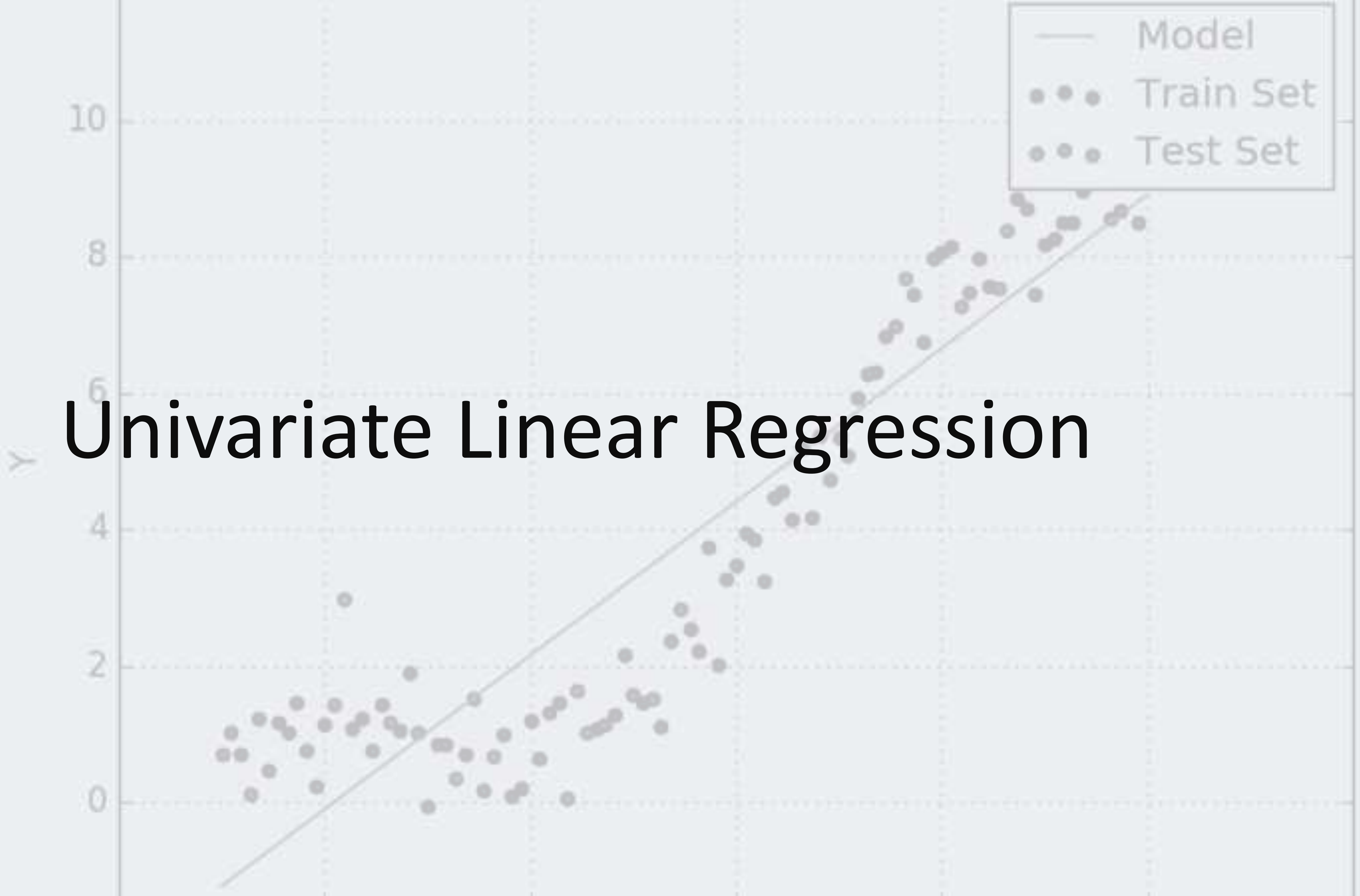
Multiple Linear Regression

$$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + \ldots + m_n x_n + c$$

Polynomial Linear Regression

$$y = m_1 x_1 + m_2 x_1^2 + m_3 x_1^3 + \ldots + m_n x_1^n + c$$

Univariate Linear Regression

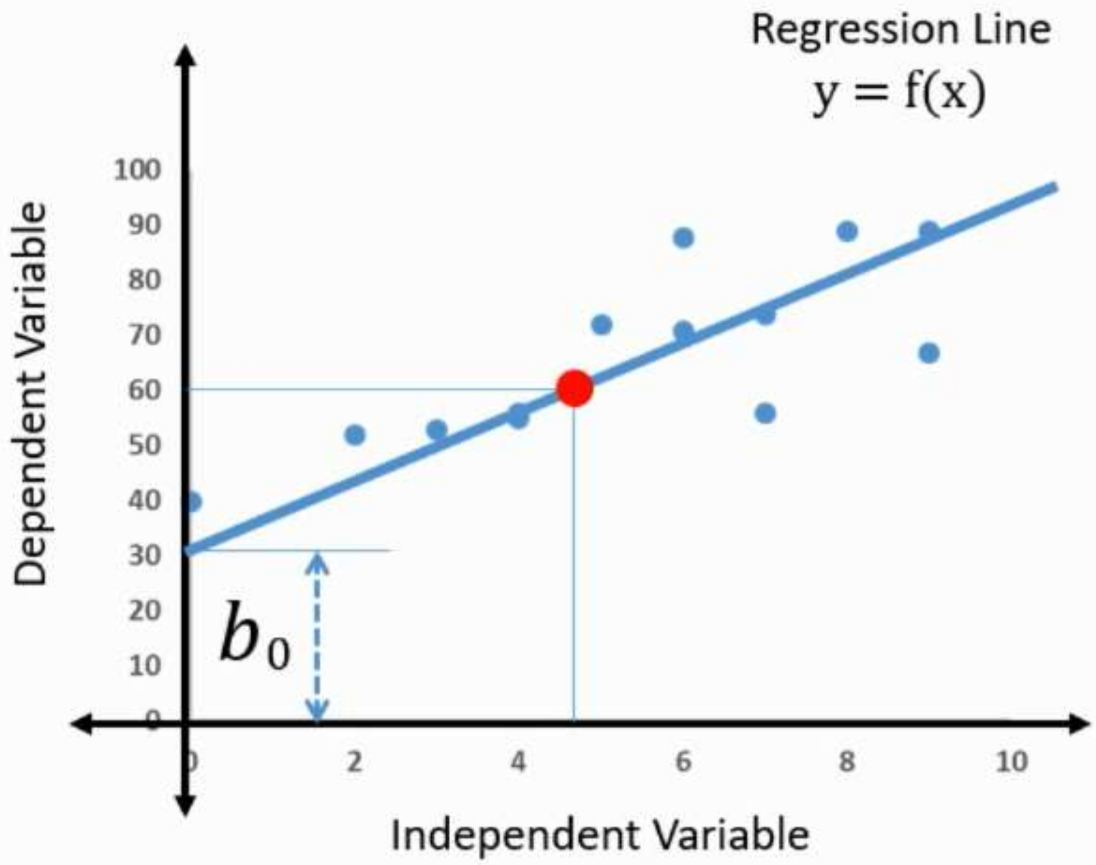


Simple Linear Regression

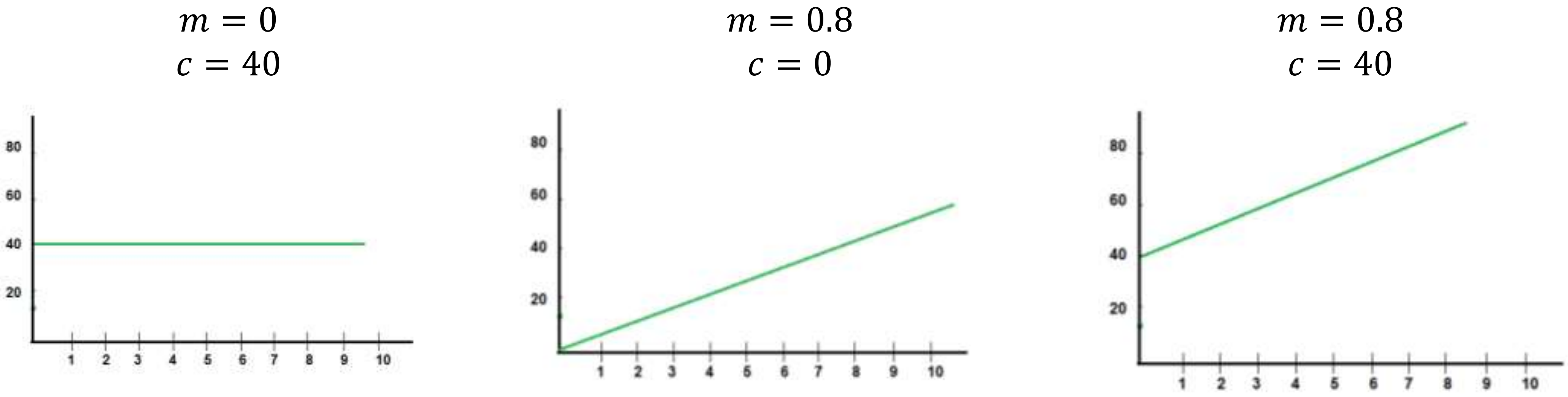
Simple Regression :

$$y = b_0 + \ b_1 x$$

Only one Dependent
Only one Independent

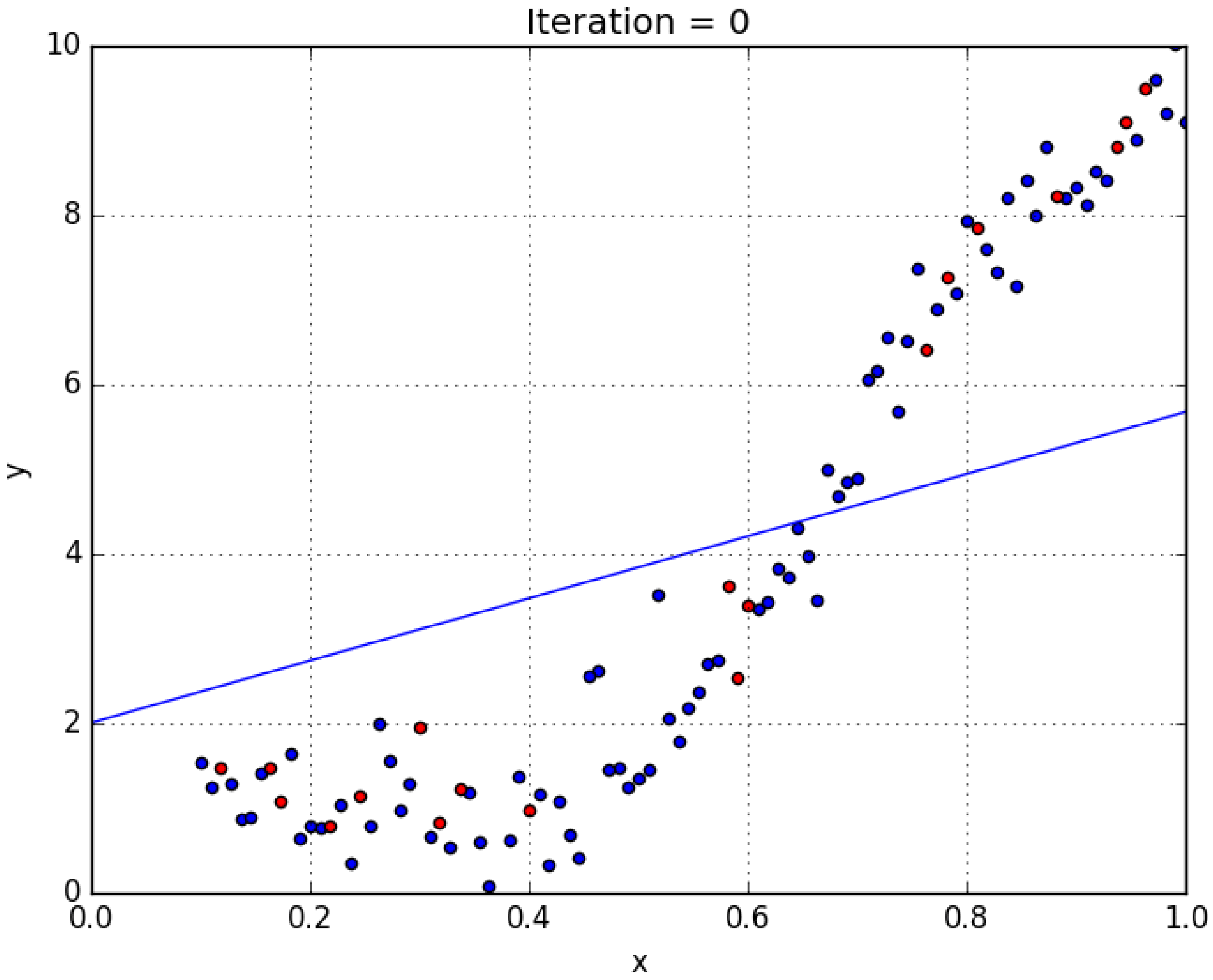


Equation of a straight line



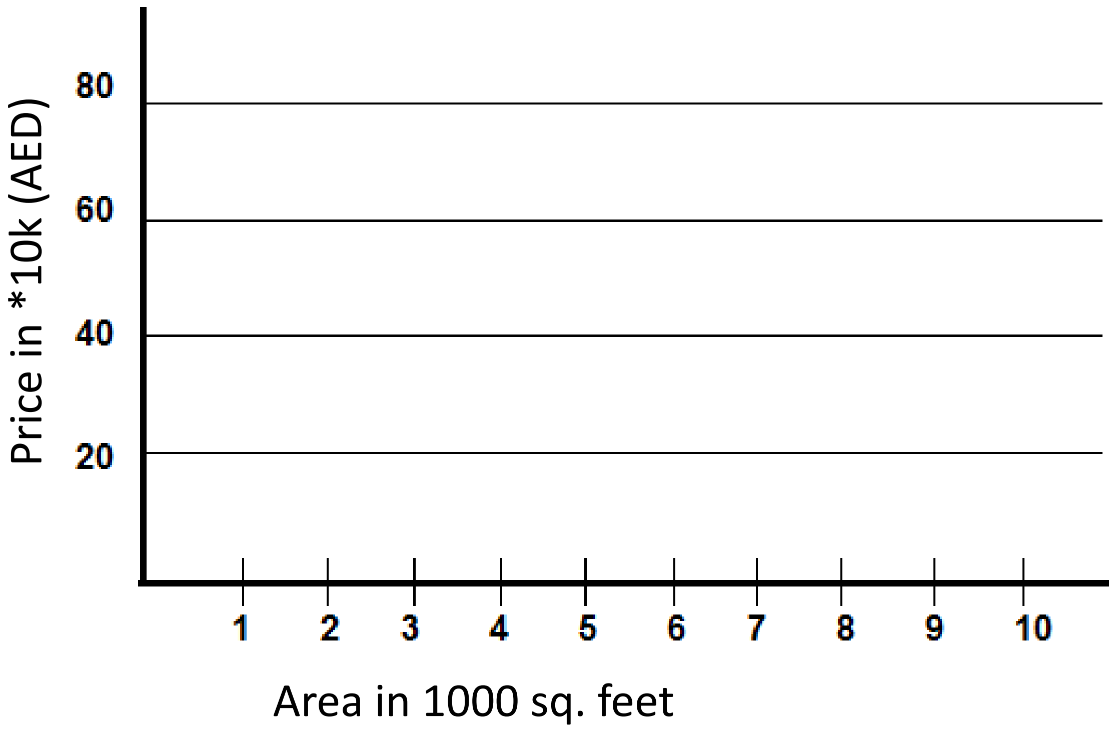
$$\hat{y} = mx + c$$

- During the training period the regression line is getting more fit.



Housing Prices Prediction

Area (sq ft)	Price In AED
1200	200,000
1800	420,000
3200	440,000
3800	250,000
4200	620,000

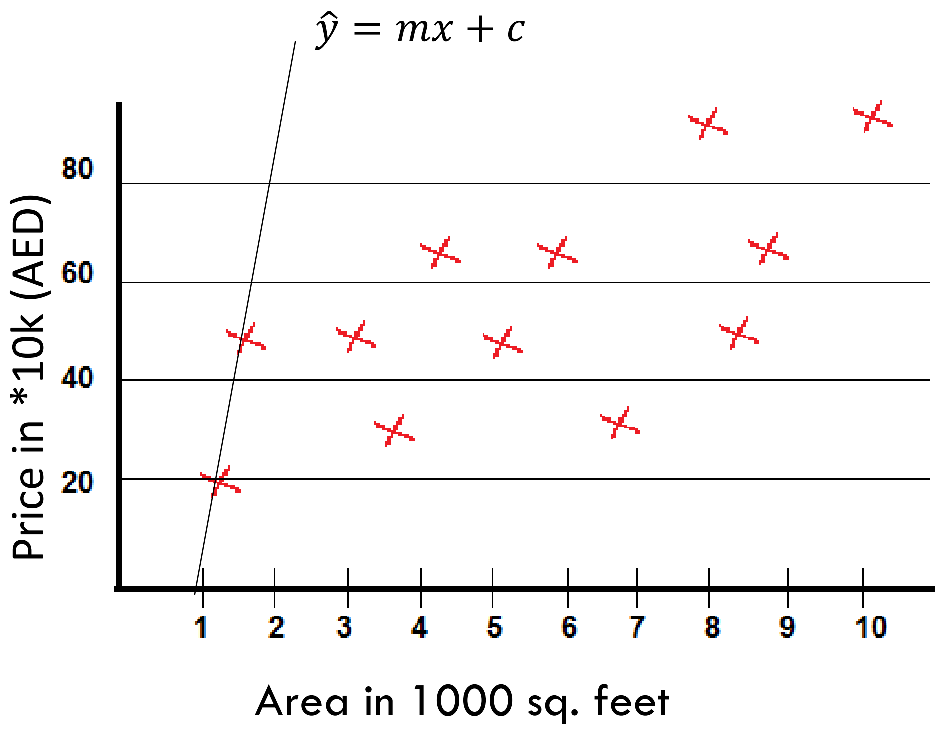


Housing Prices Prediction

Area (sq ft)	Price In AED
1200	200,000
1800	420,000
3200	440,000
3800	250,000
4200	620,000

y: Dependent Variable, criterion variable, or regressand.

x: Independent variable, predictor variables or regressors.



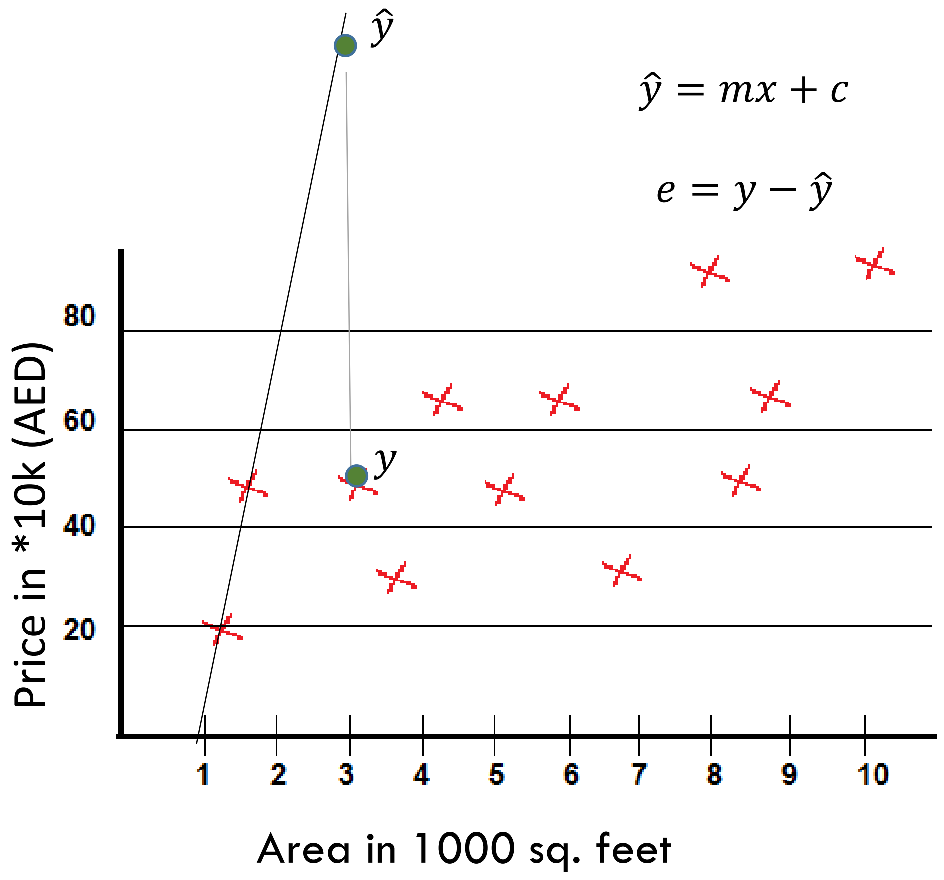
Housing Prices Prediction

Linear Regression in one Variable

Area (sq ft)	Price In AED
1200	200,000
1800	420,000
3200	440,000
3800	250,000
4200	620,000

$$\hat{y} = mx + c$$

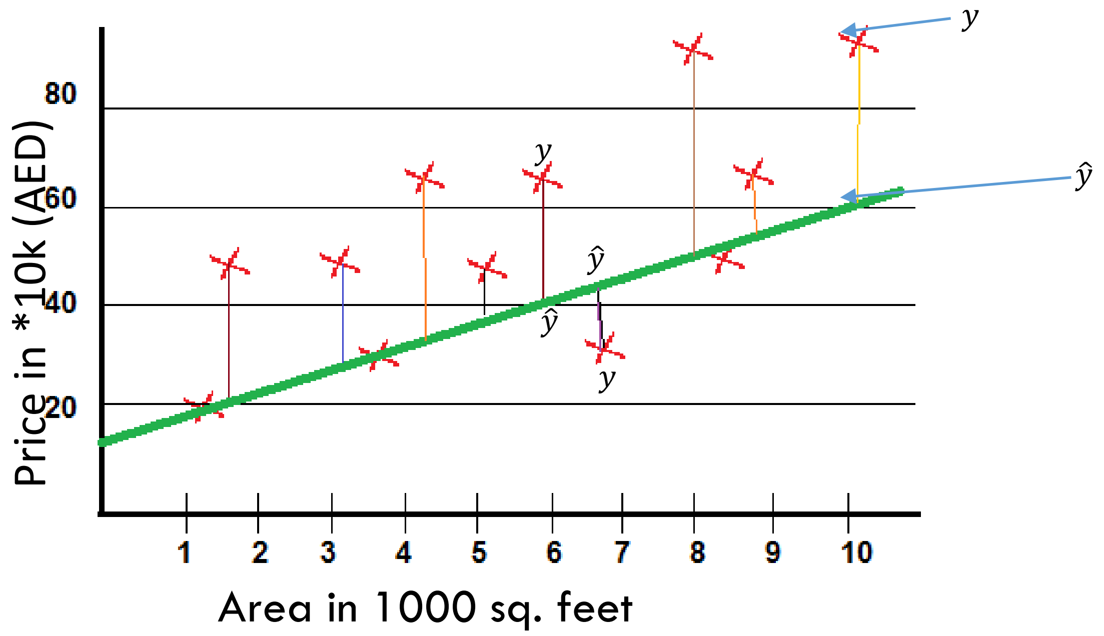
$\hat{y} = Value\ predicted\ by\ current\ Algorithm$



Housing Prices Prediction

$$minimize$$
$$(y - \hat{y})$$

$$Predictor$$
$$\hat{y} = mx + c$$



Housing Prices Prediction

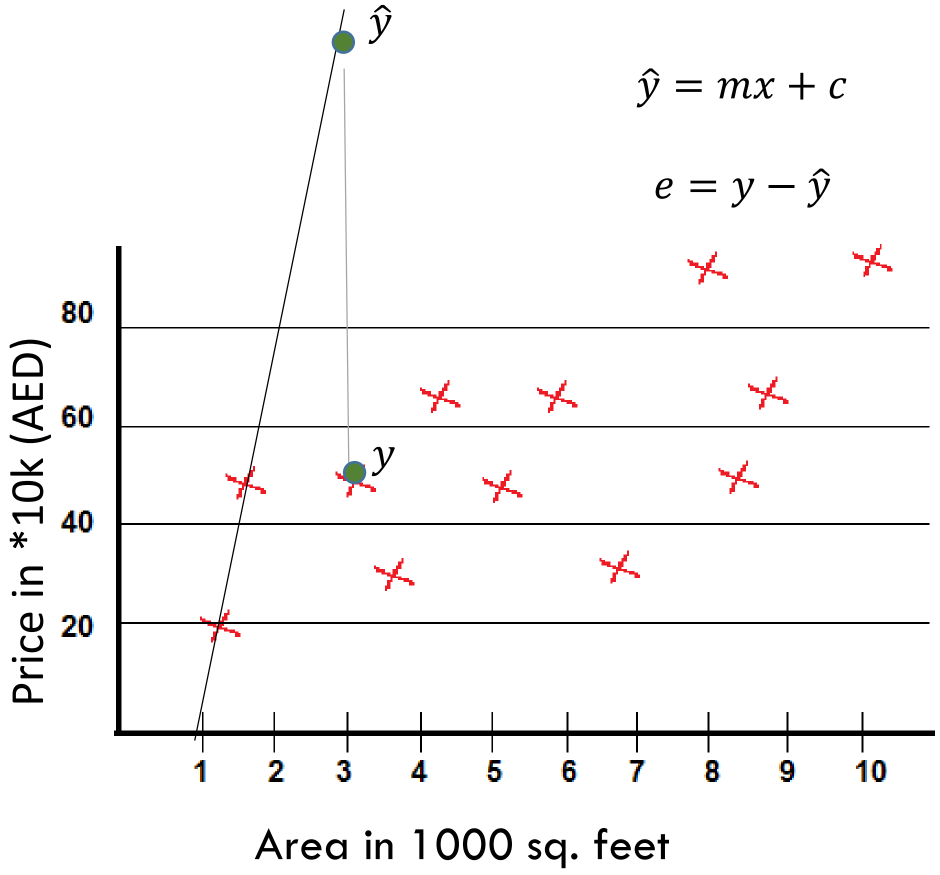
Cost Function

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

$$j(m_i, c) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Predictor

$$\hat{y} = mx + c$$



Regression Equation:

$$\hat{y} = mx + c$$

Parameters

$$m_i, c$$

Cost Function:

$$j(m_i, c\ ) = \frac{1}{2n}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \widehat{\boldsymbol{y}}_{\boldsymbol{i}})^{\boldsymbol{2}}$$

Goal

$$\underset{m_i, c}{minimize}\, J(m_i, c)$$

Objective of Linear Regression

- Establish If there is a relationship between two variables.

  Examples – relationship between housing process and area of house, no of hours of study and the marks obtained, income and spending etc.

- Prediction of new possible values

  Based on the area of house predicting the house prices in a particular month; based on number of hour studied predicting the possible marks. Sales in next 3months etc.

LINEAR REGRESSION USE CASES

Real Estate

- To model residential home prices as a function of the home's living area, bathrooms, number of bedrooms, lot size.

Medicine

- To analyze the effect of a proposed radiation treatment on reducing tumor sizes based on patient attributes such as age or weight.

Demand Forecasting

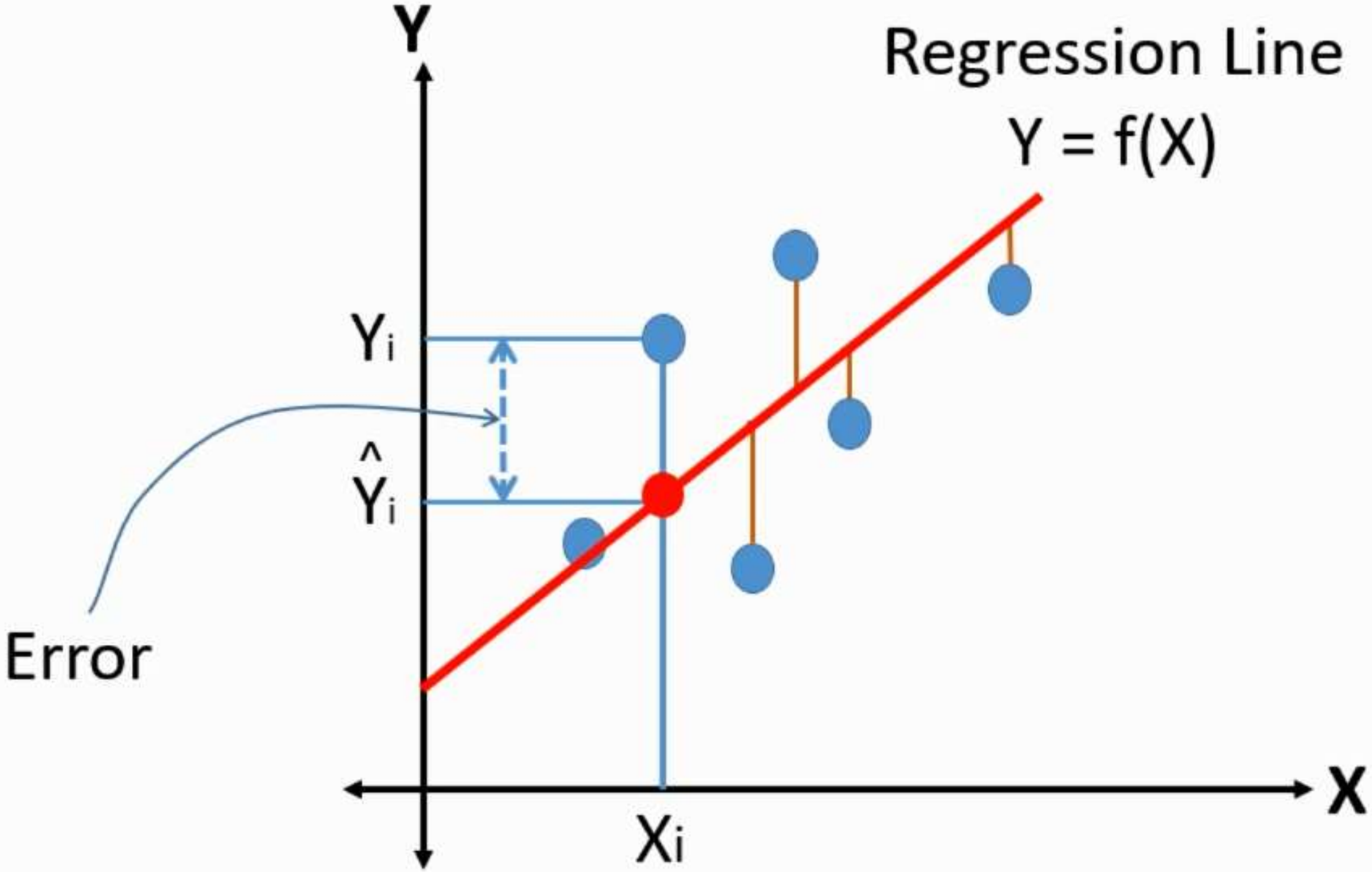
- To predict demand for goods and services. For example, restaurant chains can predict the quantity of food depending on weather.

Marketing

- To predict company’s sales based on previous month’s sales and stock prices of a company.

Regression Cost Function

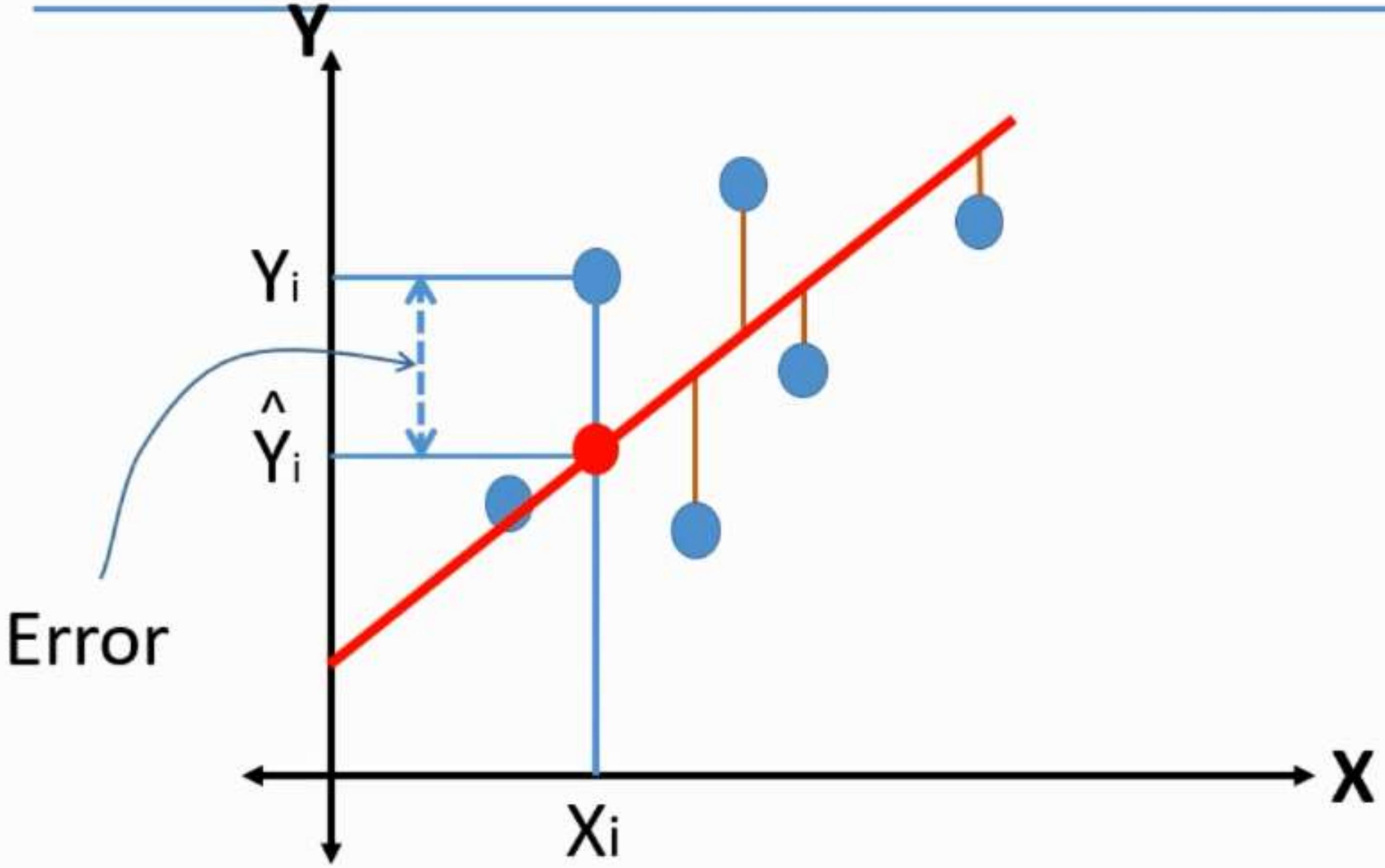
Ordinary Least Square



Minimum

$$\sum_{i=1}^{n}(yi - \hat{y}i)^2$$

Mean Absolute Error



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

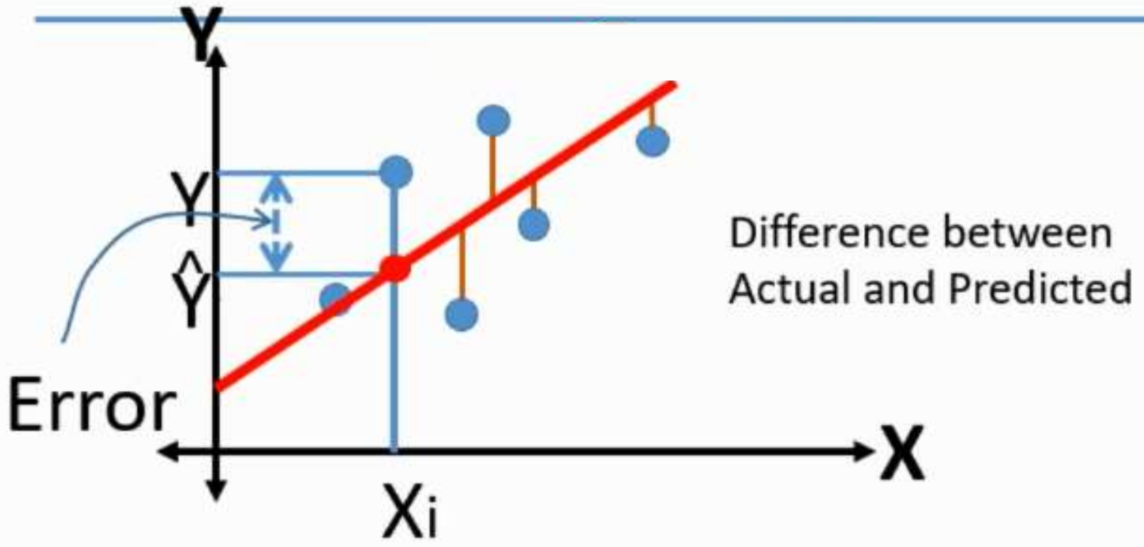
Mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes.

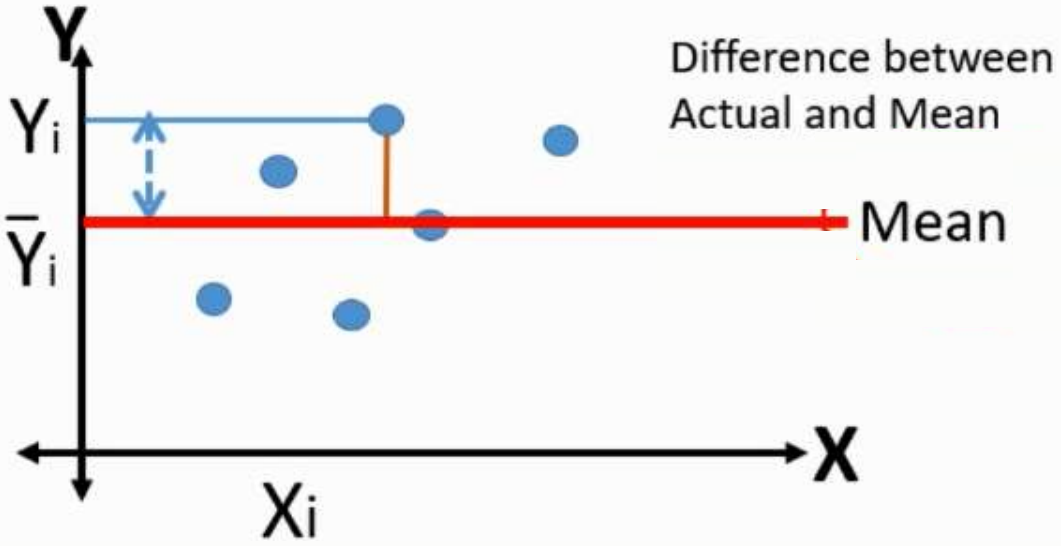
Root Mean Squared Error

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

- Very commonly used and makes for an excellent general purpose error metric for numerical predictions.
- Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors.

Relative Absolute Error





$$RAE = \frac{\sum_{i=1}^{n} |yi - \hat{y}i|}{\sum_{i=1}^{n} |yi - \bar{y}i|}$$

Demo 1: Create Linear Regression model for Predicting the salary based on the Experience.

Task: Create ML model to predict the Student test result based on Studying hours.

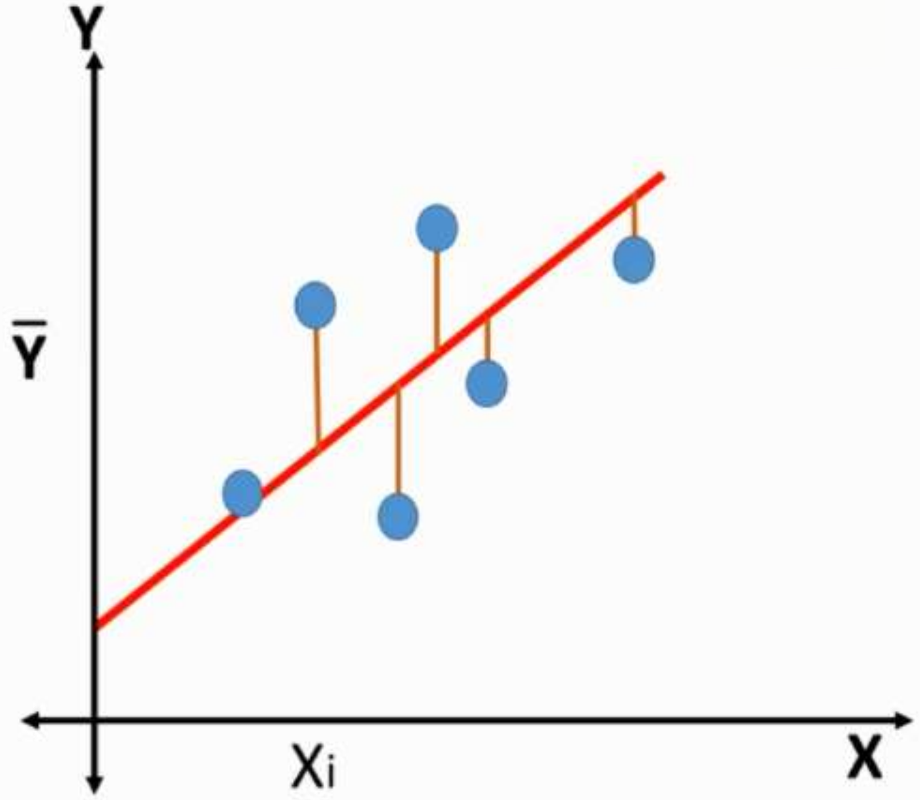
Is it a good prediction?

	0
0	49.3537
1	49.3537
2	39.2995
3	39.2995
4	84.5434
5	49.3537
6	39.2995
7	74.4892
8	59.4079

```
from sklearn.l   ar_model import LinearRegression

# Create
std_reg =          ession()

# Train the        or object on training data
std_reg.fit(X    ain, Y_train)

# pred         d   a
Y_predic  =        g.predict(X_test)
```

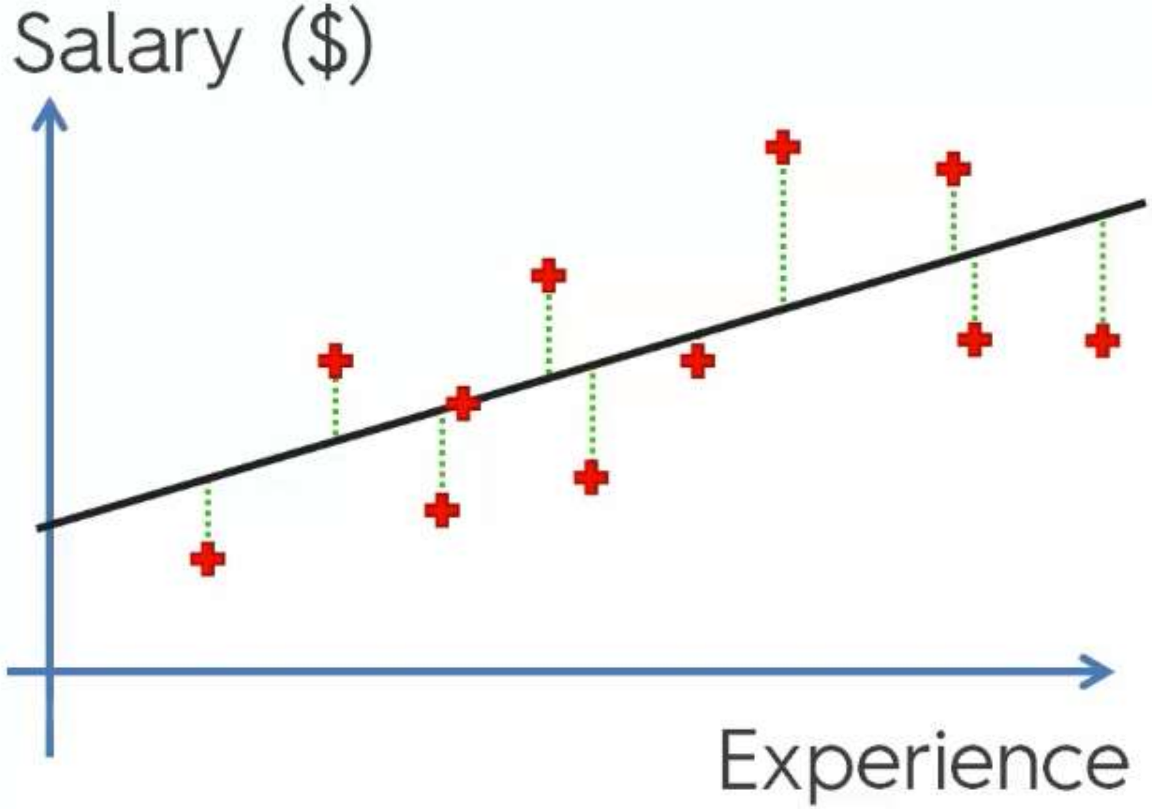
Coefficient of Determination

How much (what %) of variation in Y is described by the variation in X?



R-Square

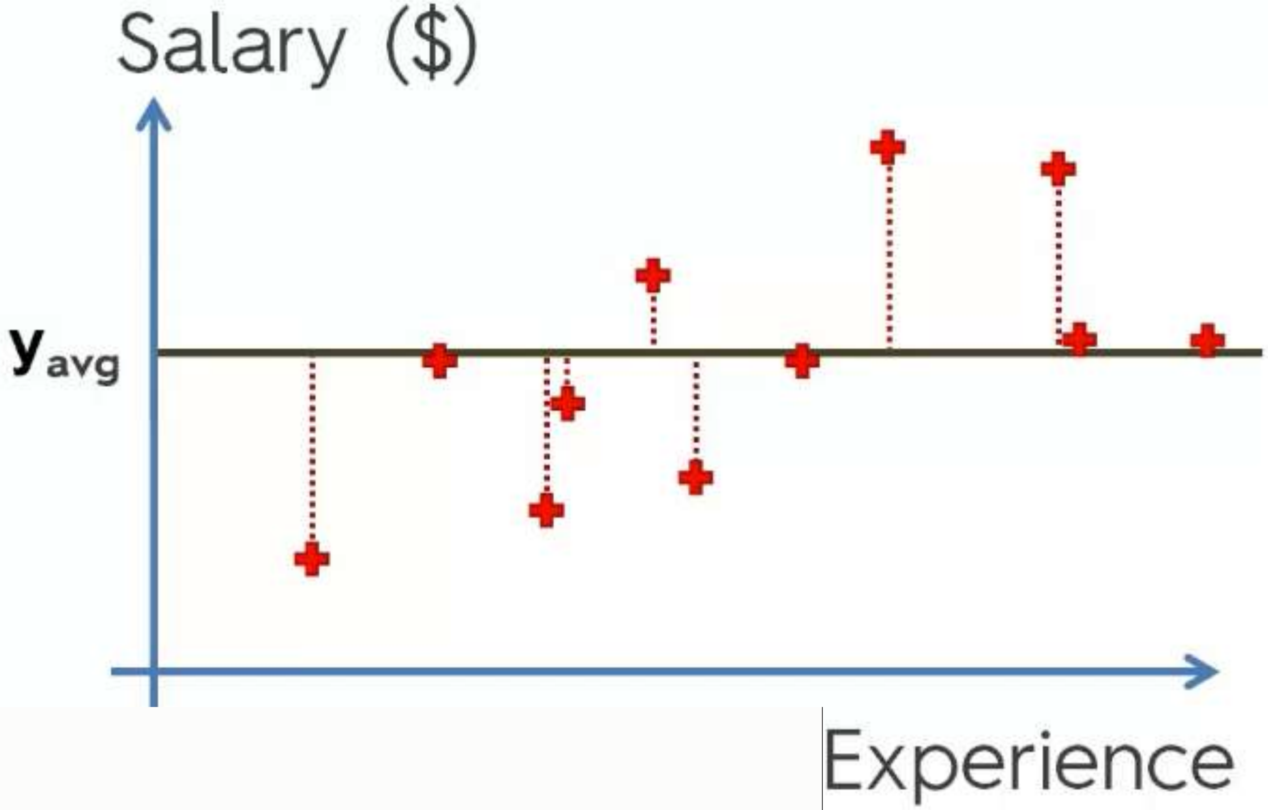
Simple Linear Regression:



SUM $(y_i - \hat{y}_i)^2 \rightarrow$ min

R-Square

Simple Linear Regression:



$$SS_{res} = SUM\ (y_i - \hat{y}_i)^2$$

$$SS_{tot} = SUM\ (y_i - y_{avg})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Demo 3: Get R-square for the previous demos.

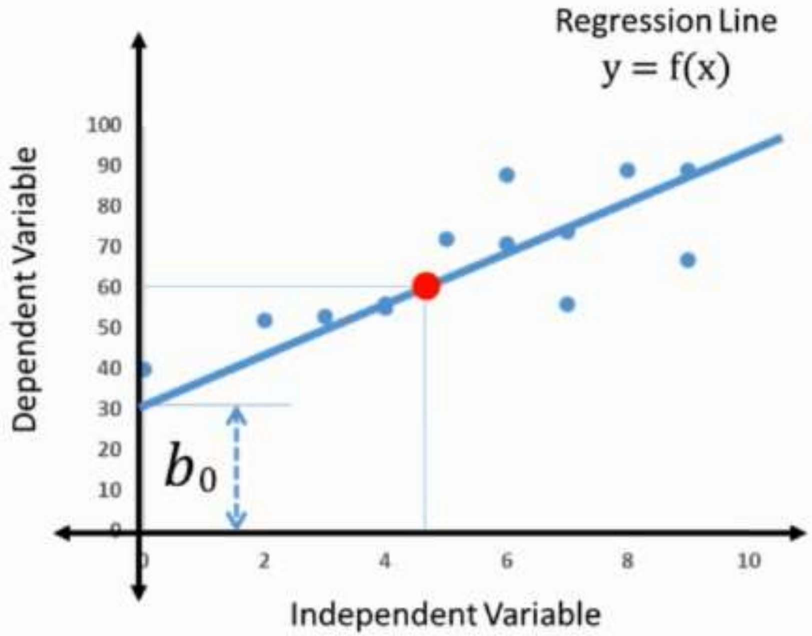
Multiple Linear Regression

Multiple Linear Regression

Simple Regression :

Only one Dependent
Only one Independent

$$y = b_0 + b_1 x$$



Multiple Linear Regression :

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

Multiple Linear Regression

Hrs Studied (X1)	Hrs Slept (X2)	Marks (Y)
0	8	40
2	8	52
3	7.5	53
4	7	55
4	9	56
5	8.5	72
6	9	71
6	7	88
7	6	56
7	7	74
8	9	89
9	6	67
9	9	89

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

Dependent Variable

Marks Obtained

Independent Variable

Hrs Studied
Hrs Slept

Demo 4: Create ML to predict Exam Result based on (Sleeping & Studying Hours)

Freedom of Wearing Shirts

- Office Wear
- Monday to Friday
- Can not repeat a shirt

Freedom of Wearing Shirts

- Monday 5
- Tuesday 4
- Wednesday 3
- Thursday 2
- Friday No Choice

Degrees of Freedom in Statistics

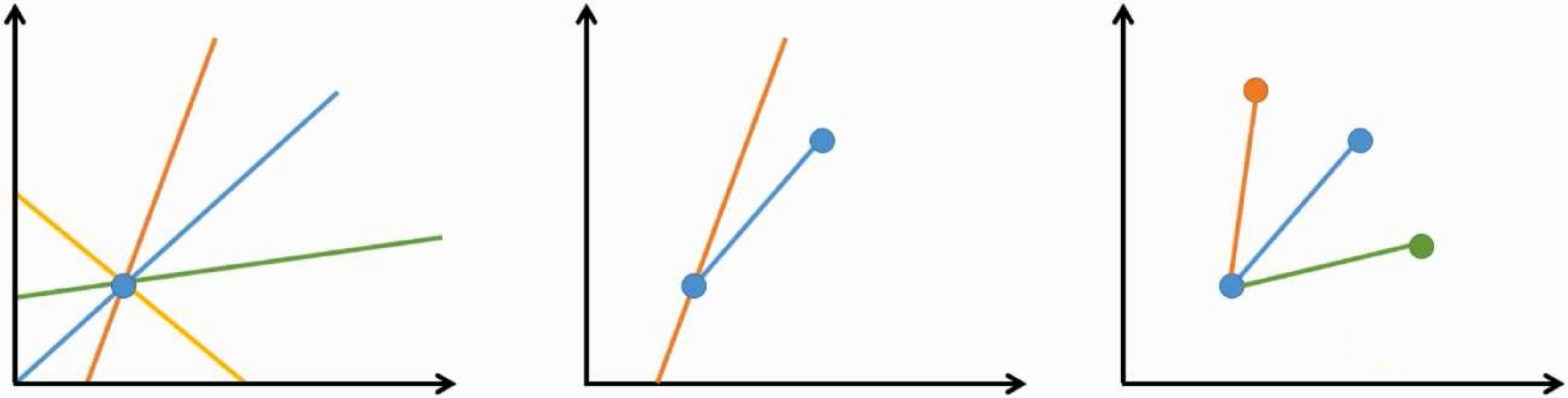
The number of values in the final calculation of a statistic that are free to vary.

OR

The minimum number of independent coordinates that can specify the position of the system completely.

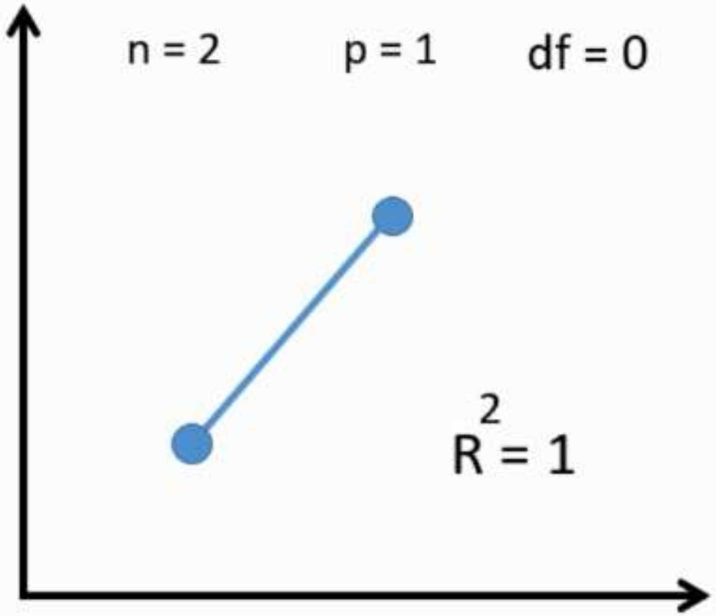
$$\mathrm{df} = \underbrace{n}_{\text{Number of Observations}} - \underbrace{p}_{\text{Number of variables}} - 1$$

Degrees of Freedom in Statistics



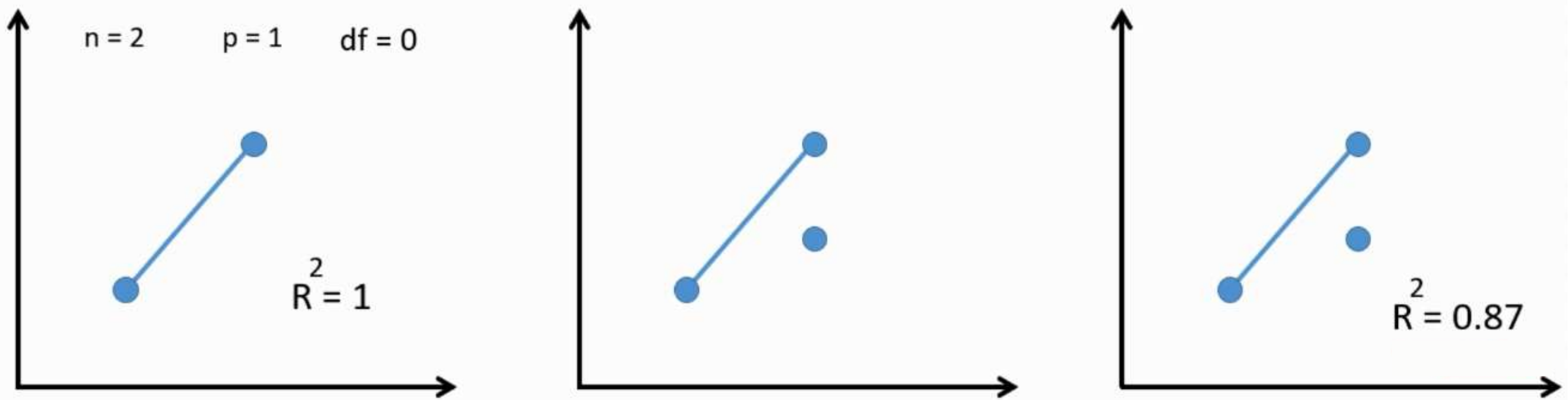
Degrees of Freedom in Statistics (n – p – 1)

$$y = b_0 + \ b_1 x_1$$



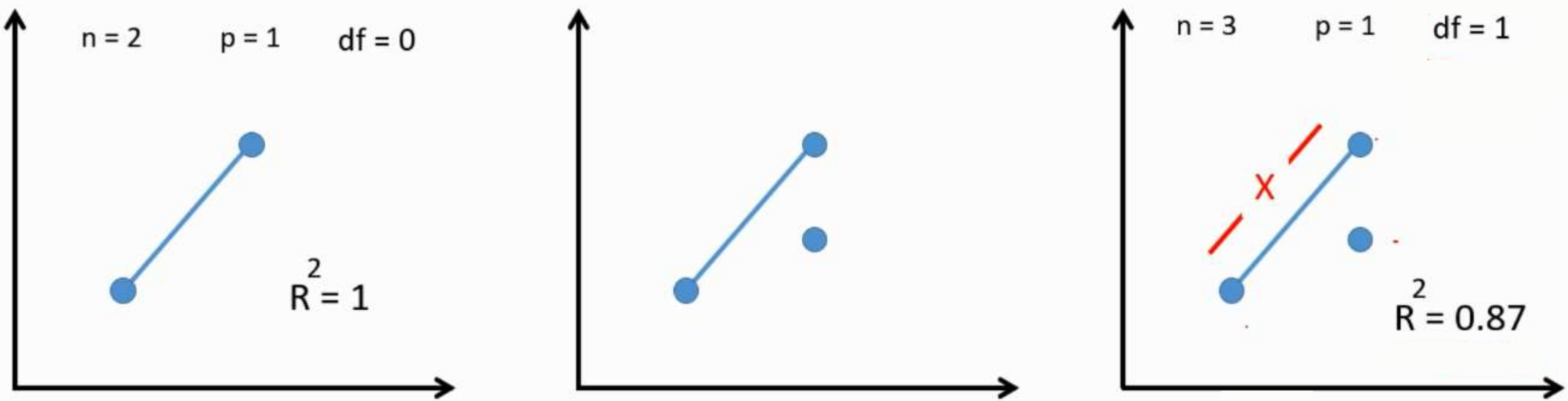
Degrees of Freedom in Statistics (n – p – 1)

$$y = b_0 + b_1 x_1$$



Degrees of Freedom in Statistics (n – p – 1)

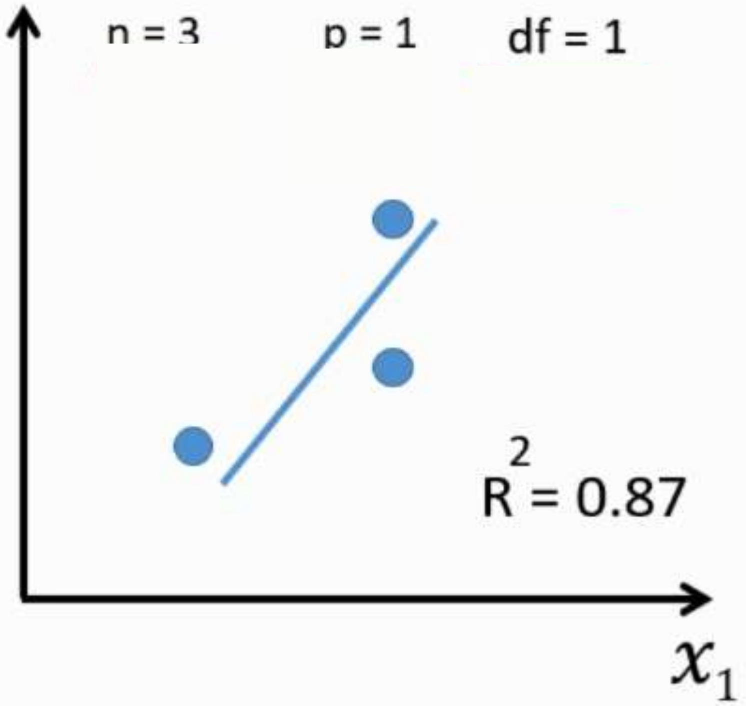
$$y = b_0 + b_1 x_1$$



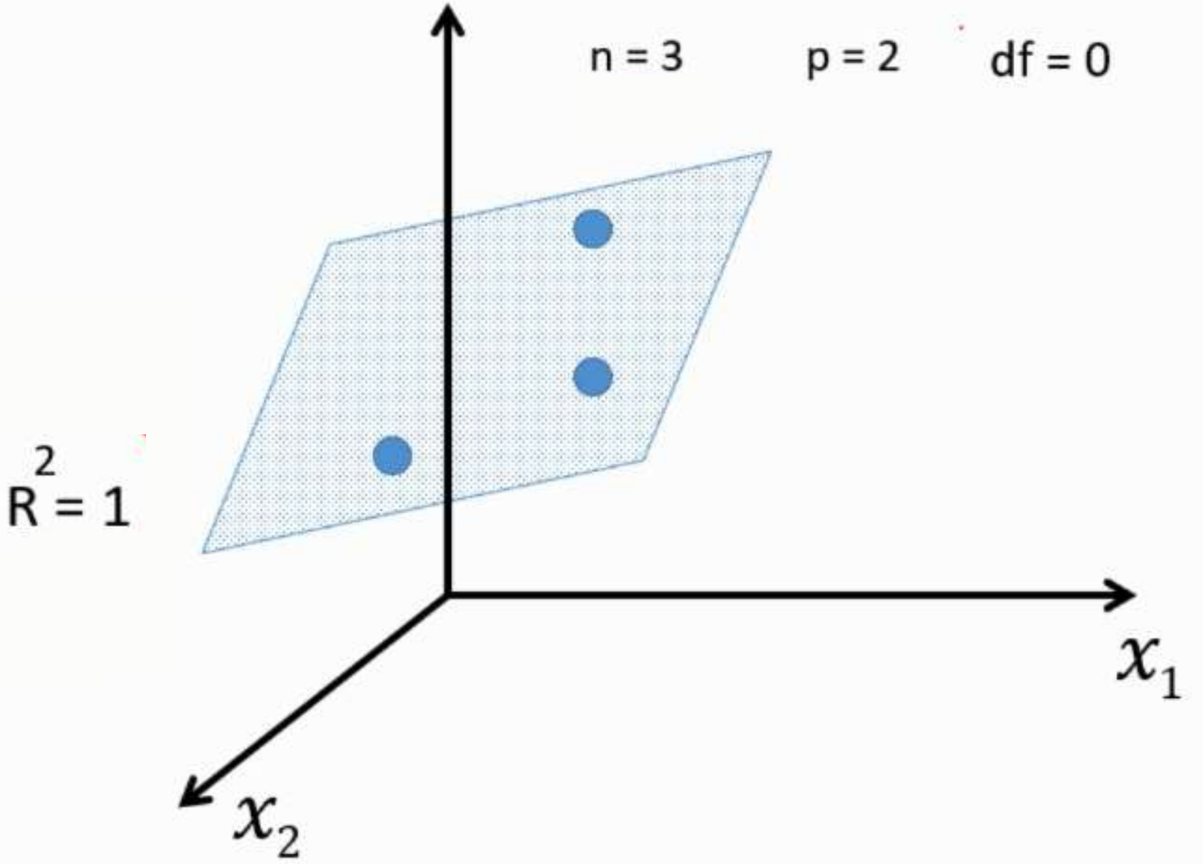
Note: As the (Degree of freedom) increase the (R-squared) decrease

Degrees of Freedom in Statistics (n – p – 1)

$$y = b_0 + b_1 x_1$$



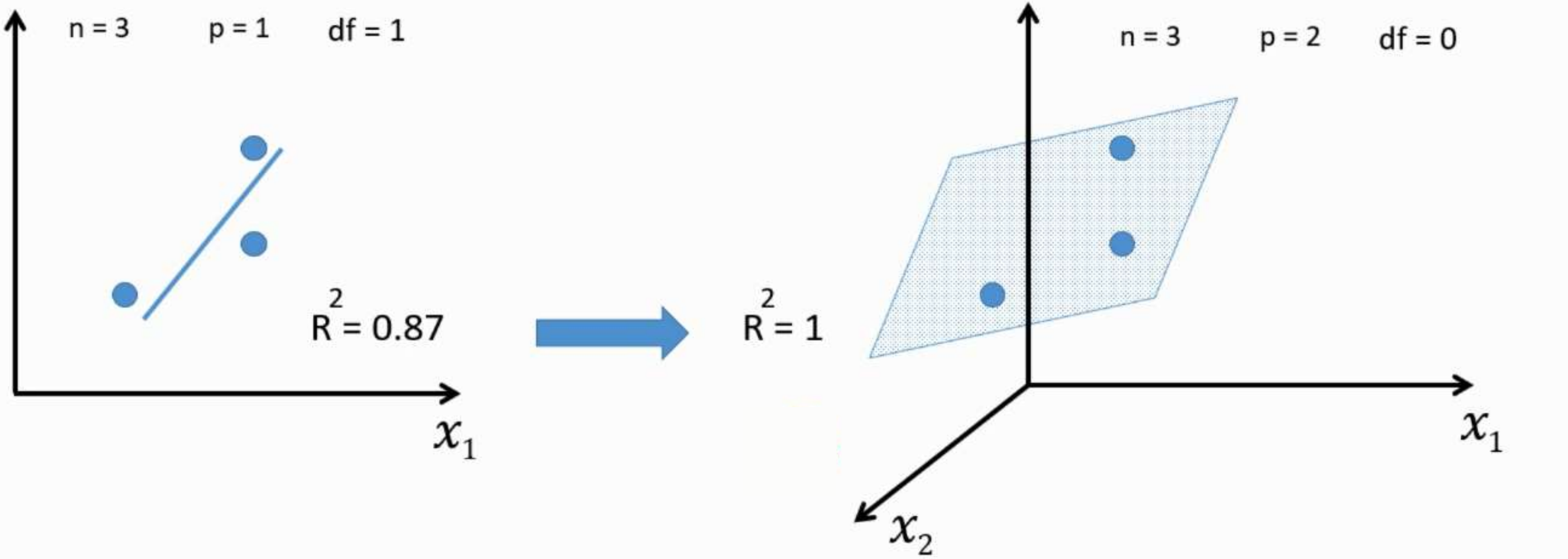
$$y = b_0 + b_1 x_1 + b_2 x_2$$



Degrees of Freedom in Statistics (n – p – 1)

$$y = b_0 + b_1 x_1$$

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



Note: Adding more variable helps to increase (R- squared) and that's what we want.

Degrees of Freedom in Statistics (n – p – 1)

Adding more variables increases value of R-Squared.

Higher the value of R-Squared,
Variation in Y is better explained by variation in X.

Let's add more variables.....

Degrees of Freedom in Statistics (n – p – 1)

Time to Finish

Driver's Experience

Engine Size

Horsepower

Number of laps

Type of Soup my grandma cooked.

Increase R-Squared somehow?

Adjusted R-Squared

Adjusted R-Squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 – Goodness of fit
(greater is better)

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

$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2$

Problem:

$\boxed{+\ b_3{}^*x_3}$

SS_{res} -> Min

R^2 will never decrease

Note: Since R-Squared will never decrease we can not know if adding variable is making the model good or bad, so incase of multivariate regression we need to use (Adjusted R-squared) to solve this issue.

Adjusted R-Squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Note: Adjusted R-squared) will increase only if the added variable is improving the model then (Adjusted R-squared) will increase, that's why we use (Adjusted R-squared) incase of multivariate regression.

$$\text{Adj } R^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

p - number of regressors

n - sample size

Adjusted R-Squared

Lower value of
Adjusted R-Squared

$$\bar{R}^2 = 1 - \left[\frac{(1 - R^2) * (n - 1)}{n - p - 1} \right]$$

Increase in this term

Lower Denominator due
to higher value of p.

If the R-Squared does no
increase significantly.

$R \quad = Sample\ R{-}Squared$

p $\quad$ = Number of independent variables

n $\quad$ = sample size or number of observations

Adjusted R-Squared

N	p	R-Squared	Adjusted R-Squared
50	10	0.80	0.75
50	12	0.82	0.76
50	15	0.83	0.75
50	20	0.84	0.73

$$\bar{R}^2 = 1 - \frac{(1 - R^2) * (n - 1)}{n - p - 1}$$

Assumptions of Multiple Linear Regression

Relationship Among Variables

- Linear Relationship
- Multicollinearity

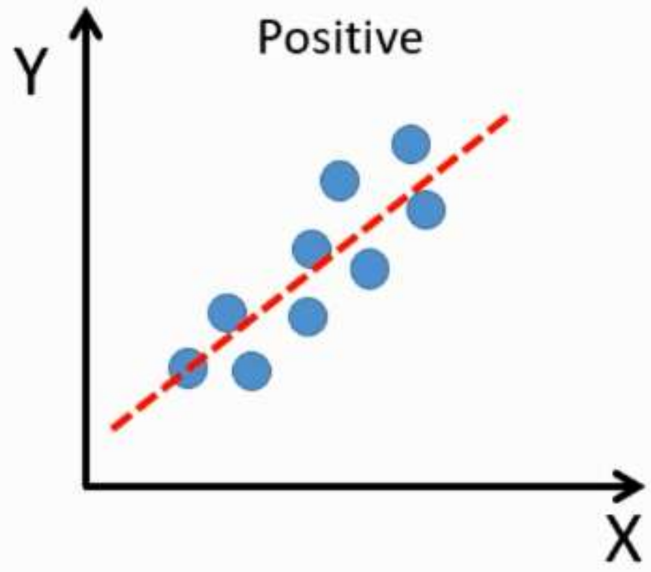
Behaviour of Data

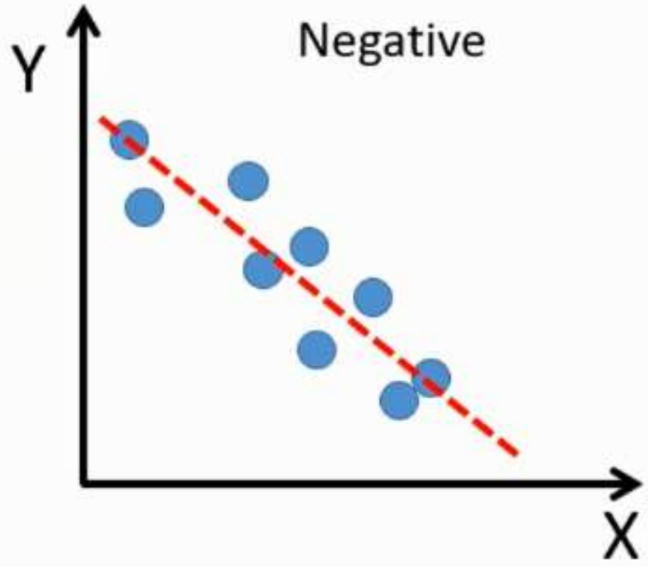
- Sample Size
- Normality

Linear Relationship

- Dependent and Independent Features have linear relationship

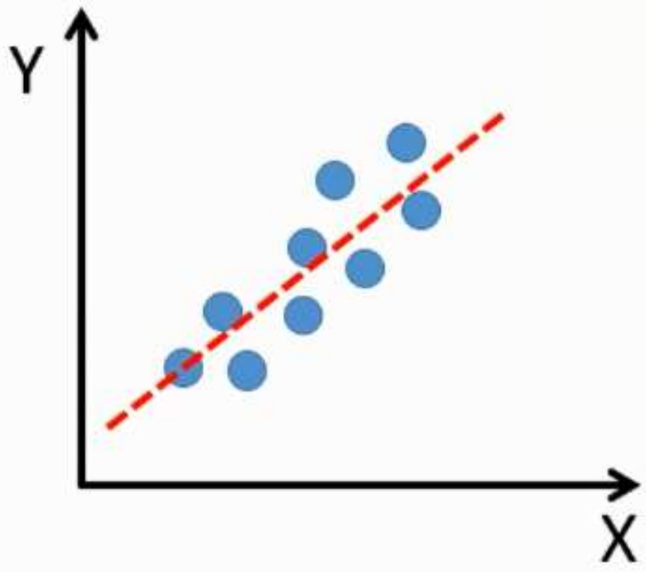
- Can be Positive or Negative correlation



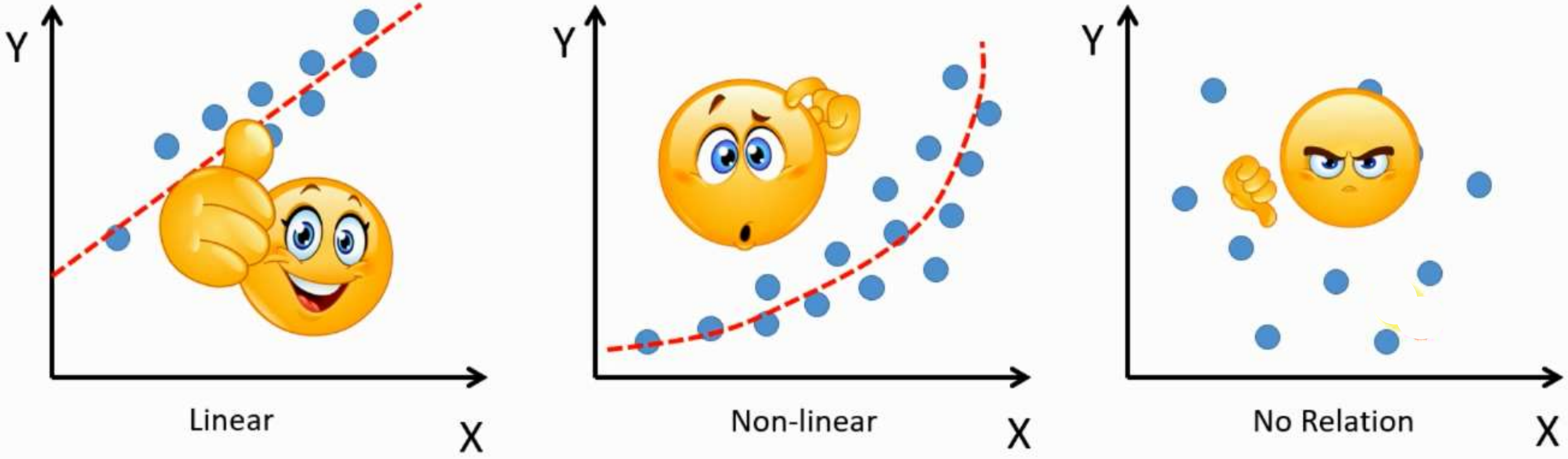


Linear Relationship

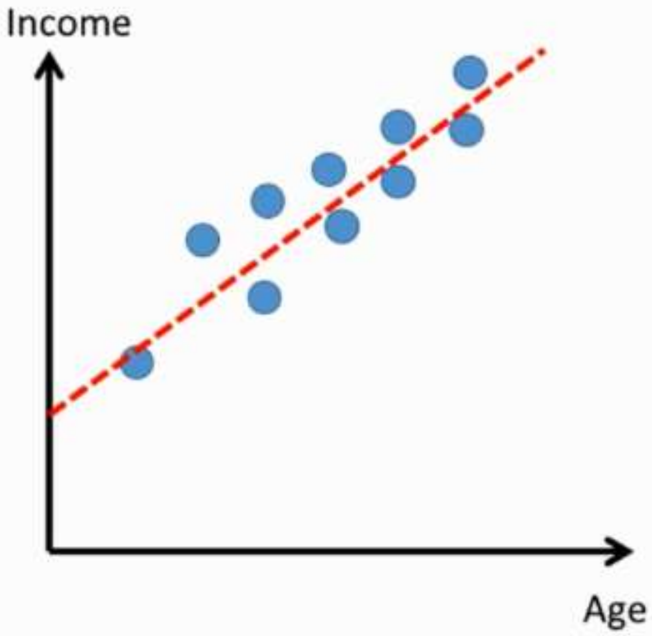
- Dependent and Independent Features have linear relationship
- Can be Positive or Negative correlation
- Can be checked using Pearson Correlation Coefficient as well as visualisation

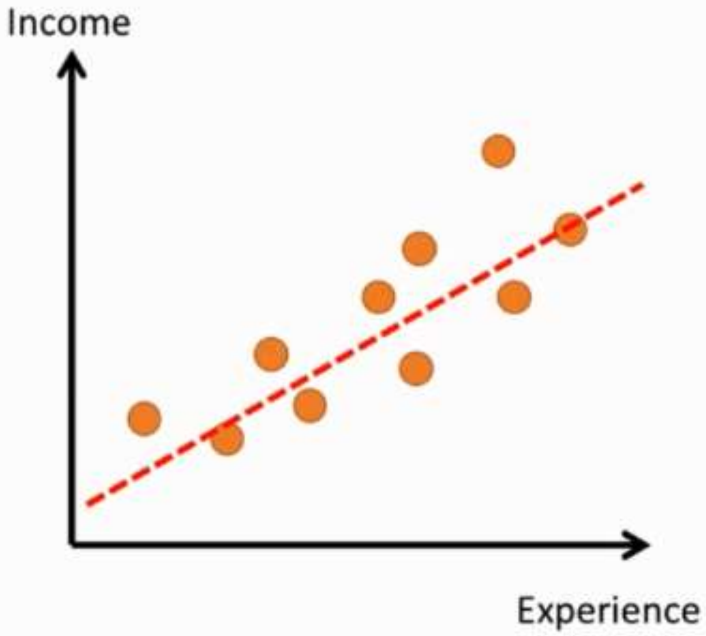


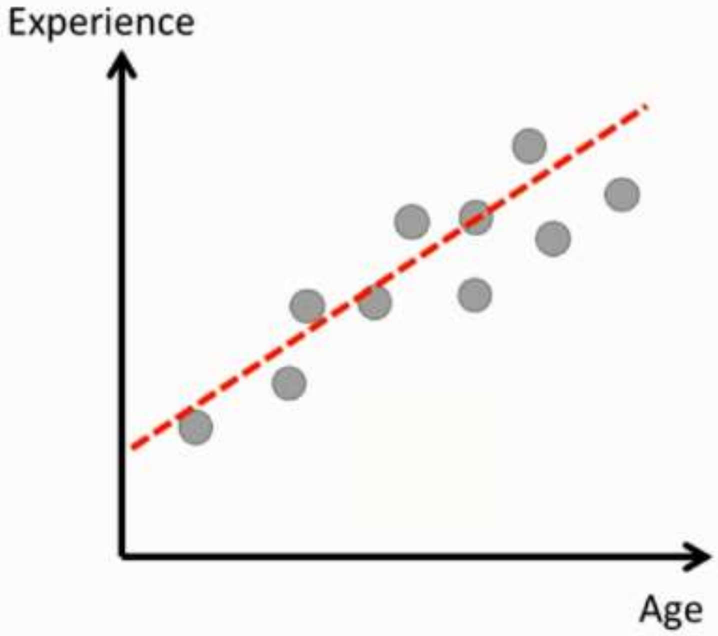
Linear Relationship



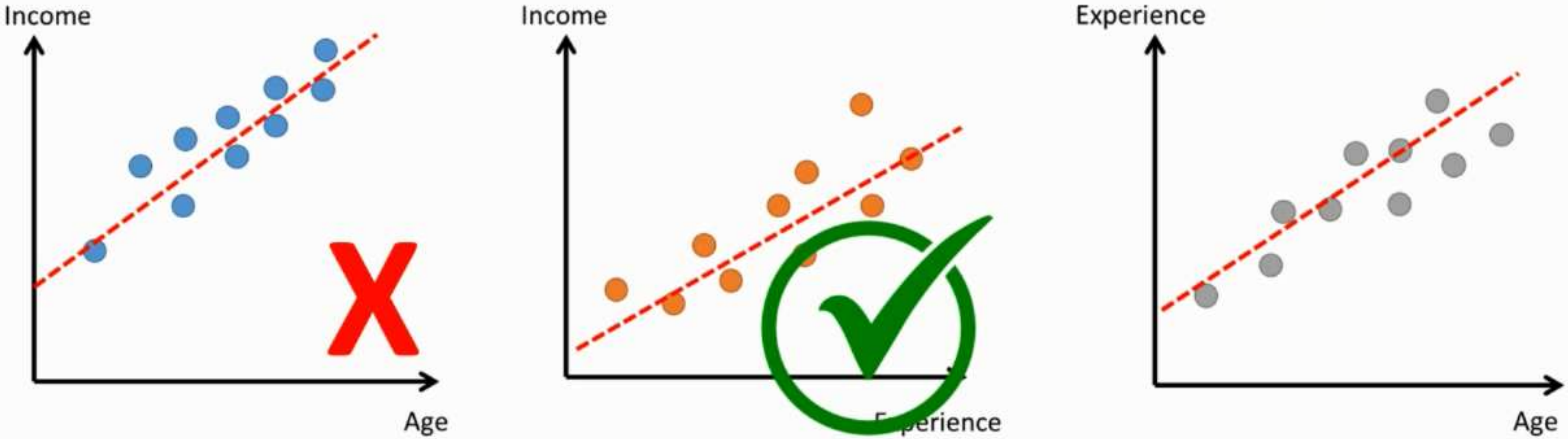
No Multicollinearity







No Multicollinearity



Since both age and experience are correlated with each other so we need to choose one of them only with the dependent variable.

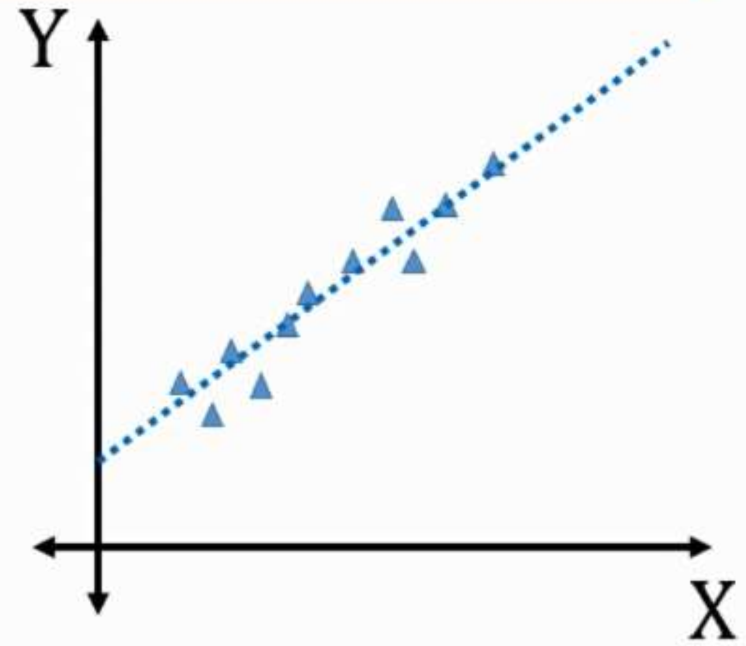
Correlation Coefficient Matrix

Age	Experience	Education Received	Salary
32	8	6	$ 8,000
40	15	8	$ 12,000
35	6	8	$ 10,000

	Age	Experience	Education Received	Salary
Age	1	0.9	0.2	0.7
Experience	0.9	1	0.15	0.72
Education Received	0.2	0.15	1	0.85
Salary	0.7	0.72	0.85	1

Statistically Correlated

- Strength of the correlation – Coefficient of Correlation
- Direction of correlation – Sign of the Coefficient



Pearson Correlation Coefficient

$$r = \frac{\sum (x - \bar{x}) * (y - \bar{y})}{(N - 1) * \sigma_x * \sigma_y}$$

Correlation Coefficient Matrix

Age	Experience	Education Received	Salary
32	8	6	$ 8,000
40	15	8	$ 12,000
35	6	8	$ 10,000

	Age	Experience	Education Received	Salary
Age	1	0.88	0.2	~~0.7~~ X
Experience	0.88	1	0.15	0.72
Education Received	0.2	0.15	1	0.85
Salary	0.7	0.72	0.85	1