Coefficient of Determination (R^2)

Data Science | CCDATSCL

R-squared is a statistical measure that tells you how well a regression model fits the data.
It tells you how well the model explains the variation in the data.

- **R-squared** is measured on a scale from **0 to 1**.
- A **value of 0** means that the model **does not explain any of the variation in the data**.
- A **value of 1** means that the model **explains all of the variation in the data**.

$$\textit{Sum of Squared Errors}$$

$$(SSE) = \sum_{i=1}^{n}(y_i - y_{predict})^2$$

$$\textit{Sum of Squared Total}\ (SST) = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

Sum of Squares Error (SSE)

- SSE represents sum of squares error, also known as **residual sum of squares**.
- It is the difference between the **observed value** and the **predicted value**.
- Usually, the lower the sum of squares error better model the regression. SSE is that part of the total variation which is not modeled by the regression line.

$$SSE = \sum_{i=1}^{n}\left(y_i - y_{predict}\right)^2$$

where:

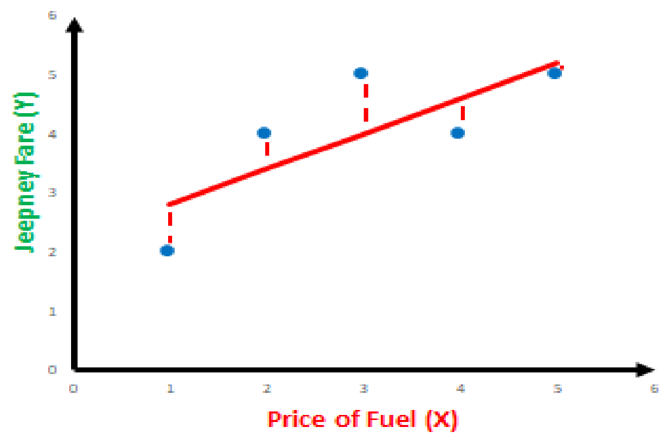
y_i is the one of the values of the **dependent variable**

$y_{predict}$ is one of the **predicted values**

Price of Fuel (X)	Jeepney Fare (Y)	Predicted Jeepney Fare ($Y_{predict}$)	$Y - Y_{predict}$	$(Y - Y_{predict})^2$
1	2	2.8	-0.8	0.64
2	4	3.4	0.6	0.36
3	5	4	1	1
4	4	4.6	-0.6	0.36
5	5	5.2	-0.2	0.04

$$SSE = \sum_{i=1}^{n}(y_i - y_{predict})^2$$

SSE = 2.4



Sum of Squares Total (SST)

- SST represents the total sum of squares. It is the squared values of the dependent variable to the sample mean.
- In other words, the total sum of squares measures **the variation in a sample**.

$$SST = \sum_{i=1}^{n}(y_i - y_{mean})^2$$

where:

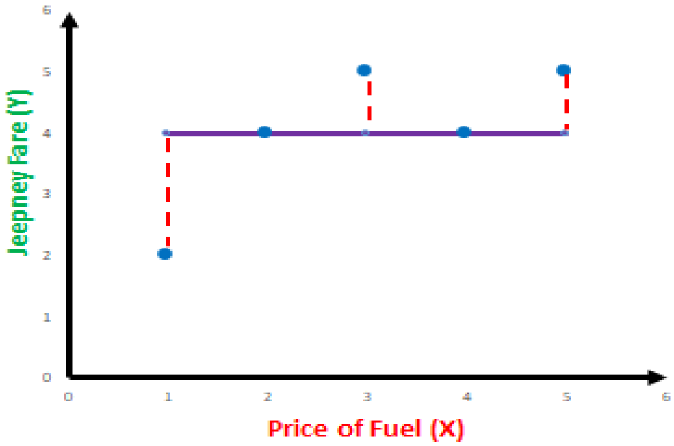
y_i is the one of the values of the **dependent variable**

y_{mean} is the **mean of the dependent variables**

Price of Fuel (X)	Jeepney Fare (Y)	Predicted Jeepney Fare ($Y_{predict}$)	$Y - Y_{predict}$	$(Y - Y_{mean})^2$
1	2	2.8	-0.8	4
2	4	3.4	0.6	0
3	5	4	1	1
4	4	4.6	-0.6	0
5	5	5.2	-0.2	1

$$\boldsymbol{SST} = \sum_{i=1}^{n} (y_i - y_{mean})^2$$

SST = 6.0



$$\boldsymbol{R^2 = 1 - \frac{2.4}{6}}$$

$$\boldsymbol{R^2 = 0.60}$$

We can say that the **price of fuel (X)** and **jeepney fare (Y)** relationship accounts for **60%** of the variation