STATISTICS FOR DATA SCIENCE

Power Test &
Simple Linear Regression

Dr. Karthiyayini
Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

Unit 5 : Power Test & Simple Linear Regression

Session : 6 (Continued Session)

Sub Topic : Least Squares Line

Dr. Karthiyayini
Department of Science & Humanities

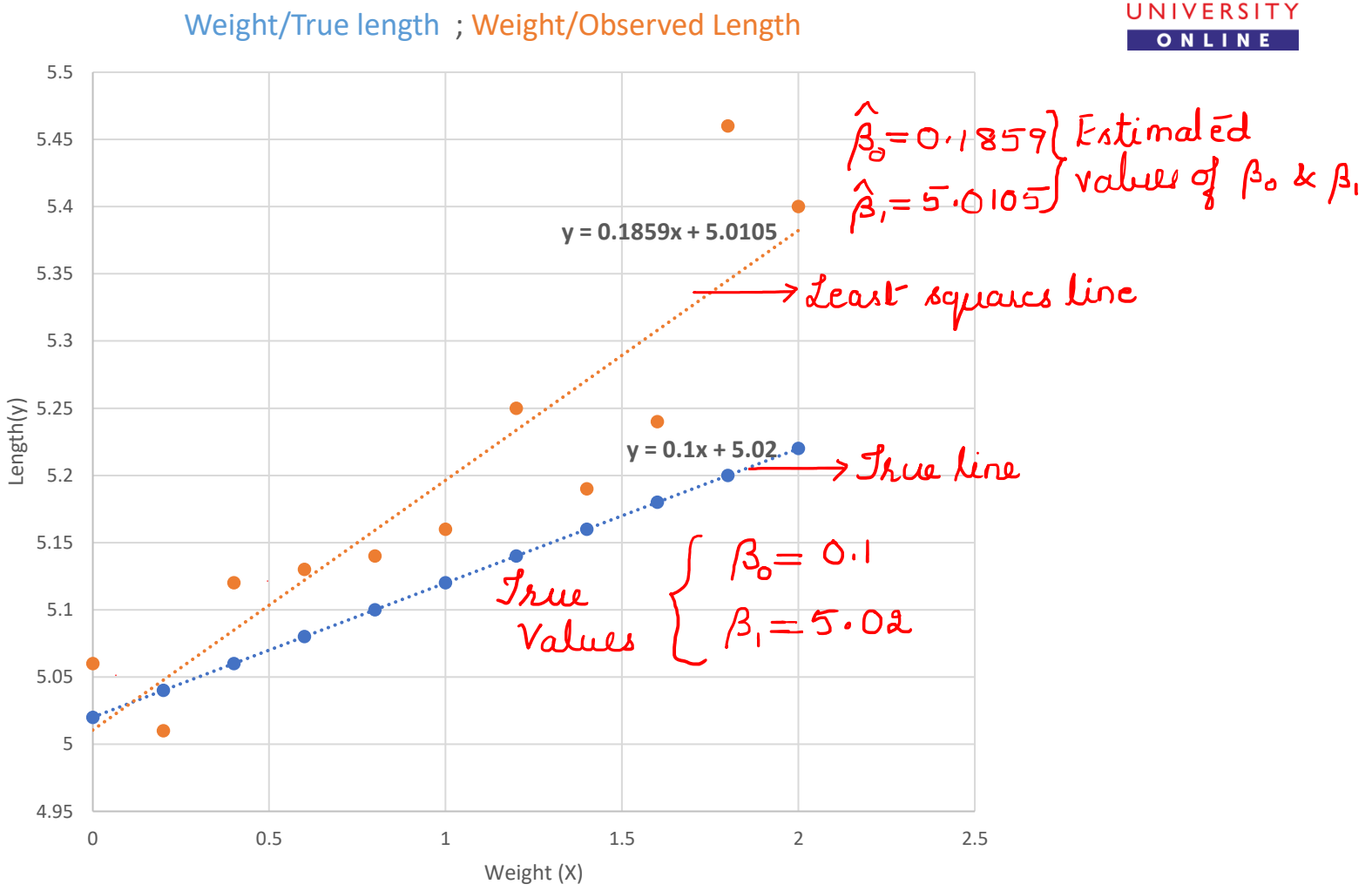
Some Observations :

- The Estimates are not the same as true values
- The Residuals are not the same as the Errors.
- Don’t extrapolate outside the range of the data.
- Don’t use the Least Squares line when the data aren’t linear.

STATISTICS FOR DATA SCIENCE

The Estimates are not the same as true values

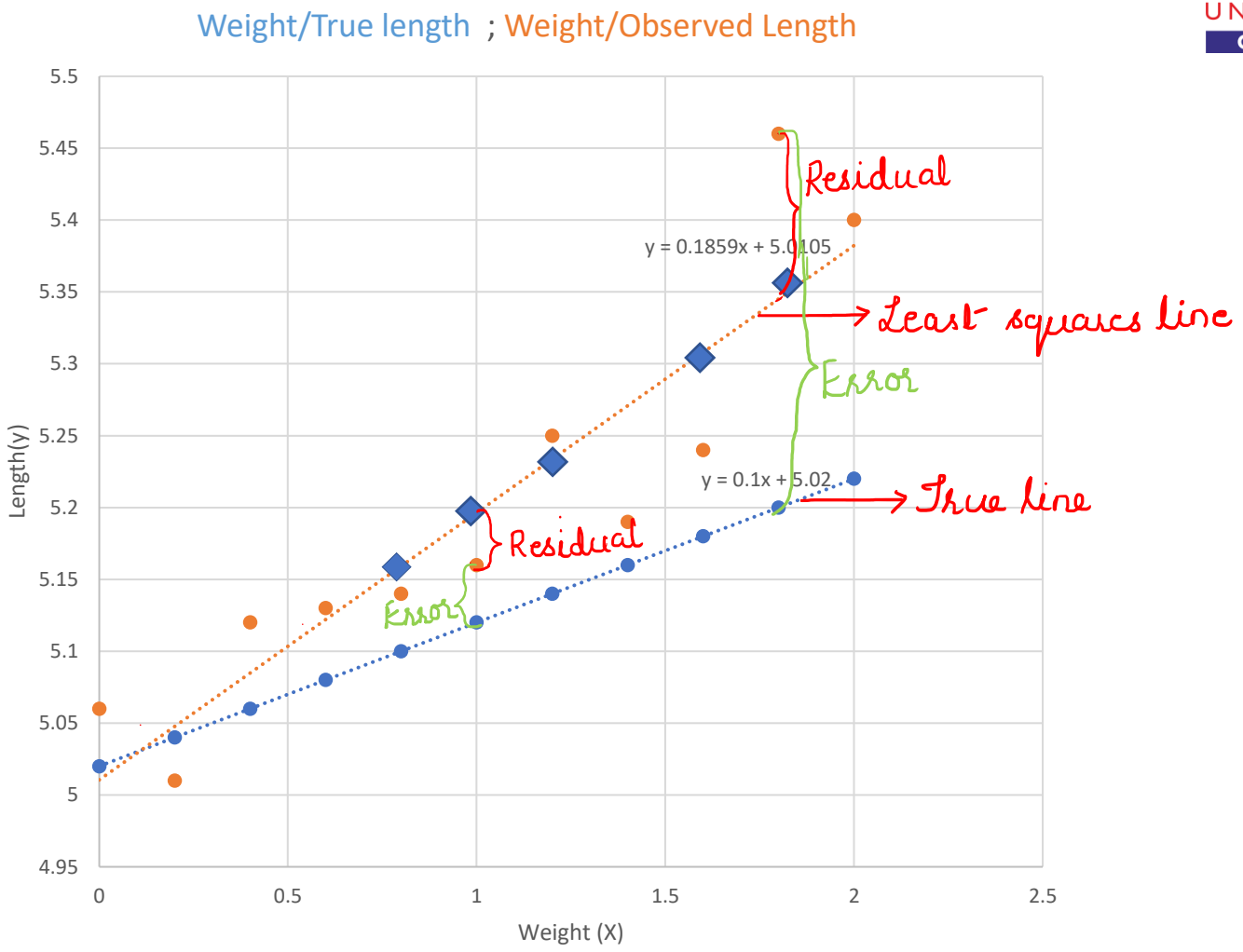
Weight (lb) (x)	True Length $(in.)$ (y)	Length $(in.)$ (y)
0.0	5.02	5.06
0.2	5.04	5.01
0.4	5.06	5.12
0.6	5.08	5.13
0.8	5.10	5.14
1.0	5.12	5.16
1.2	5.14	5.25
1.4	5.16	5.19
1.6	5.18	5.24
1.8	5.20	5.46
2.0	5.22	5.40



STATISTICS FOR DATA SCIENCE

The Residuals are not the same as Errors

Weight (lb) (x)	Length $(in.)$ (y)	Length $(in.)$ (y)
0.0	5.02	5.06
0.2	5.04	5.01
0.4	5.06	5.12
0.6	5.08	5.13
0.8	5.10	5.14
1.0	5.12	5.16
1.2	5.14	5.25
1.4	5.16	5.19
1.6	5.18	5.24
1.8	5.20	5.46
2.0	5.22	5.40



STATISTICS FOR DATA SCIENCE

Don’t Extrapolate outside the range of the data!!

❖The details pertaining to the no. of hours spent by students in preparing for an entrance exam and the marks scored (on a scale of (0 – 100) is provided in the following table.

Using these values,

i. Estimate the marks scored by a student who has spent 2.35 hours.

ii. Predict the marks that a student can score if he/she invests 20 hours.

SL No.	No. of hours spent	Marks Scored
1	6	82
2	10	88
3	2	56
4	4	64
5	6	77
6	7	92
7	0	23
8	1	41
9	8	80
10	5	59
11	3	47

Solution : i. 45.43 marks
ii. 160 marks
{Refer previous session}

In the second case, if a student puts in 20 hours of effort the predicted marks 160 is not feasible since the marks range is (0 to 100).

Therefore extrapolating outside the range of the data may not give useful/logical predictions.

STATISTICS FOR DATA SCIENCE

Don’t Extrapolate outside the range of the data!!

Weight (lb) (x)	Length $(in.)$ (y)	Weight (lb) (x)	Length $(in.)$ (y)
0.0	5.06	2.0	5.40
0.2	5.01	2.2	5.57
0.4	5.12	2.4	5.47
0.6	5.13	2.6	5.53
0.8	5.14	2.8	5.61
1.0	5.16	3.0	5.59
1.2	5.25	3.2	5.61
1.4	5.19	3.4	5.75
1.6	5.24	3.6	5.68
1.8	5.46	3.8	5.80

Here, the least squares line is given by $y = 0.2046x + 4.997$
{Refer to Example problem in the previous session}

For weight $x = 100lb$,
length $y = (0.2046)(100) + 4.9997 = 25.46$

Observations :

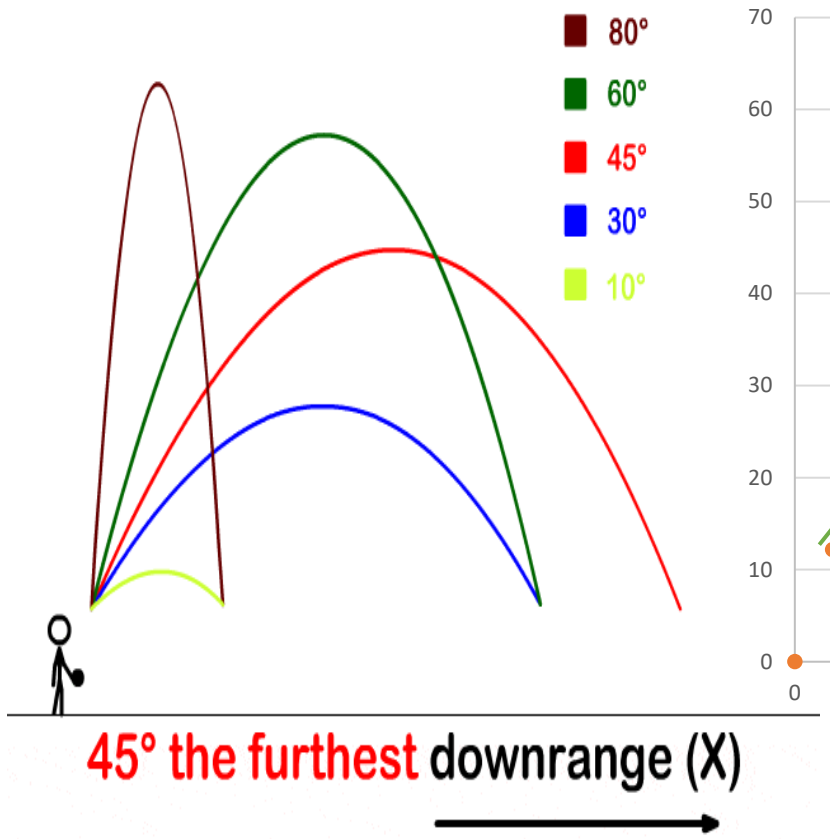
- The spring may not withstand this much weight and would stretch out of shape in which case the Hooke’s law will not hold.

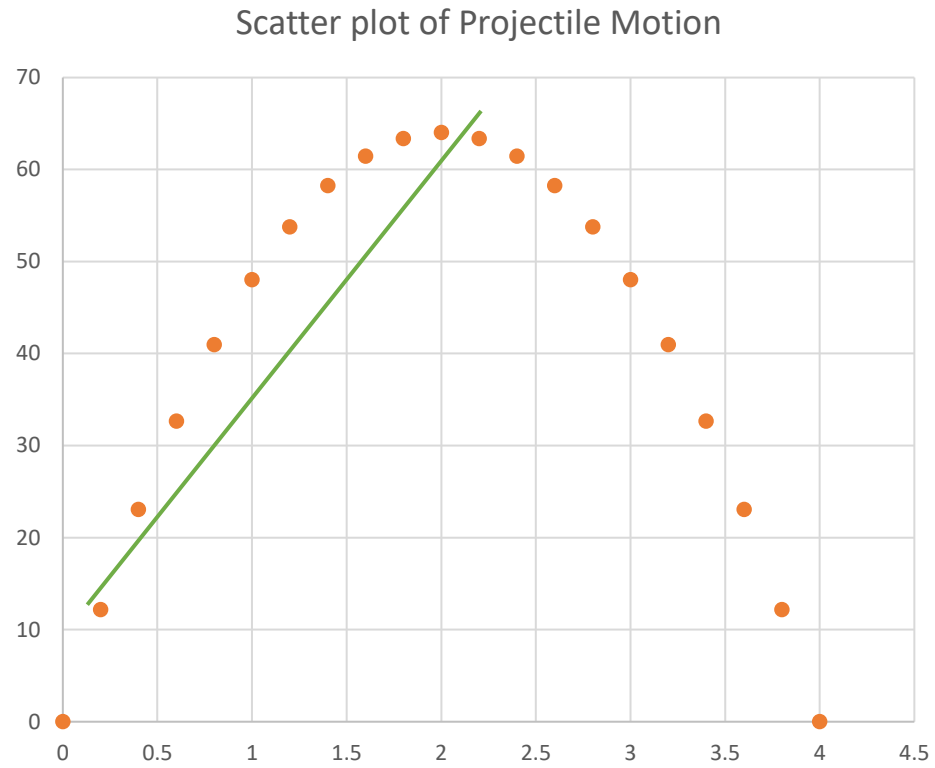
- Therefore, if we want to know how the spring will respond to a load Of 100lb then we need to include 100lb or more in our data set.

- Hence do not extrapolate outside the range of the data.

STATISTICS FOR DATA SCIENCE

Don’t use the Least Squares Line when the data aren’t linear





❖ We have seen that in the case of the projectile motion, the correlation coefficient between height and time is zero which is not true since the height of the falling body varies with time.

❖ So using correlation coefficient in case of non linear relationships lead to invalid results.

❖ Similarly, it is not advisable to use the Least squares line for non linear data will give misleading predictions and hence the use of Least squares for non linear data must be avoided.

Note : In some cases the Least – Squares line can be used for *non linear data,* but only after *variable transformation* is applied.

Source : stickmanphysics.com

Alternate form of the Least Squares Line

Consider the correlation co-efficient given by,

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Multiplying both sides by $\frac{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$ we obtain,

$$r.\frac{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \widehat{\beta_1}$$

$$\Rightarrow \widehat{\beta_1} = r.\frac{S_y}{S_x}$$

Alternate form of the Least Squares Line

Also, consider the least squares line given by,

$$y = \bar{y} - \widehat{\beta_1}\bar{x} + \widehat{\beta_1}x$$

$$\Rightarrow (y - \bar{y}) = \widehat{\beta_1}(x - \bar{x})$$

$$\Rightarrow (y - \bar{y}) = r.\frac{S_y}{S_x}(x - \bar{x})$$

This implies that the least squares line passes through the center of mass of the scatterplot $(\bar{x}, \bar{y})$, with slope $\widehat{\beta_1} = r.\frac{S_y}{S_x}$

Measuring goodness of fit

- A goodness of fit statistic is a quantity that measures how well a model explains a given set of data.
- A linear model fits well if there is a strong relationship between the variables involved.
- The strength of a linear relationship can be measured by considering,

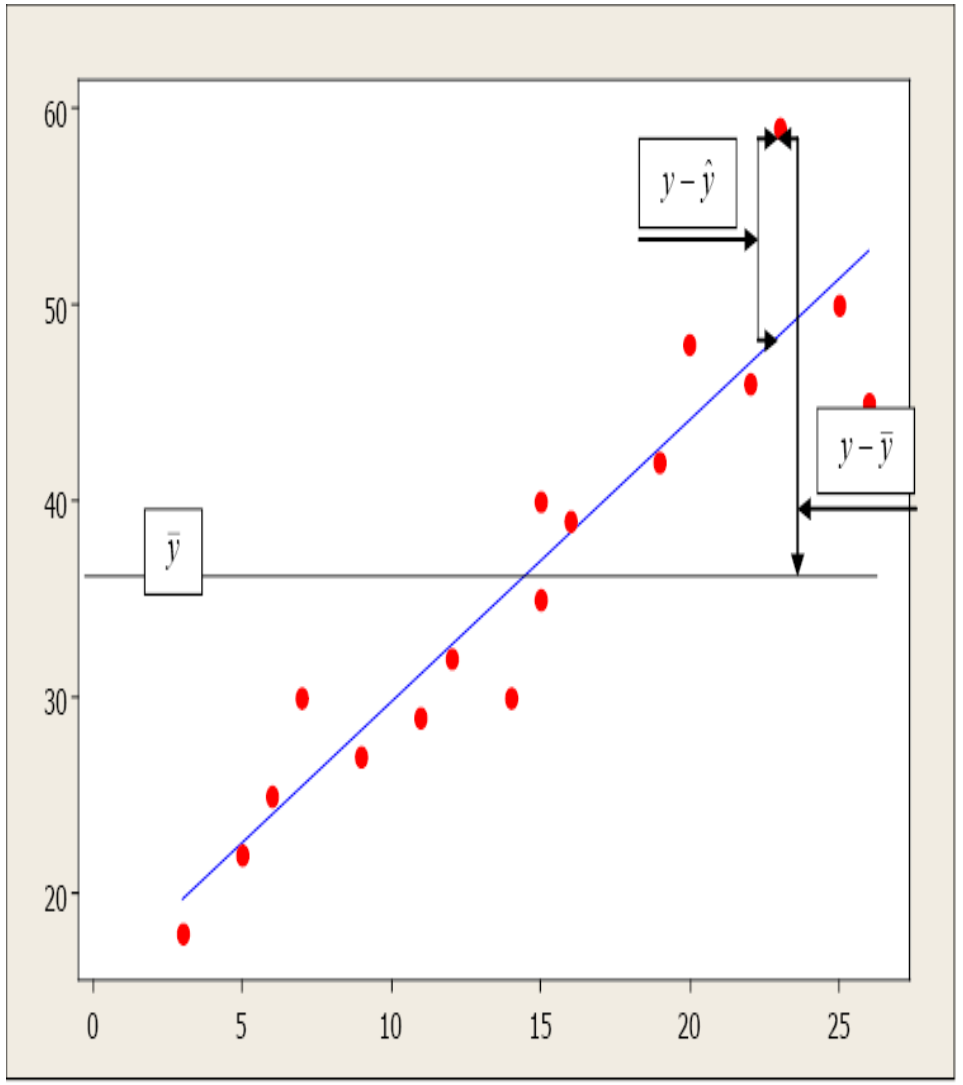
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

- The above relation is also referred to as a goodness-of-fit statistic.
- The draw back of this statistic relation is that it cannot be used to compare the goodness-of-fit of two models which have different data set. (That is, data sets having different units)
- Hence we use the relation, $r^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2 - \sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

  which is obtained by using the correlation coefficient.
- This is also referred to as the ***co-efficient of determination.***

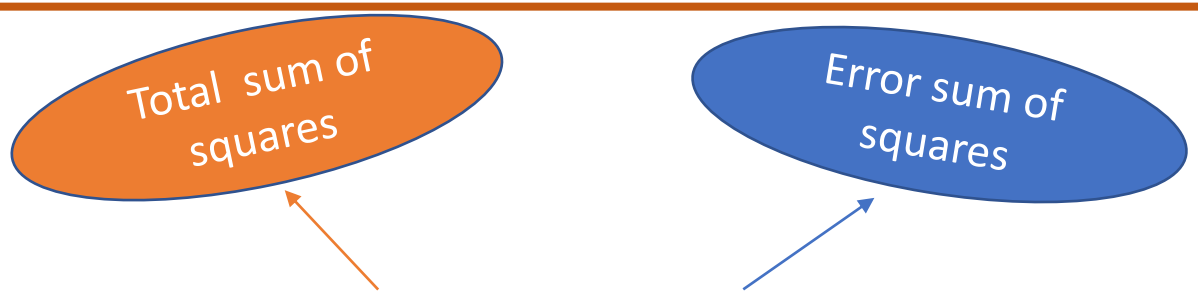
Visualisation of r^2

- $y_i - \bar{y}$: Vertical distance from the point (x_i, y_i) to the horizontal line $y = \bar{y}$.
- $y_i - \hat{y}_i$: Vertical distance from the point (x_i, y_i) to the least squares line.
- $\sum_{i=1}^{n}(y_i - \bar{y})^2$: Measures the overall spread of the points around the line $y = \bar{y}$.
- $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: Measures the overall spread of the points around the least squares line.
- $\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: Measures the reduction in the spread of the points obtained by using the least squares line rather than the line $y = \bar{y}$.



Source : courses.lumenlearning.com

Some special terminologies!



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

- $\sum_{i=1}^{n}(y_i-\bar{y})^2 - \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$: Regression sum of squares

- Therefore, Total sum of squares = Regression sum of squares + Error sum of squares

- And , $r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}}$

- r^2 is also referred to as the *proportion of the variance in y explained by Regression.*

More about r^2

- The coefficient of determination **Coefficient of determination**, in statistics, R^2 (or r^2), a measure that assesses the ability of a model to predict or explain an outcome in the linear regression setting.

- More specifically, R^2 indicates the proportion of the variance in the dependent variable (Y) that is predicted or explained by linear regression and the predictor variable (X, also known as the independent variable).

More about r^2

- Is a quantity that indicates how well a statistical model fits a data set. In other words, it is a statistical measure of how close the observed data are to the fitted regression line.
- It explains how much variation in the dependent variable y is characterized by a variation in the independent variable x.
- It is used to forecast or predict the possible outcomes.
- Its value lies between 0 and 1.
- The higher the value of r^2 , the better the prediction.

PES UNIVERSITY ONLINE

THANK YOU

Dr. Karthiyayini
Department of Science & Humanities

Karthiyayini.roy@pes.edu

+91 80 6618 6651