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

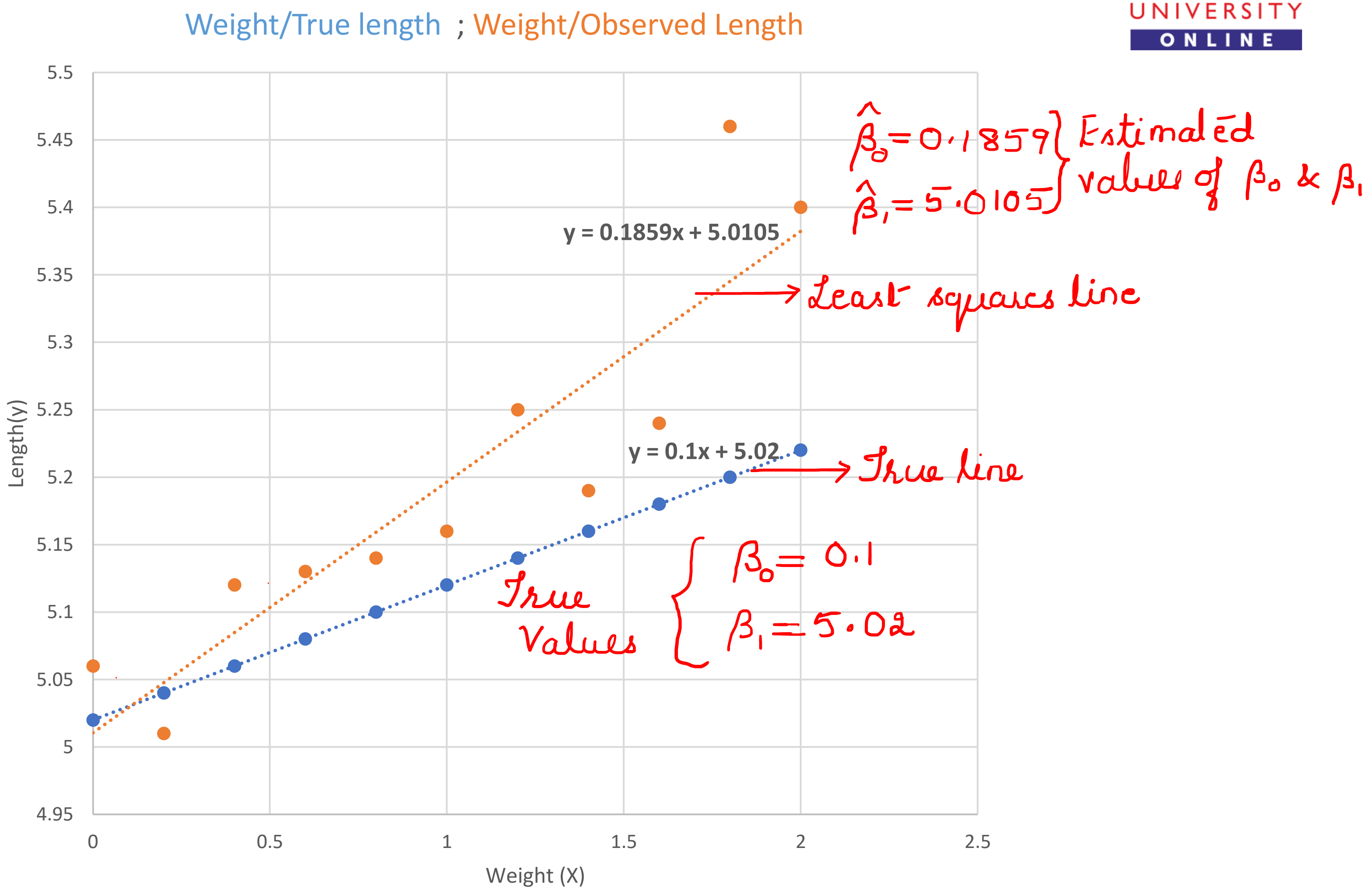
Dr. Karthiyayini
Department of Science & Humanities

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.

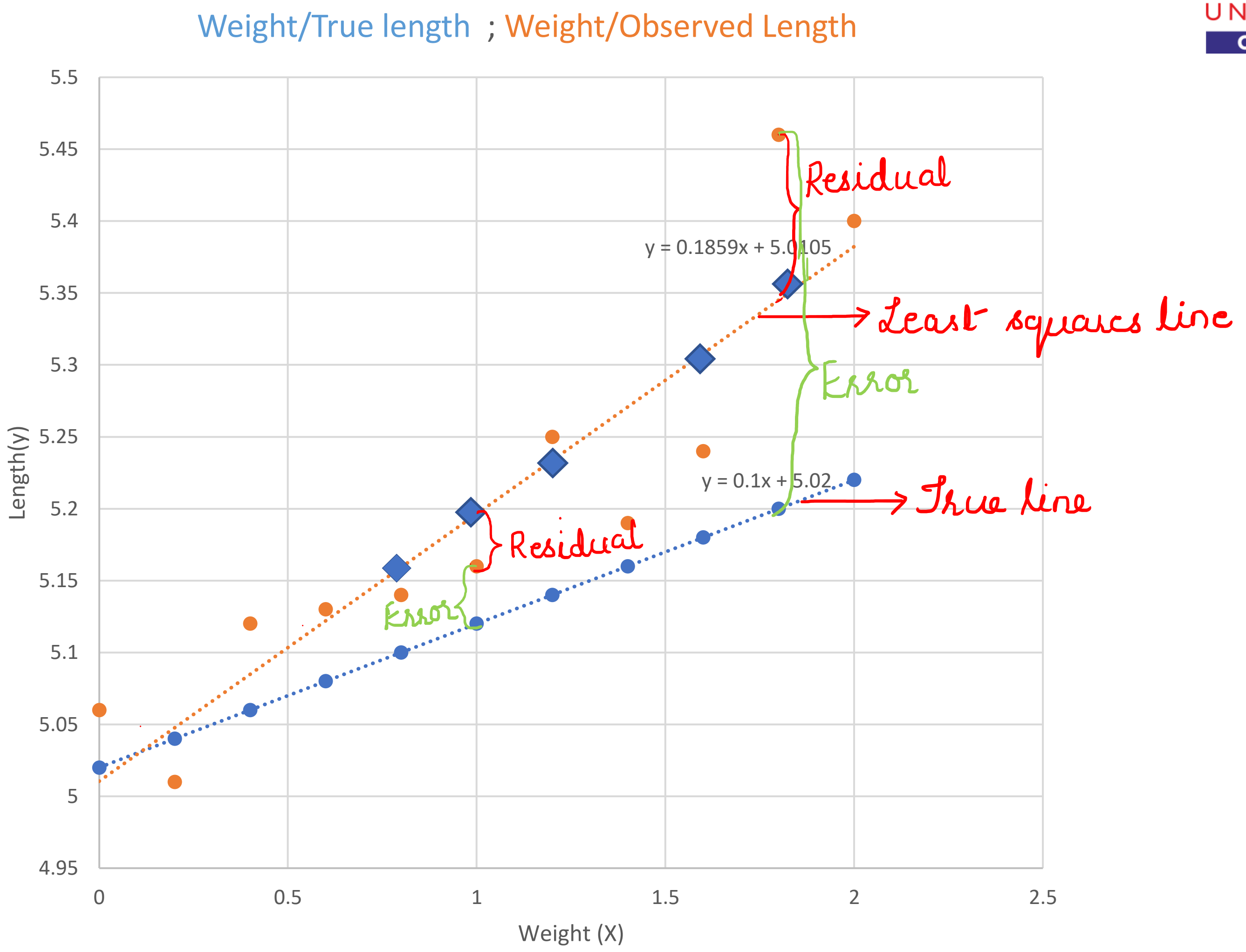
The Estimates are not the same as true values

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



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



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.

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

Using these values,

i. Estimate the marks scored by a student who has spent 2.35 hours. $\boxed{45.43}$

ii. Predict the marks that a student can score if he/she invests 20 hours. $\boxed{160}$

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

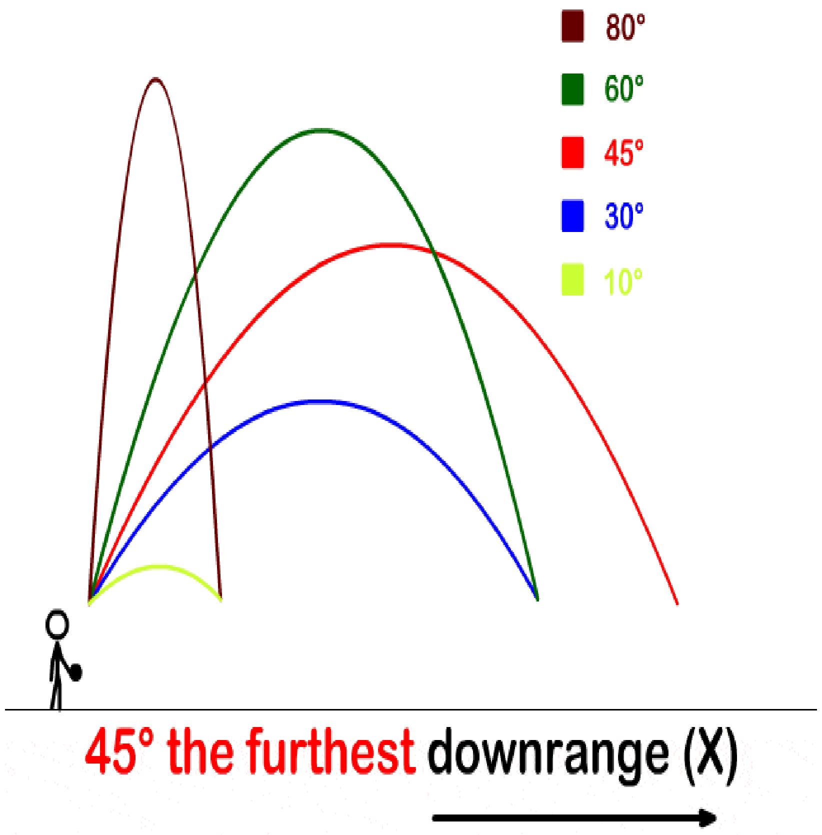
Least Square line : $y = 0.2046x + 4.997$

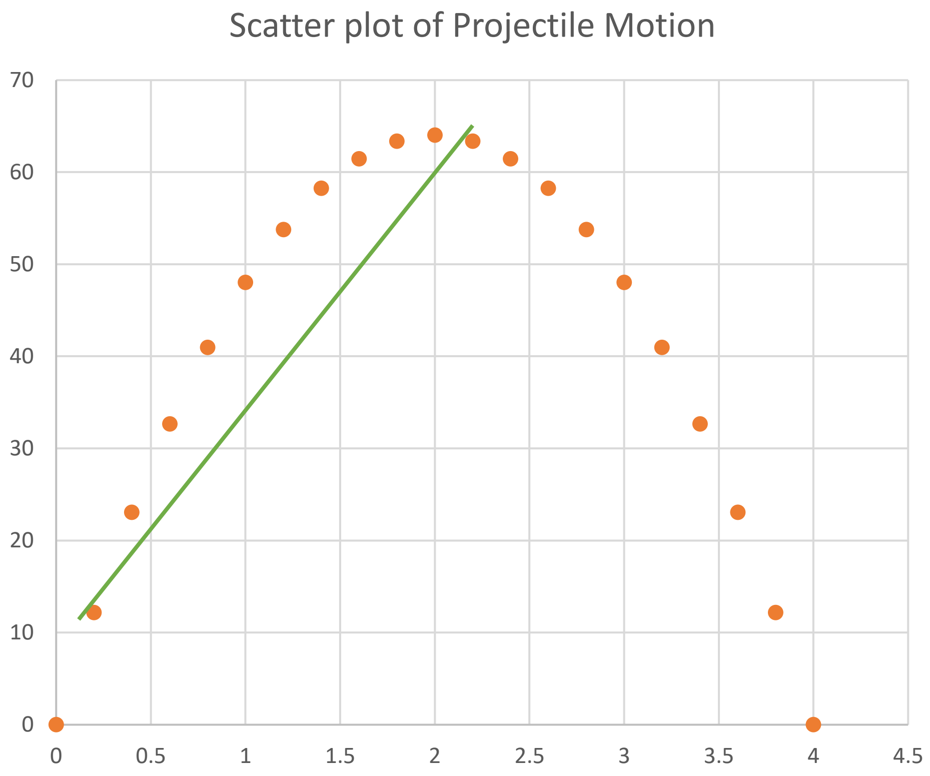
For weight, $x = 100$ lb,

Length, $y = (0.2046)(100) + 4.997$

$= 25.46$ in.

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





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

Measuring goodness of fit

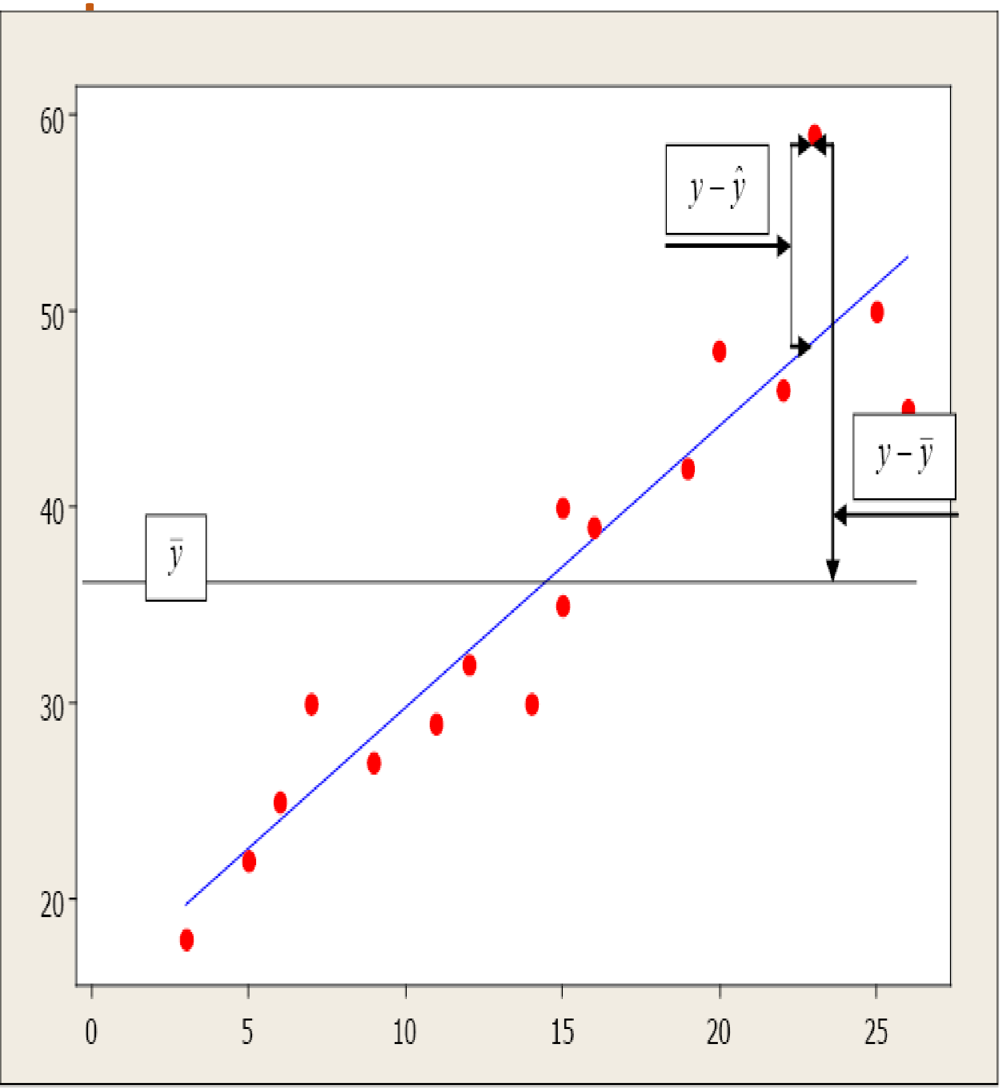
- A goodness of fit statistic is a quantity that measures how well a model explains a given set of data.
- A linear model fits well if there is a strong relationship between the variables involved.
- The strength of a linear relationship can be measured by considering,

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

- 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



$y-\hat{y}$: distance of (x_i, y_i) from the least squares line.

$y-\bar{y}$: distance of (x_i, y_i) from the line $y=\bar{y}$.

$(y-\hat{y})^2$: gives the overall spread of the data around the least squares line.

$(y-\bar{y})^2$: gives the overall spread of the data around the line $y=\bar{y}$

$\sum_{i=1}^{n}(y_i-\bar{y})^2 - \sum_{i=1}^{n}(y_i-\hat{y})^2$: goodness of fit statistic

Source : courses.lumenlearning.com

Some special terminologies!

Total sum of squares

Error sum of squares

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

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

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

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

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

More about r^2

- 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 $\boldsymbol{r^2}$, the better the prediction.

THANK YOU

Dr. Karthiyayini
Department of Science & Humanities

Karthiyayini.roy@pes.edu

+91 80 6618 6651