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

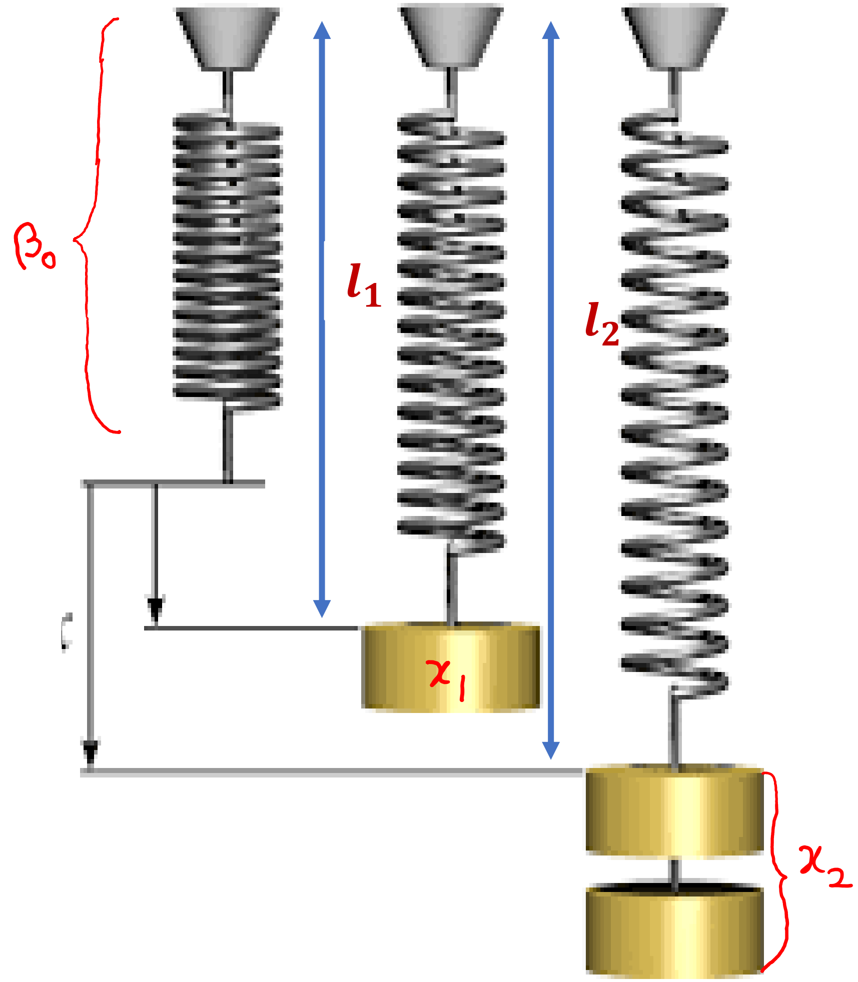
Session : 6

Sub Topic : Least Squares Line

Dr. Karthiyayini
Department of Science & Humanities

- How to compute the Least Squares Line
- Residuals and Errors
- Measuring Goodness of fit

How to compute the Least – Squares Line ???



Weights = x_i

Corresponding length = l_i

$\therefore$ l_i = actual length of the spring + stretched length

= β_0 + stretched length

Hooke's Law:

strain (deformation of spring) $\propto$ stress applied

$\Rightarrow$ stretched length $\propto$ weights applied

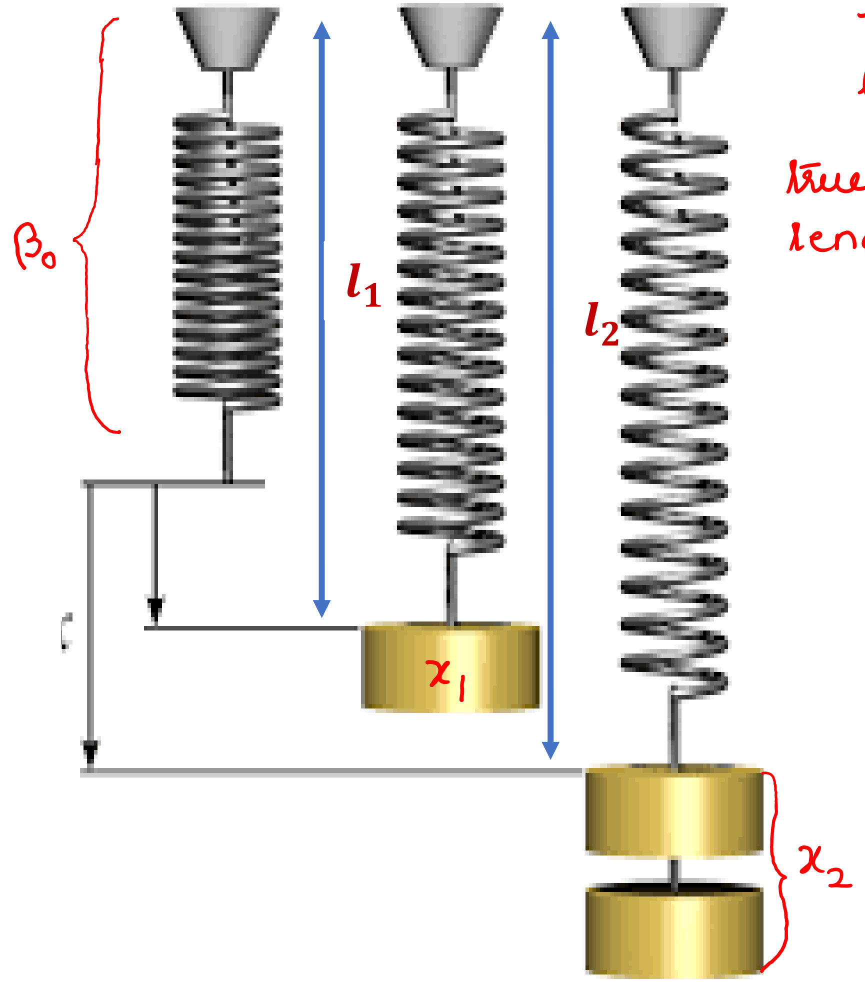
= kx_i $\longrightarrow$ proportionality constant.

= $\beta_1 x_i$ $\longrightarrow$ spring constant.

$\therefore$ Length, $\boxed{l_i = \beta_0 + \beta_1 x_i}$ $\longrightarrow$ 1

Source : Internet

How to compute the Least – Squares Line ???



Source : Internet

Therefore, when weight x_i is applied, the corresponding length,

true length $\leftarrow l_i = \beta_0 + \beta_1 x_i \longrightarrow (1)$

β_0 → initial length of the spring

β_1 → spring constant

Observed / measured length ≠ true length due to errors

Let y_i = observed length corresponding to the weight x_i

= true length + some error

$= l_i + \varepsilon_i$

$\therefore \boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i} \longrightarrow (2)$ Linear Model

y_i → dependent variable

x_i → independent variable

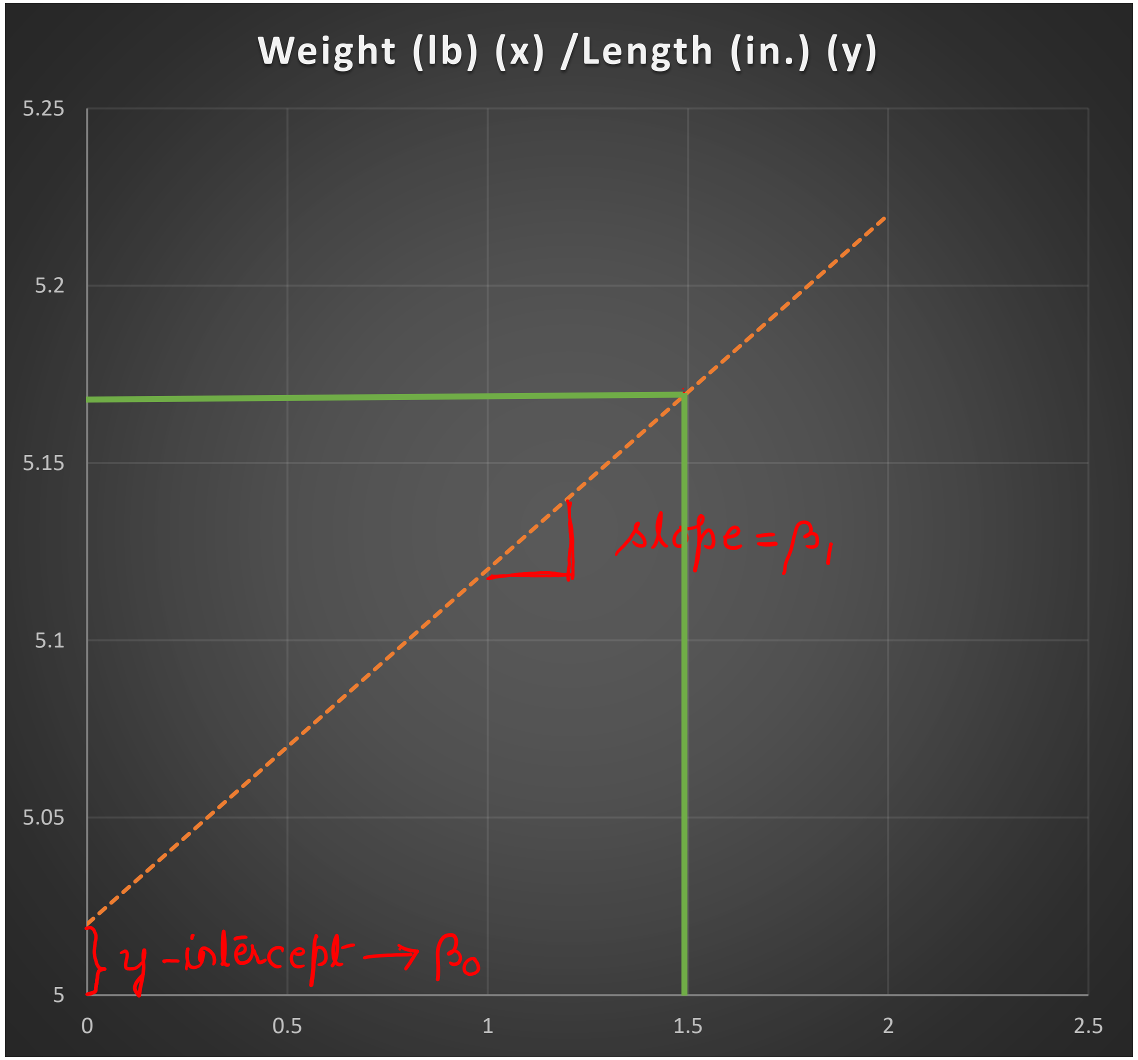
ε_i → error corresponding to weight x_i

β_0 & $\beta_1 \longrightarrow$ regression co-efficients.

STATISTICS FOR DATA SCIENCE

Scenario # 1 : No Errors!!

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



Observed / Measured length = true length

ie, $y_i = l_i$

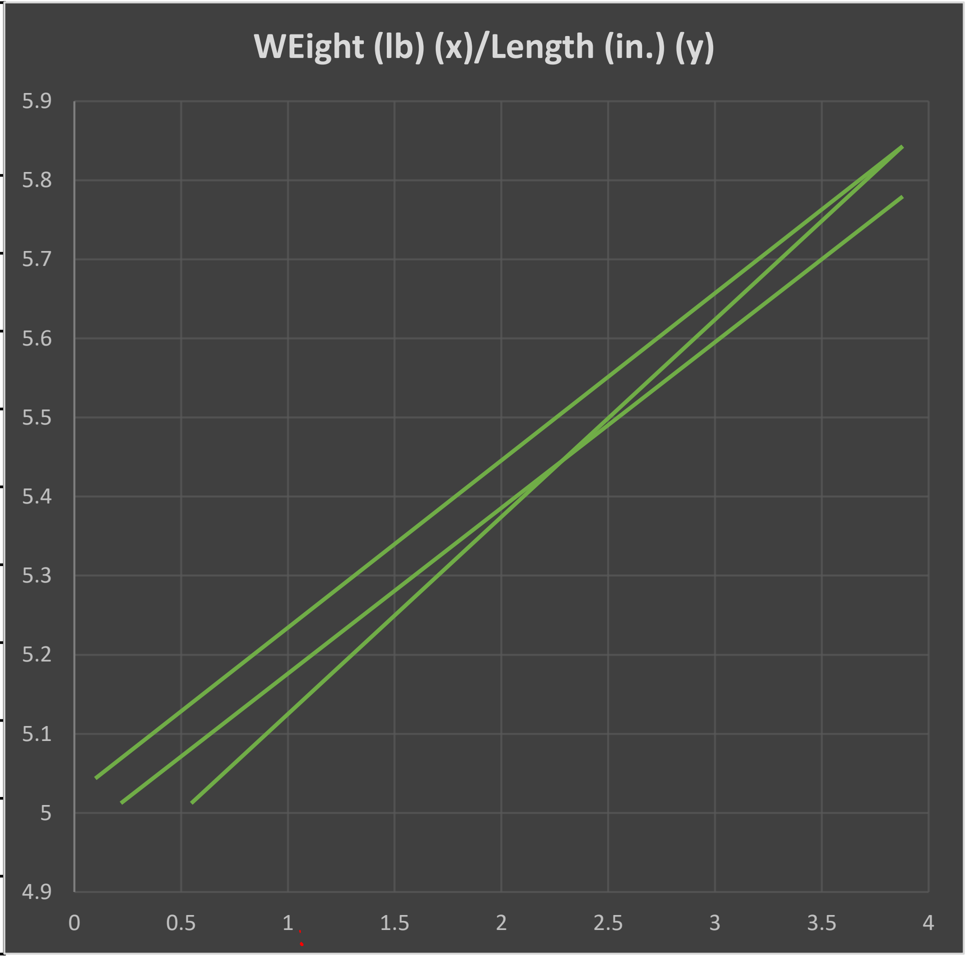
or $y_i = \beta_0 + \beta_1 x_i$

β_0 → y-intercept

β_1 → slope

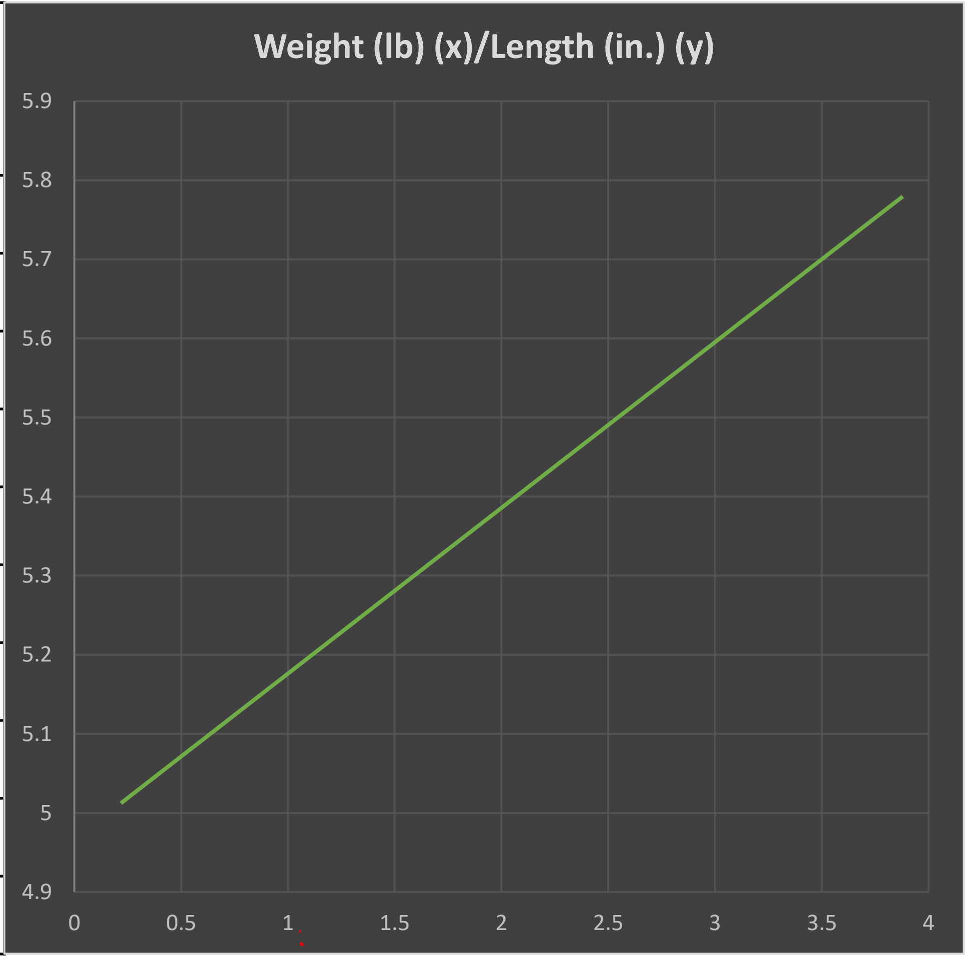
Scenario #2 : Measurement has Errors!!

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



Scenario #2 : Measurement has Errors!!

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



$\sum_{i=1}^{n} e_i^2 \rightarrow$ minimum

Residual

Residual :

❖ $e_i = y_{observed} - y_{predicted}$

$= y_i - \hat{y}_i$

$= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

$e_i \rightarrow$ Residual associated with point (x_i, y_i).

Observed values of y

$\rightarrow$ y values given in data

Predicted values of y

$\rightarrow$ y values on the least square line

Examples:

1) For $x = 1$, $y = 5.16$ and $\hat{y} = 5.21$

$\therefore$ Residual $= y - \hat{y}$

$= 5.16 - 5.21$

$= -0.05$

2) For $x = 3.8$, $y = 5.80$ and $\hat{y} = 5.78$

$\therefore$ Residual $= y - \hat{y}$

$= 5.80 - 5.78$

$= 0.02$.

Least Square Line :

NOTE : The least square line is defined to be the line for which the sum of squared residuals is minimum.

❖ *That is, it is the line for which $\sum_{i=1}^{n} e_i{}^2$ is minimum.*

$$\therefore \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

❖ *Using some Mathematical computations it can be shown that,*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Least Squares Line : Summary

Scenario #1 : If there is no measurement error then the data points lie on the straight line $y = \beta_0 + \beta_1 x$ and values of β_0 and β_1 can be obtained easily by calculating the slope and the intercept.

Scenario #2 : If there is a measurement error ε_i , then

- the exact value of β_0 and β_1 cannot be determined
- the values of β_0 and β_1 are computed by calculating the least square line.
- The least square line is given by $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$

  where

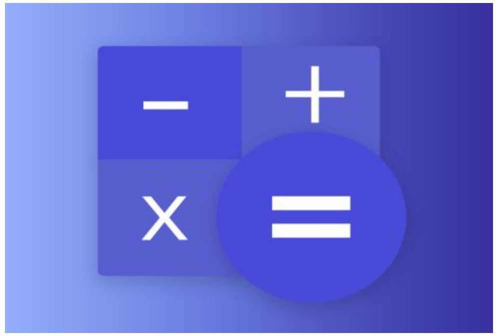
- $\widehat{\beta_0} \rightarrow$ the y $-$ intercept of the least square line

  $\rightarrow$ gives an estimate of β_0 , the initial length of the spring.

- $\widehat{\beta_1}$ $\rightarrow$the slope of the least square line

  $\rightarrow$ gives an estimate of the actual value of the spring constant β_1 .

Computing formulas



Remark :

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

- $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$

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

For computational purposes we use the equivalent formula that is specified in the RHS.

Sources :uplabs.com

Try This !!!

Using the Hooke's law data given in the table

i. Compute the least squares estimates of the spring constant and the unloaded length of the spring.

ii. Write the equation of the least squares line.

iii. Estimate the length of the spring under a load of 1.3 lb.

iv. Estimate the length of the spring under a load of 1.4 lb.

v. Obtain the Residuals corresponding to all the points (x_i, y_i).

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

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.

THANK YOU

Dr. Karthiyayini
Department of Science & Humanities

Karthiyayini.roy@pes.edu

+91 80 6618 6651