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

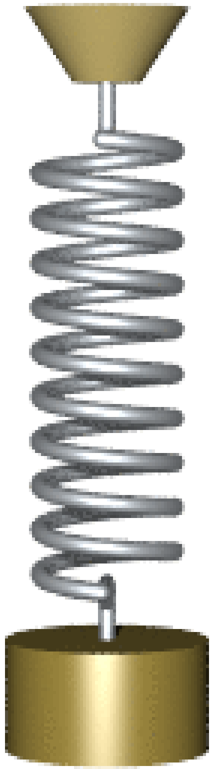
- How to compute the Least Squares Line
- Residuals and Errors
- Measuring Goodness of fit

How to compute the Least – Squares Line ???

- Consider a spring that is hung vertically with the top end fixed.
- Let some weights be hung one at a time at the other end .
- Let x_i = the weights
- Let l_i = length of the spring due to the load x_i
- Then by the Hooke's law we have,

$$l_i = \beta_0 + \beta_1 x_i \text{…………….}(1)$$

where β_0 = the initial length of the spring when the spring is not loaded and β_1 = spring constant.



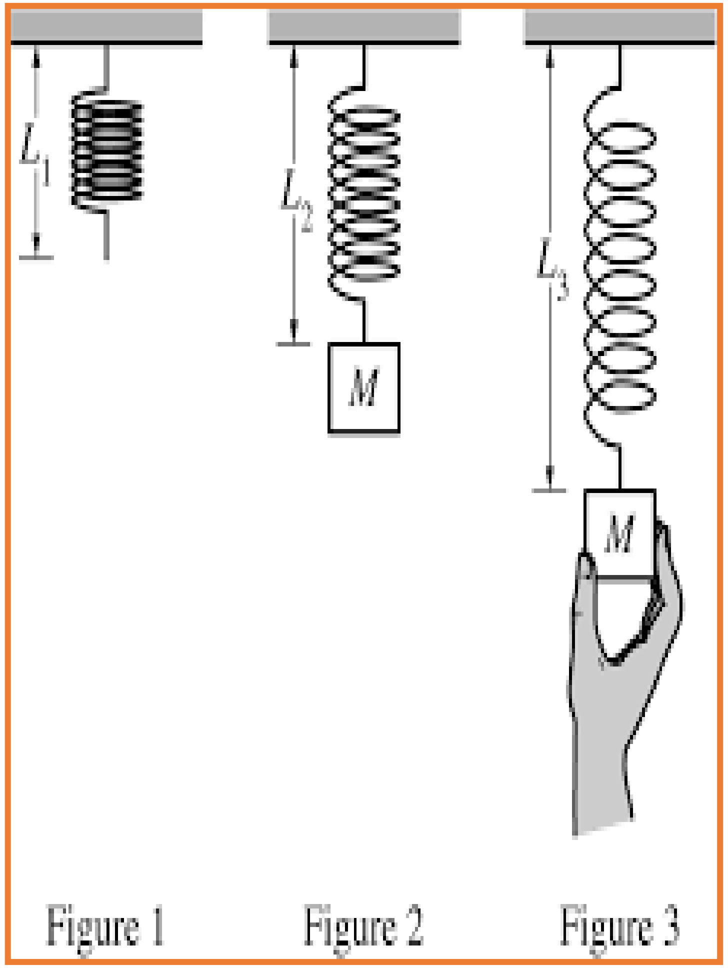
Sources : en.wikipedia.org/wiki/Oscillation

How to compute the Least – Squares Line ???

- Let y_i = measured length of the spring due to the load x_i
- Then $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$……….(2) where ε_i= the error in the ith measurement.

Equation (2) is called as a ***Linear Model*** where

- x_i is the independent variable
- y_i is the dependent variable
- β_0 and β_1 are the regression coefficients
- ε_i is the error



Source : Internet

How to compute the Least – Squares Line ???

- Let y_i = measured length of the spring due to the load x_i
- Then,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ldots\ldots\ldots(2)$$

where ε_i= the error in the ith measurement.

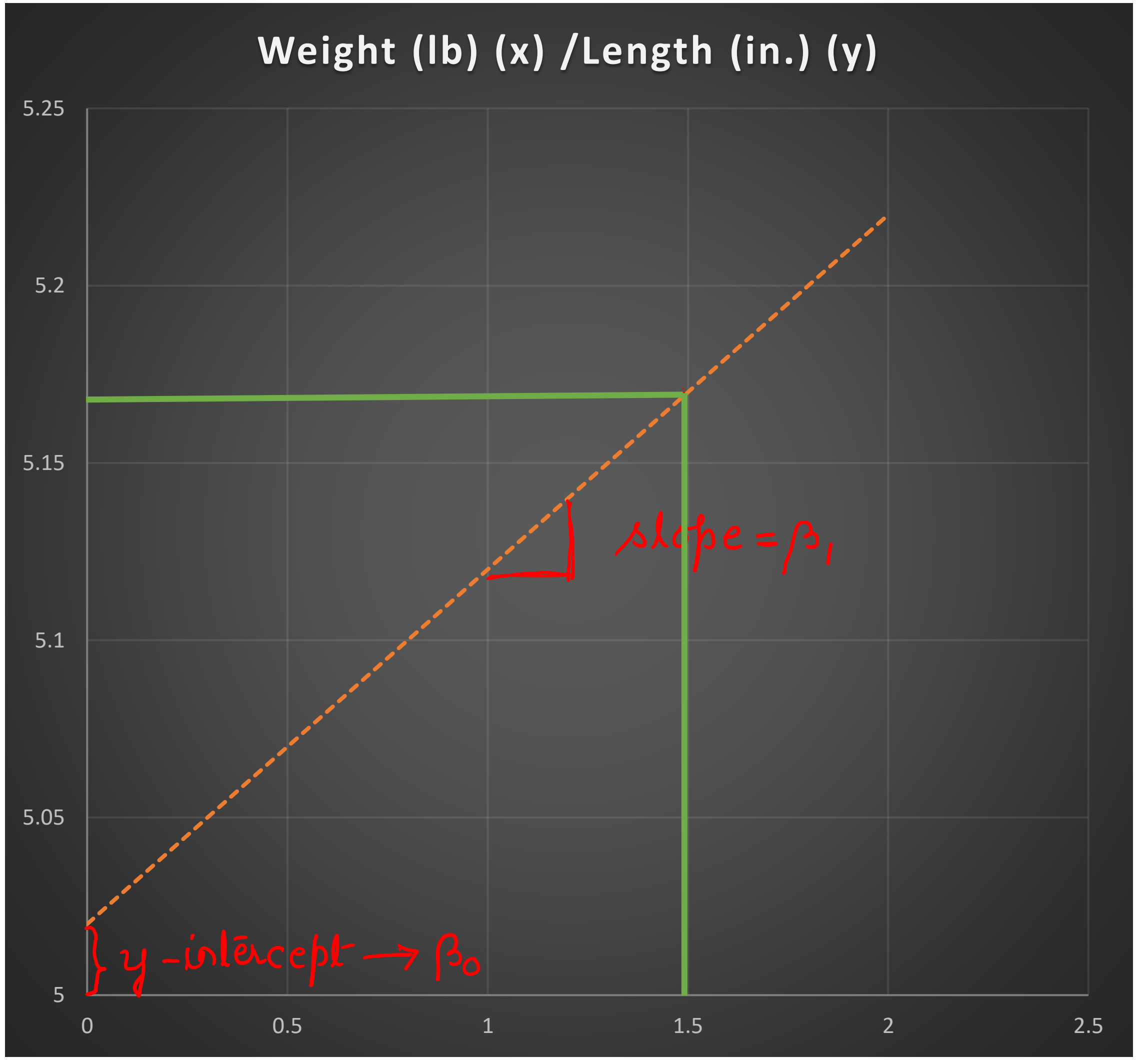
Equation (2) is called as a ***Linear Model***

where

- x_i is the independent variable
- y_i is the dependent variable
- β_0 and β_1 are the regression coefficients
- ε_i is the error

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



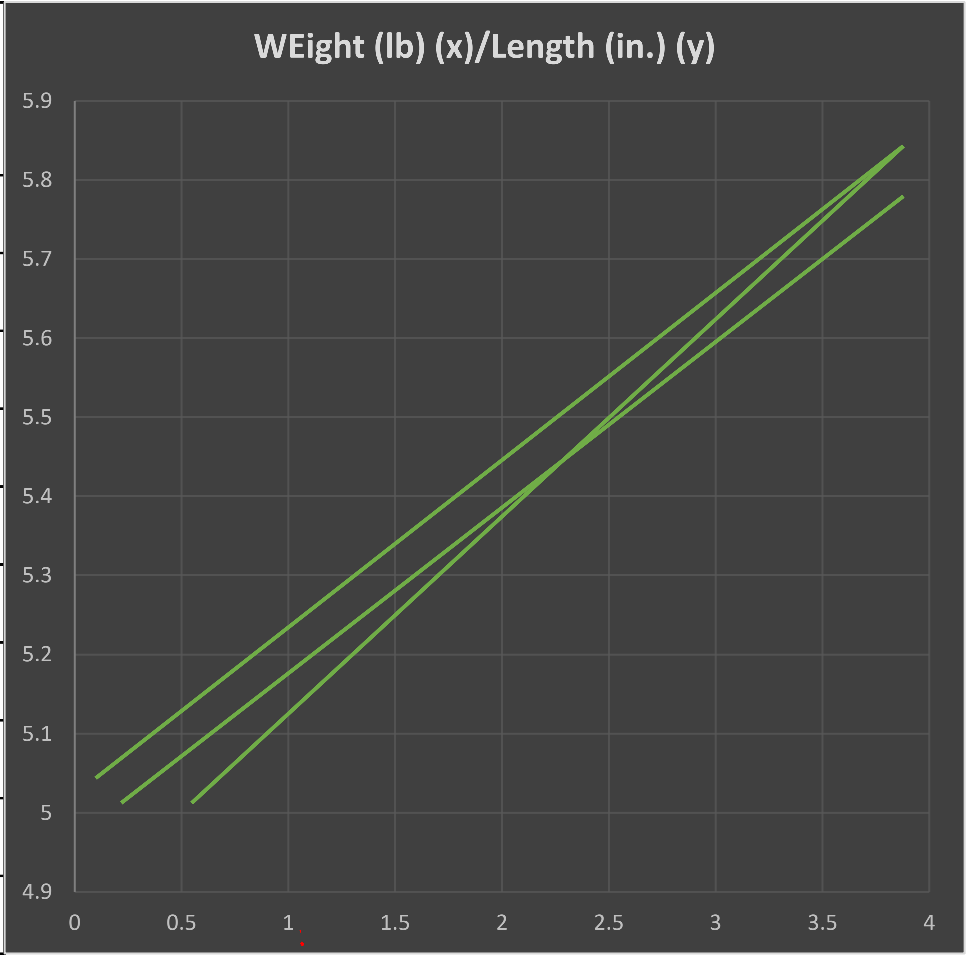
- If there are no errors in measuring then, all the data points lie on a straight line and observed length/ measured length = True length
- That is, $y_i = l_i$
  $\Rightarrow y_i = \beta_0 + \beta_1 x_i$

  where β_0 is the y – intercept

  and β_1 is the slope.

Scenario #2 : Measurement has Errors!!

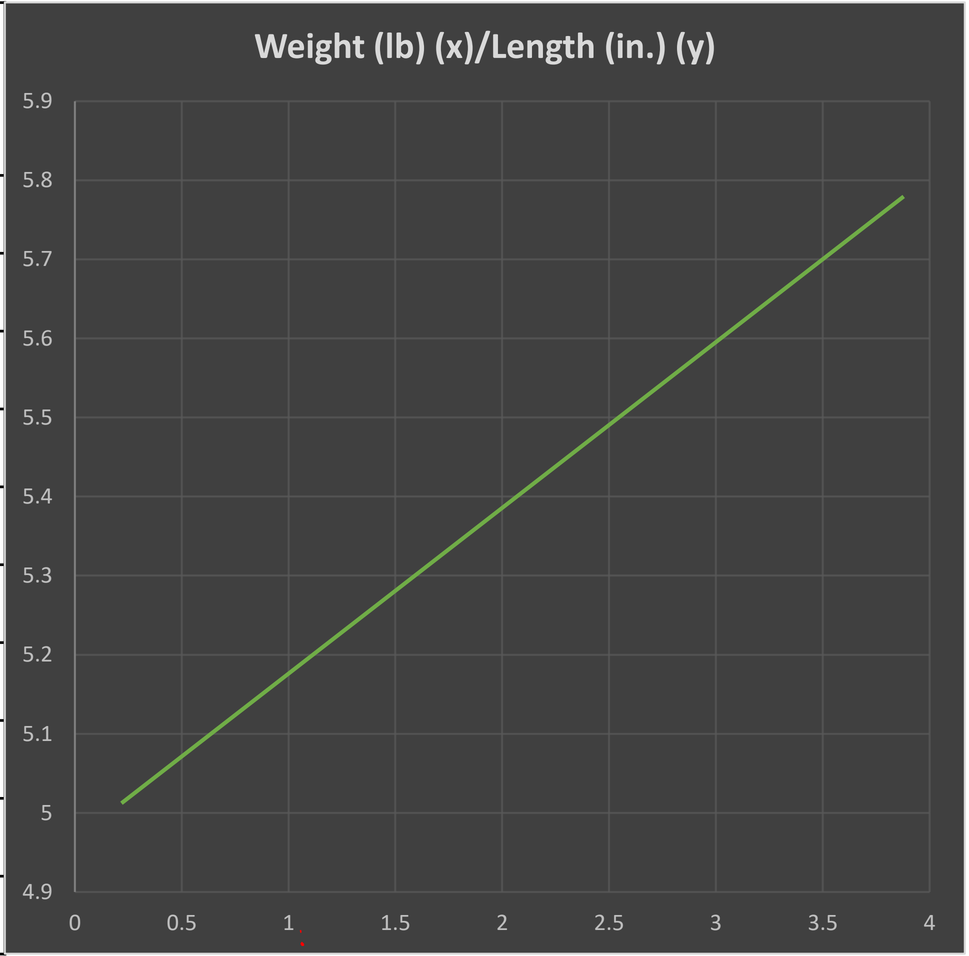
Weight (lb) (x)	Length $(in.)$ (y)	Weight (lb) (x)	Length $(in.)$ (y)
0.0	5.06	2.0	5.40
0.2	5.01	2.2	5.57
0.4	5.12	2.4	5.47
0.6	5.13	2.6	5.53
0.8	5.14	2.8	5.61
1.0	5.16	3.0	5.59
1.2	5.25	3.2	5.61
1.4	5.19	3.4	5.75
1.6	5.24	3.6	5.68
1.8	5.46	3.8	5.80



- ❖ If there are errors in measuring then the data points do not lie on a straight line.
- ❖ Different people may draw the line at various places as shown in the Scatter which may lead to a difference in the analysis made.
- ❖ Hence we have the least squares line.

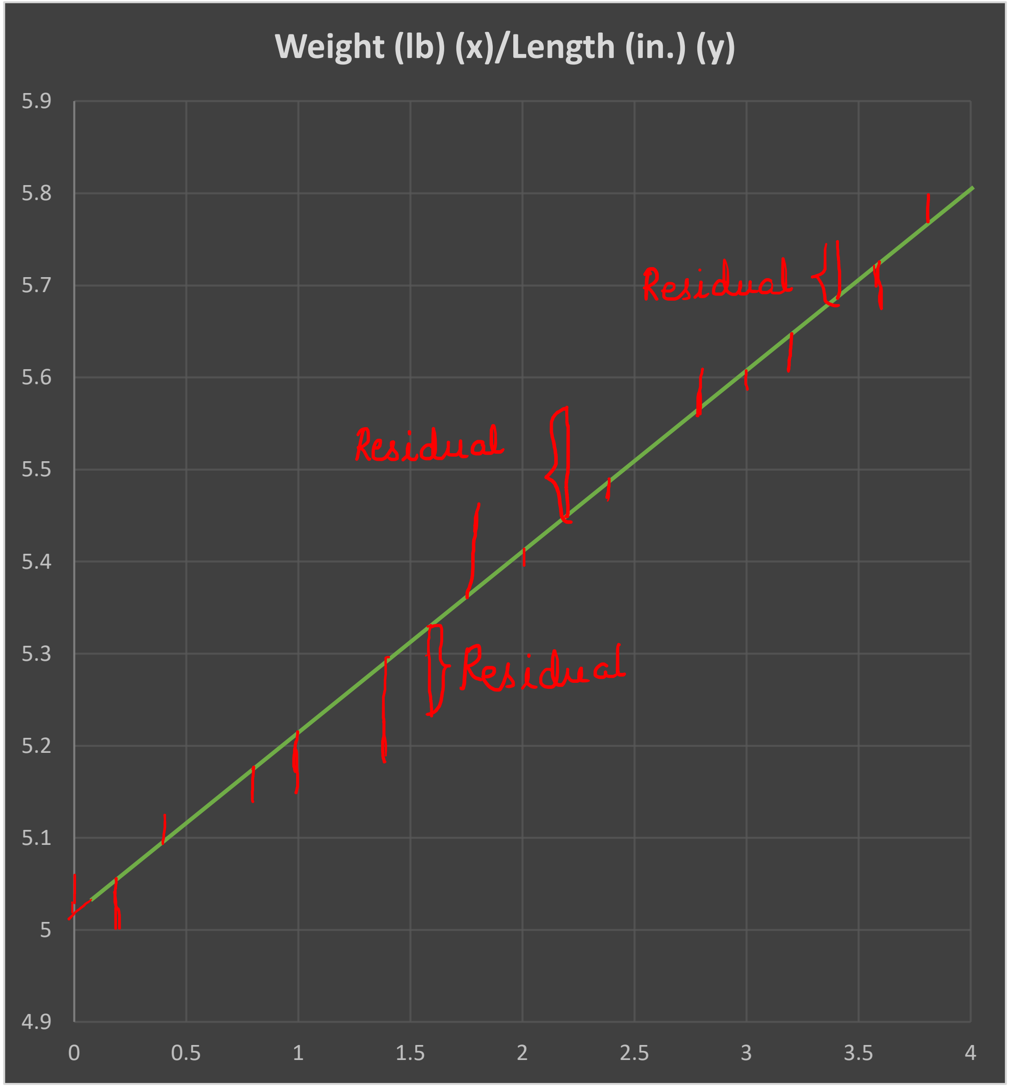
Scenario #2 : Measurement has Errors!!

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



- ❖ The least squares line is the line for which the sum of squares of the residual is minimum.
- ❖ That is $\sum_{i=1}^{n} e_i^{\,2}$ is minimum where e_i is the residual.

Residual :



❖The residual is defined as the difference in the observed and the predicted values of y.

❖That is, $e_i = y_{observed} - y_{predicted}$

$$= y_i - \widehat{y_i}$$

$$= y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i$$

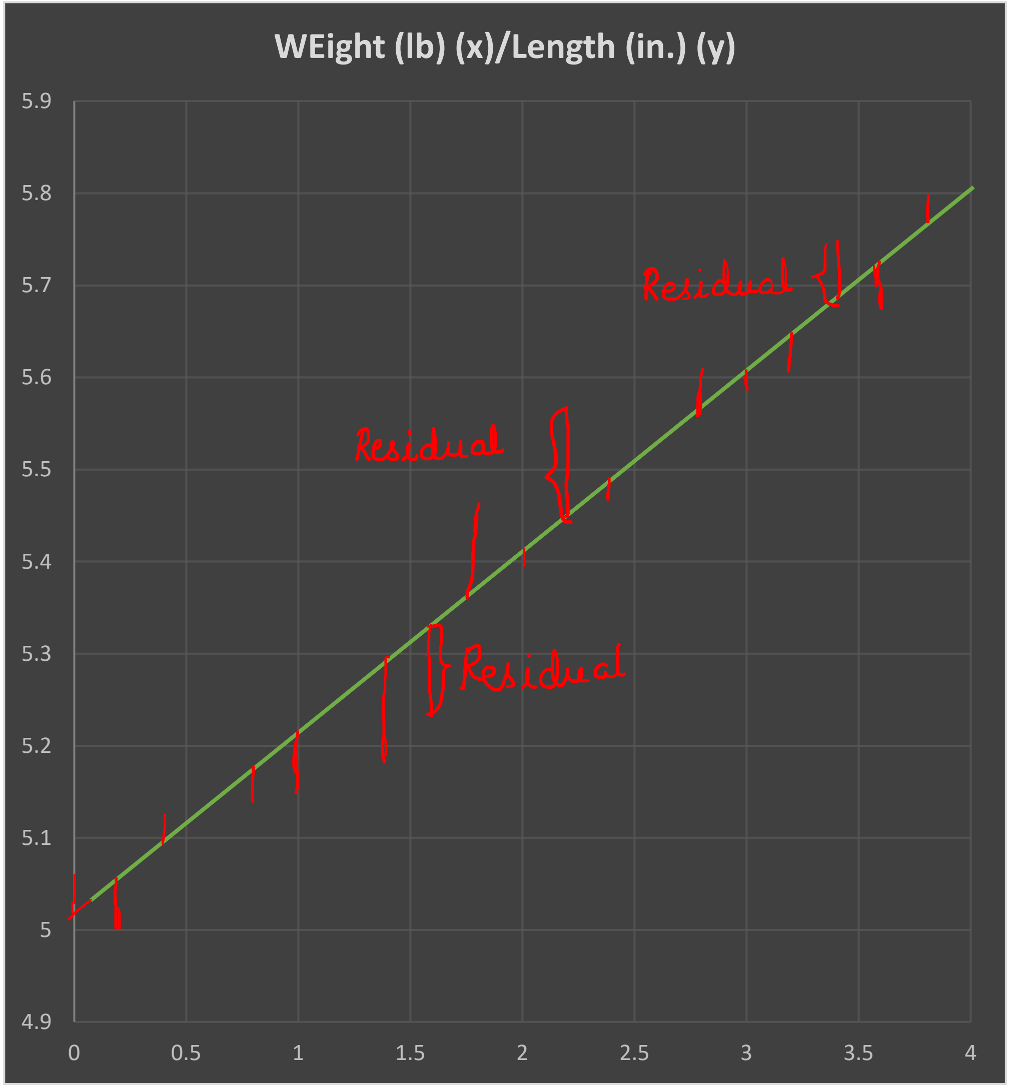
❖ Here e_i is called as the residual associated with the point (x_i, y_i)

Examples :

1. For $x = 1, y = 5.16\ and\ \hat{y} = 5.21,$
   Residual = $y - \hat{y} = 5.16 - 5.21 = -0.05$

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

$$\Rightarrow \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2$$

❖Using some Mathematical computations it can be shown that,

- $\widehat{\boldsymbol{\beta}_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $\widehat{\boldsymbol{\beta}_0} = \bar{y} - \widehat{\boldsymbol{\beta}_1}\bar{x}$

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

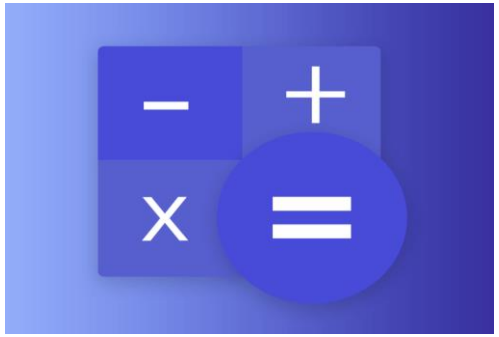
- $\widehat{\beta_0}$ → the y − intercept of the least square line

  → gives an estimate of β_0 , the initial length of the spring.

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

  → 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