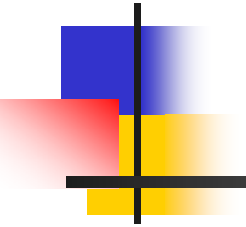




Linear Regression



Major: Industrial Engineering

Authors: Autar Kaw, Luke Snyder

What is Regression?

What is regression? Given n data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ best fit $y = f(x)$ to the data. The best fit is generally based on minimizing the sum of the square of the residuals, S_r .

Residual at a point is

$$\varepsilon_i = y_i - f(x_i)$$

Sum of the square of the residuals

$$S_r = \sum_{i=1}^n (y_i - f(x_i))^2$$

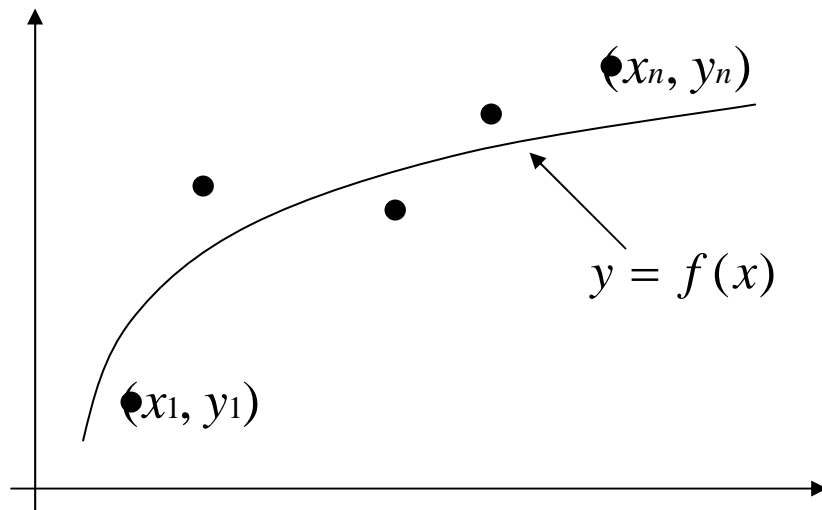


Figure. Basic model for regression

Linear Regression-Criterion#1

Given n data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ best fit $y = a_0 + a_1 x$ to the data.

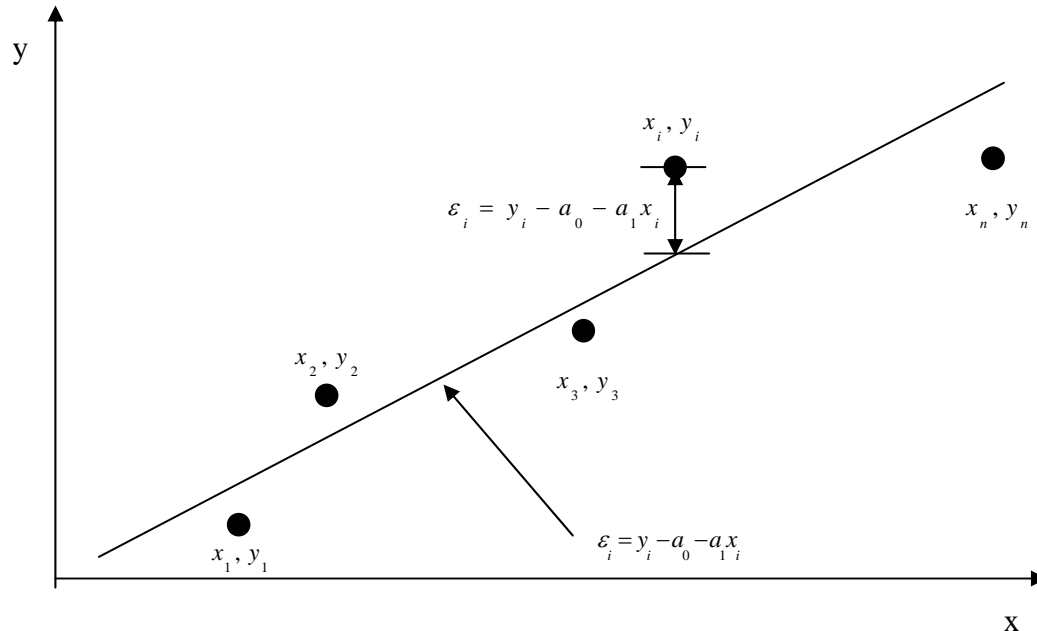


Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i .

Does minimizing $\sum_{i=1}^n \varepsilon_i$ work as a criterion, where $\varepsilon_i = y_i - (a_0 + a_1 x_i)$

Example for Criterion#1

Example: Given the data points $(2,4)$, $(3,6)$, $(2,6)$ and $(3,8)$, best fit the data to a straight line using Criterion#1

Table. Data Points

x	y
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0

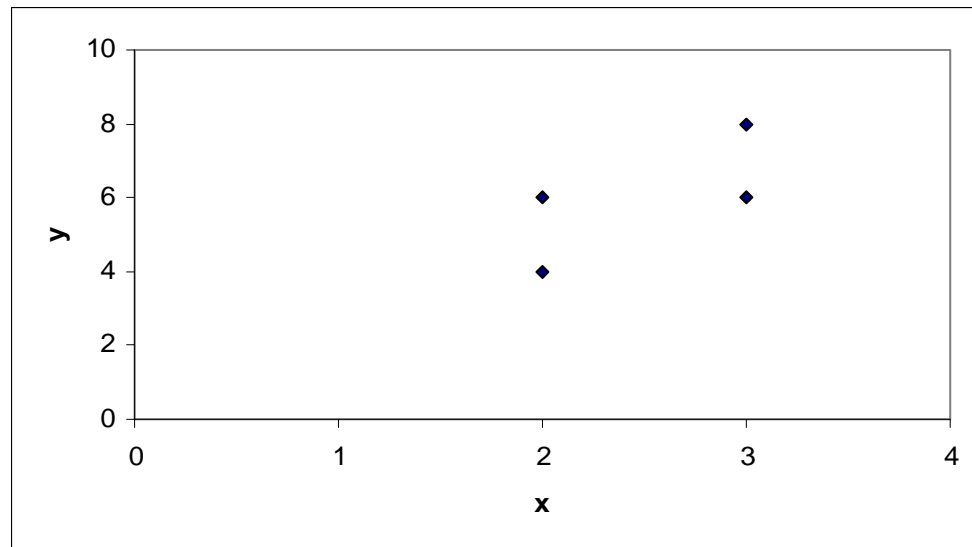


Figure. Data points for y vs. x data.

Linear Regression-Criteria#1

Using $y=4x-4$ as the regression curve

Table. Residuals at each point for regression model $y = 4x - 4$.

x	y	$y_{\text{predicted}}$	$\varepsilon = y - y_{\text{predicted}}$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	-2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
			$\sum_{i=1}^4 \varepsilon_i = 0$

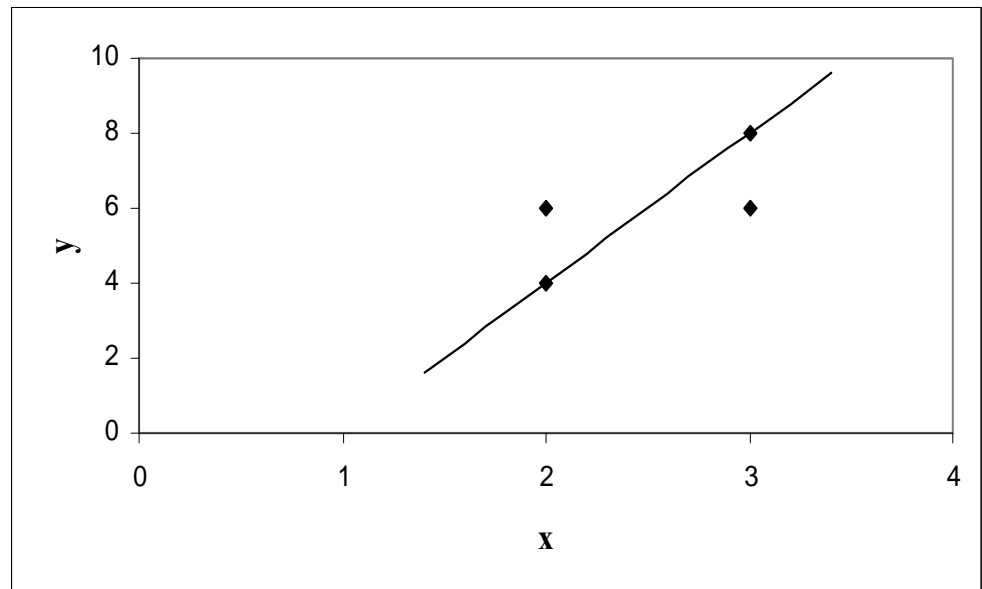


Figure. Regression curve for $y=4x-4$, y vs. x data

Linear Regression-Criteria#1

Using $y=6$ as a regression curve

Table. Residuals at each point for $y=6$

x	y	$y_{\text{predicted}}$	$\varepsilon = y - y_{\text{predicted}}$
2.0	4.0	6.0	-2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
			$\sum_{i=1}^4 \varepsilon_i = 0$

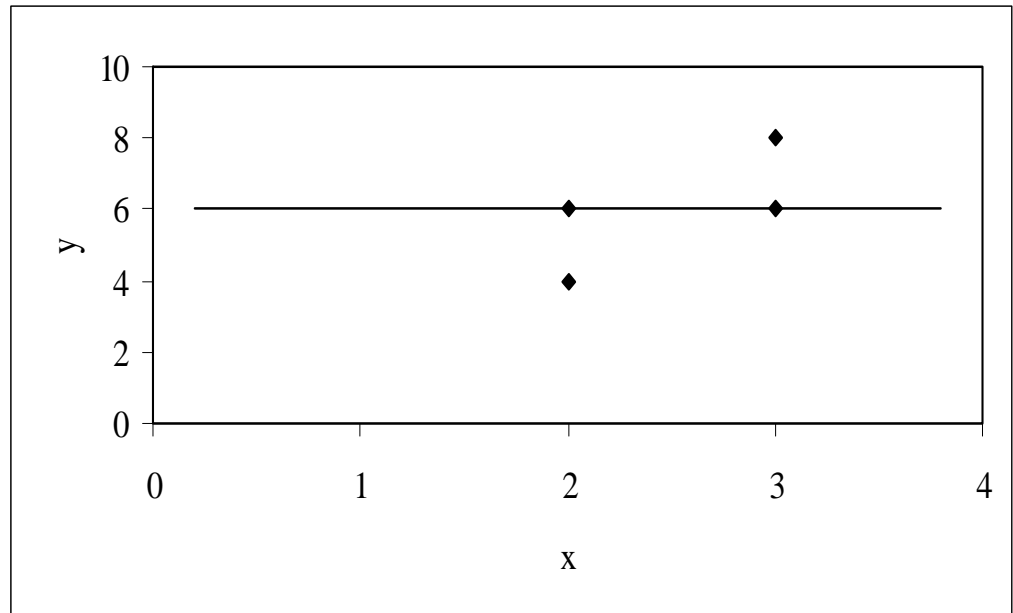


Figure. Regression curve for $y=6$, y vs. x data



Linear Regression – Criterion #1

$$\sum_{i=1}^4 \varepsilon_i = 0 \quad \text{for both regression models of } y=4x-4 \text{ and } y=6.$$

The sum of the residuals is as small as possible, that is zero, but the regression model is not unique.

Hence the above criterion of minimizing the sum of the residuals is a bad criterion.

Linear Regression-Criterion#2

Will minimizing $\sum_{i=1}^n |\epsilon_i|$ work any better?

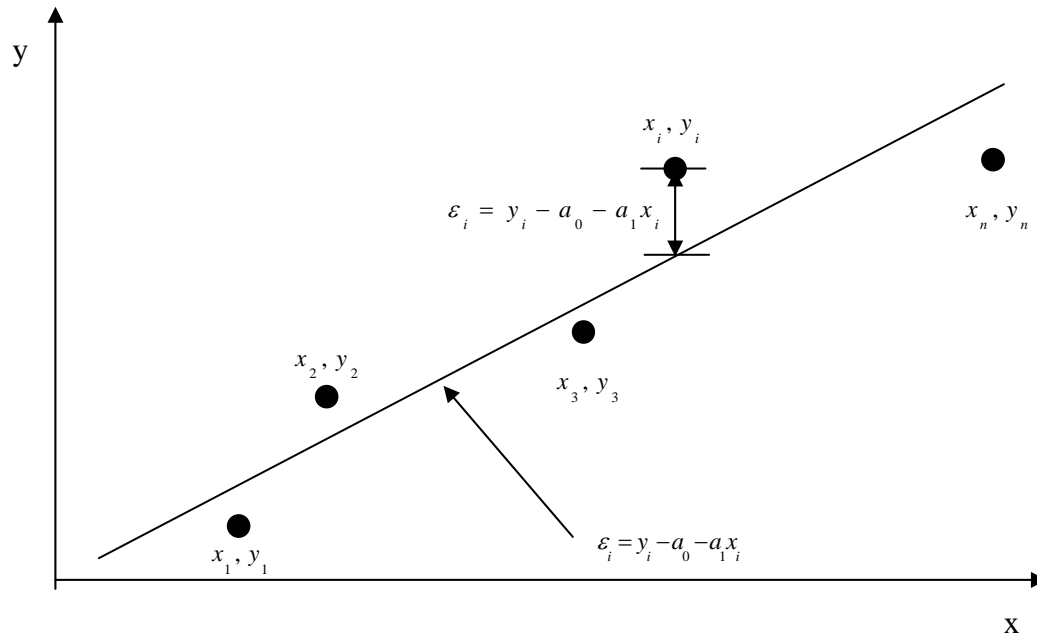


Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i .

Linear Regression-Criteria 2

Using $y=4x-4$ as the regression curve

Table. The absolute residuals employing the $y=4x-4$ regression model

x	y	$y_{\text{predicted}}$	$ \varepsilon = y - y_{\text{predicted}} $
2.0	4.0	4.0	0.0
3.0	6.0	8.0	2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
			$\sum_{i=1}^4 \varepsilon_i = 4$

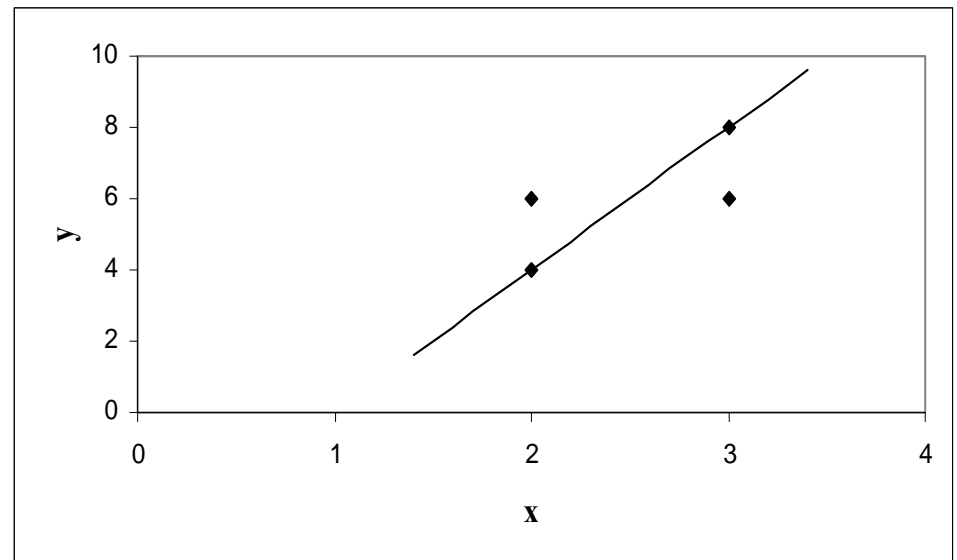


Figure. Regression curve for $y=4x-4$, y vs. x data

Linear Regression-Criteria#2

Using $y=6$ as a regression curve

Table. Absolute residuals employing the $y=6$ model

x	y	$y_{\text{predicted}}$	$ \varepsilon = y - y_{\text{predicted}} $
2.0	4.0	6.0	2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
			$\sum_{i=1}^4 \varepsilon_i = 4$

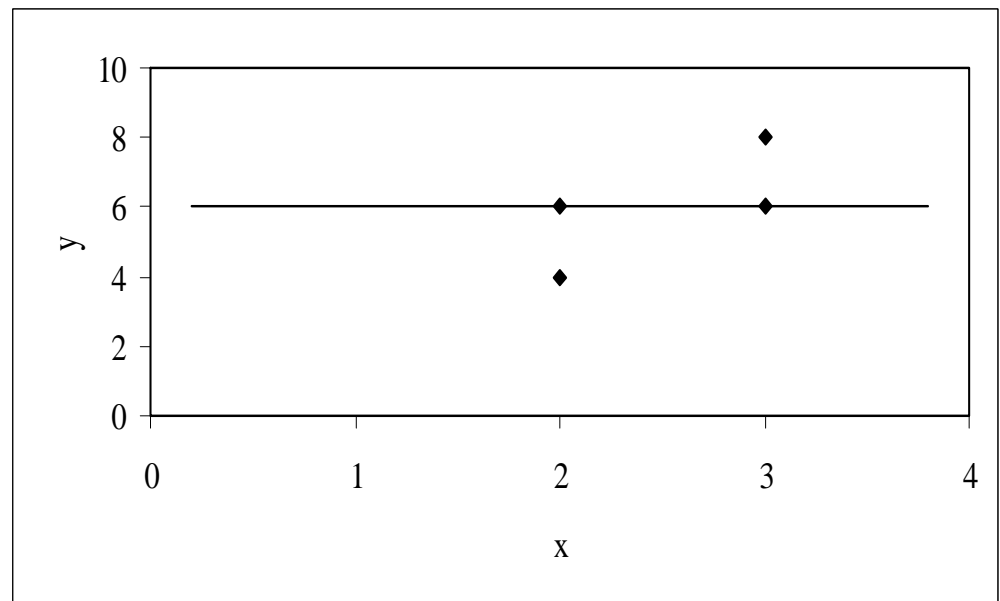


Figure. Regression curve for $y=6$, y vs. x data



Linear Regression-Criterion#2

$$\sum_{i=1}^4 |\varepsilon_i| = 4 \text{ for both regression models of } y=4x-4 \text{ and } y=6.$$

The sum of the errors has been made as small as possible, that is 4, but the regression model is not unique.

Hence the above criterion of minimizing the sum of the absolute value of the residuals is also a bad criterion.

Can you find a regression line for which $\sum_{i=1}^4 |\varepsilon_i| < 4$ and has unique regression coefficients?

Least Squares Criterion

The least squares criterion minimizes the sum of the square of the residuals in the model, and also produces a unique line.

$$S_r = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2$$

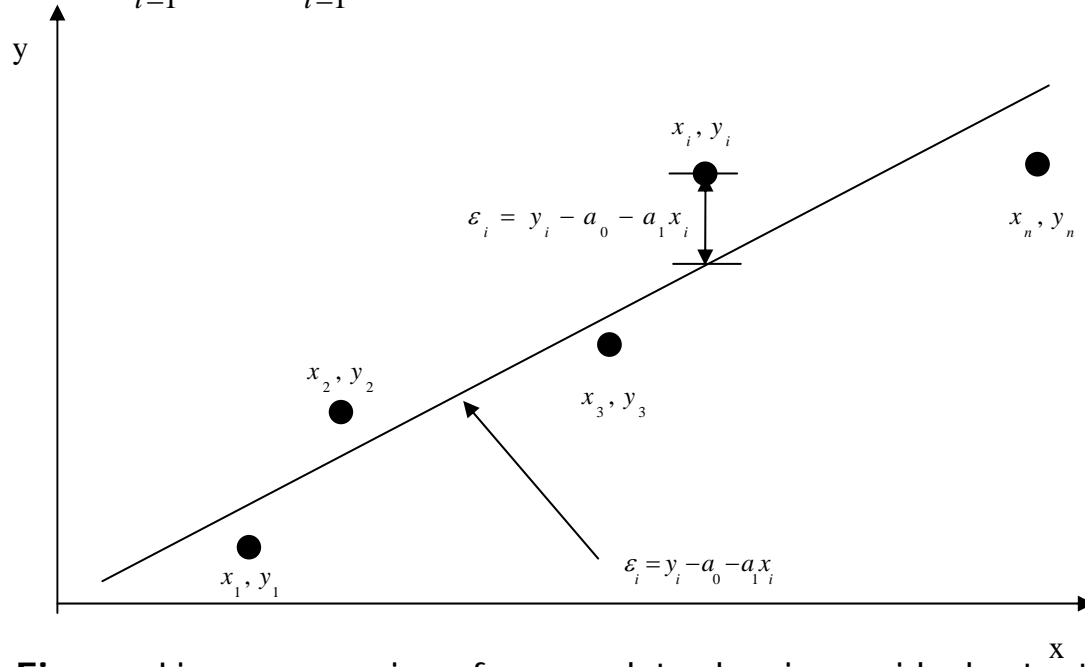


Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i .



Finding Constants of Linear Model

Minimize the sum of the square of the residuals: $S_r = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2$

To find a_0 and a_1 we minimize S_r with respect to a_1 and a_0 .

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i)(-1) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i)(-x_i) = 0$$

giving

$$\sum_{i=1}^n a_0 + \sum_{i=1}^n a_1 x_i = \sum_{i=1}^n y_i$$

$$(a_0 = \bar{y} - a_1 \bar{x})$$

$$\sum_{i=1}^n a_0 x_i + \sum_{i=1}^n a_1 x_i^2 = \sum_{i=1}^n y_i x_i$$



Finding Constants of Linear Model

Solving for a_0 and a_1 directly yields,

$$a_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

and

$$a_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad (a_0 = \bar{y} - a_1 \bar{x})$$

Example 1

As machines are used over long periods of time, the output product can get off target. Below is the average value of how much off target a product is getting manufactured as a function of machine use.

Table. Data points for h vs. t

Hours of Machine Use, t	Millimeters Off Target, h
30	1.10
33	1.21
34	1.25
35	1.23
39	1.30
44	1.40
45	1.42

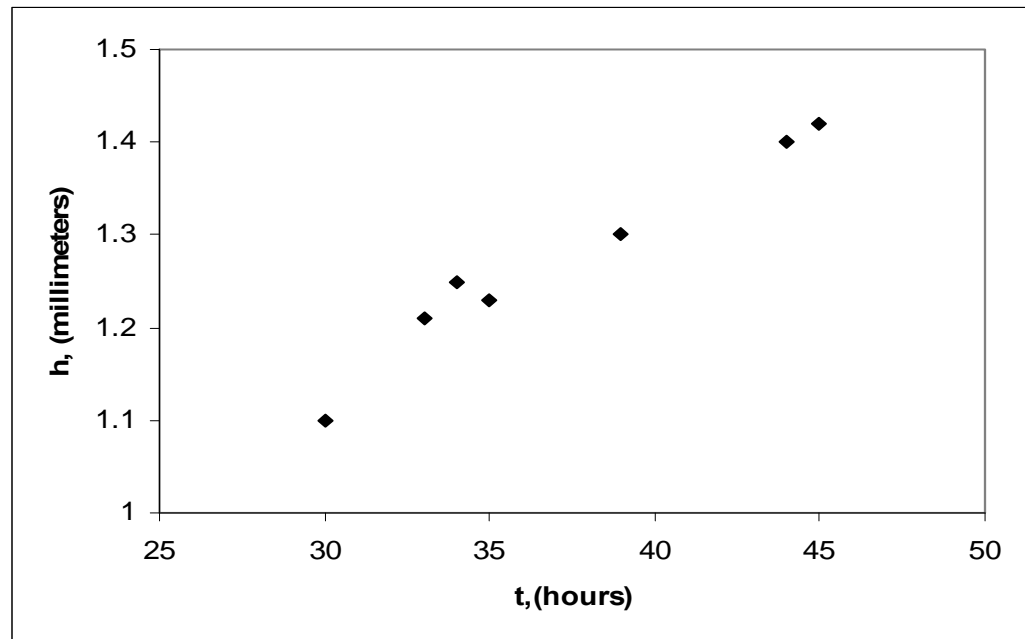


Figure. Data points for h vs. t data

Example 1 cont.

Regress the data to $h = a_0 + a_1 t$ and find when the product will be 2mm off target.

Table. Summation data for linear regression

t	h	t^2	$t \times h$
<i>Hours</i>	<i>Millimeter</i> <i>s</i>	<i>Hours²</i>	<i>Millimeter-Hour</i>
30	1.10	900	33
33	1.21	1089	39.93
34	1.25	1156	42.50
35	1.23	1225	43.05
39	1.30	1521	50.70
44	1.40	1936	61.6
45	1.42	2025	63.9
$\sum_{i=1}^7$ 263	8.91	9852	334.67

With $n = 7$

$$\begin{aligned}
 a_1 &= \frac{n \sum_{i=1}^7 t_i h_i - \sum_{i=1}^7 t_i \sum_{i=1}^7 h_i}{n \sum_{i=1}^7 t_i^2 - \left(\sum_{i=1}^7 t_i \right)^2} \\
 &= \frac{7(334.67) - (263)(8.91)}{7(9852) - (263)^2} \\
 &= 0.003122 \text{ mm-h}
 \end{aligned}$$

$\sum_{i=1}^7$



Example 1 cont.

The value for a_0 can then be found using $a_0 = \bar{h} - a_1 \bar{t}$ where

$$\bar{h} = \frac{\sum_{i=1}^7 h_i}{n} = 1.2728 \text{ mm} \qquad \bar{t} = \frac{\sum_{i=1}^7 t_i}{n} = 37.57 \text{ hours}$$

$$\begin{aligned} a_0 &= \bar{h} - a_1 \bar{t} \\ &= 1.2728 - (0.003122)(37.57) \\ &= 1.15551 \text{ mm-h} \end{aligned}$$

Example 1 cont.

The linear regression model is now given by $h = 1.15551 + 0.003122 \times t$

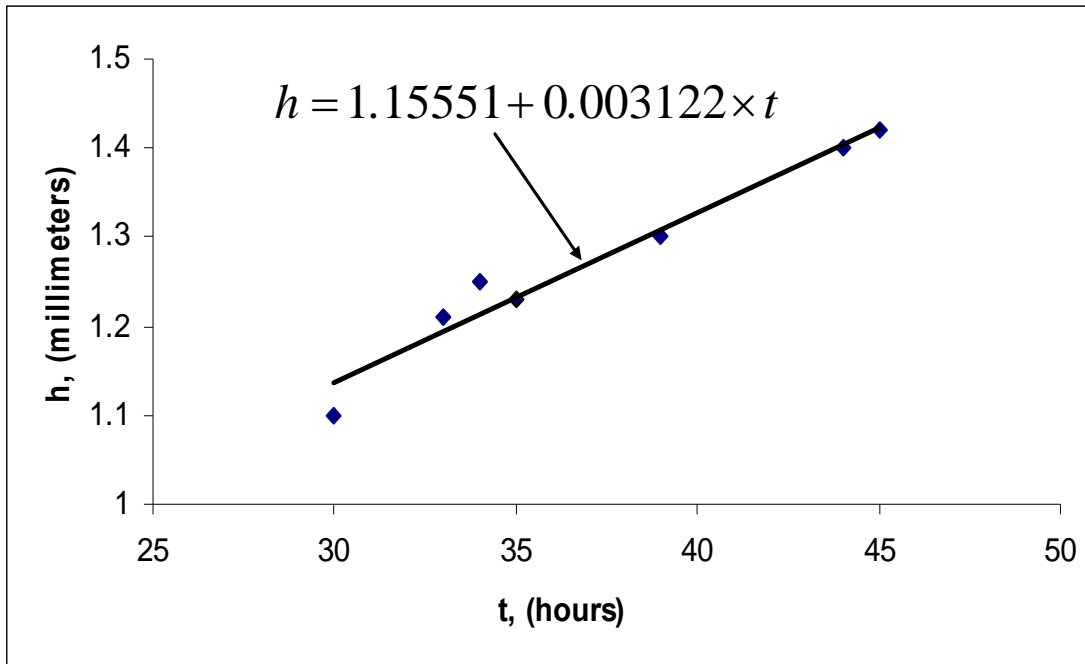


Figure. Linear regression of hours of use vs. millimeters off target.



Example 1 cont.

Solving for when $h = 2mm$ yields

$$h = 1.15551 + 0.003122 \times t$$

$$2 = 1.15551 + 0.003122 \times t$$

$$t = \frac{2 - 1.15551}{0.003122}$$

$$t = 270.496 \text{ hours}$$

Example 2

To find the longitudinal modulus of composite, the following data is collected. Find the longitudinal modulus, E using the regression model

Table. Stress vs. Strain data

Strain (%)	Stress (MPa)
0	0
0.183	306
0.36	612
0.5324	917
0.702	1223
0.867	1529
1.0244	1835
1.1774	2140
1.329	2446
1.479	2752
1.5	2767
1.56	2896

$\sigma = E\varepsilon$ and the sum of the square of the residuals.

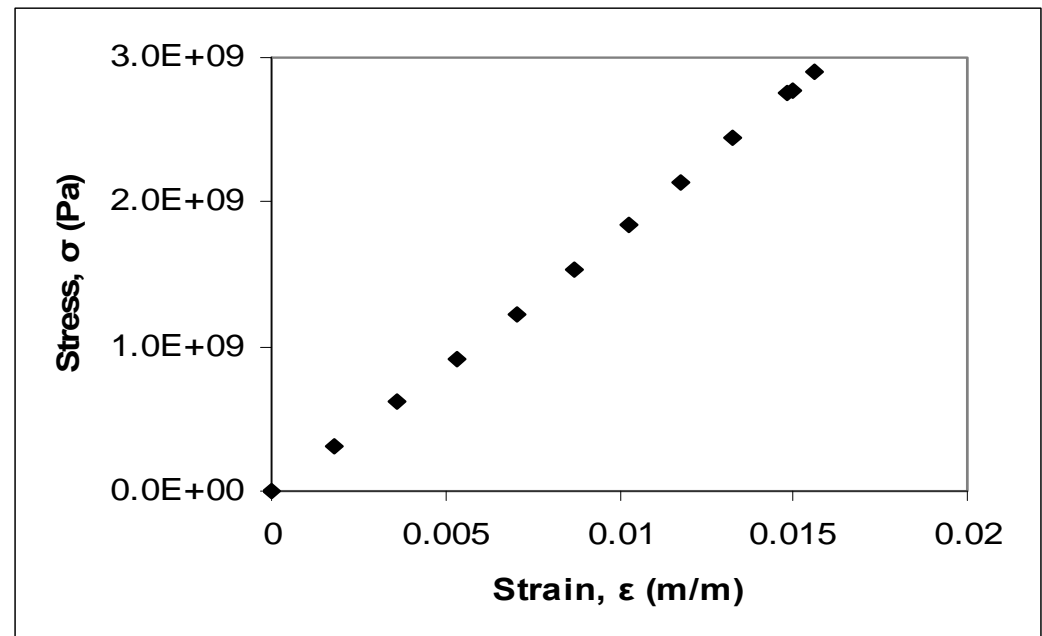


Figure. Data points for Stress vs. Strain data



Example 2 cont.

Residual at each point is given by

$$\gamma_i = \sigma_i - E\varepsilon_i$$

The sum of the square of the residuals then is

$$\begin{aligned} S_r &= \sum_{i=1}^n \gamma_i^2 \\ &= \sum_{i=1}^n (\sigma_i - E\varepsilon_i)^2 \end{aligned}$$

Differentiate with respect to E

$$\frac{\partial S_r}{\partial E} = \sum_{i=1}^n 2(\sigma_i - E\varepsilon_i)(-\varepsilon_i) = 0$$

Therefore

$$E = \frac{\sum_{i=1}^n \sigma_i \varepsilon_i}{\sum_{i=1}^n \varepsilon_i^2}$$

Example 2 cont.

Table. Summation data for regression model

i	ϵ	σ	ϵ^2	$\epsilon\sigma$
1	0.0000	0.0000	0.0000	0.0000
2	1.8300×10^{-3}	3.0600×10^8	3.3489×10^{-6}	5.5998×10^5
3	3.6000×10^{-3}	6.1200×10^8	1.2960×10^{-5}	2.2032×10^6
4	5.3240×10^{-3}	9.1700×10^8	2.8345×10^{-5}	4.8821×10^6
5	7.0200×10^{-3}	1.2230×10^9	4.9280×10^{-5}	8.5855×10^6
6	8.6700×10^{-3}	1.5290×10^9	7.5169×10^{-5}	1.3256×10^7
7	1.0244×10^{-2}	1.8350×10^9	1.0494×10^{-4}	1.8798×10^7
8	1.1774×10^{-2}	2.1400×10^9	1.3863×10^{-4}	2.5196×10^7
9	1.3290×10^{-2}	2.4460×10^9	1.7662×10^{-4}	3.2507×10^7
10	1.4790×10^{-2}	2.7520×10^9	2.1874×10^{-4}	4.0702×10^7
11	1.5000×10^{-2}	2.7670×10^9	2.2500×10^{-4}	4.1505×10^7
12	1.5600×10^{-2}	2.8960×10^9	2.4336×10^{-4}	4.5178×10^7
$\sum_{i=1}^{12}$			1.2764×10^{-3}	2.3337×10^8

With

$$\sum_{i=1}^{12} \epsilon_i^2 = 1.2764 \times 10^{-3}$$

and

$$\sum_{i=1}^{12} \sigma_i \epsilon_i = 2.3337 \times 10^8$$

Using

$$\begin{aligned}
 E &= \frac{\sum_{i=1}^{12} \sigma_i \epsilon_i}{\sum_{i=1}^{12} \epsilon_i^2} \\
 &= \frac{2.3337 \times 10^8}{1.2764 \times 10^{-3}} \\
 &= 182.83 \text{ GPa}
 \end{aligned}$$

Example 2 Results

The equation $\sigma = 182.823\varepsilon$ describes the data.

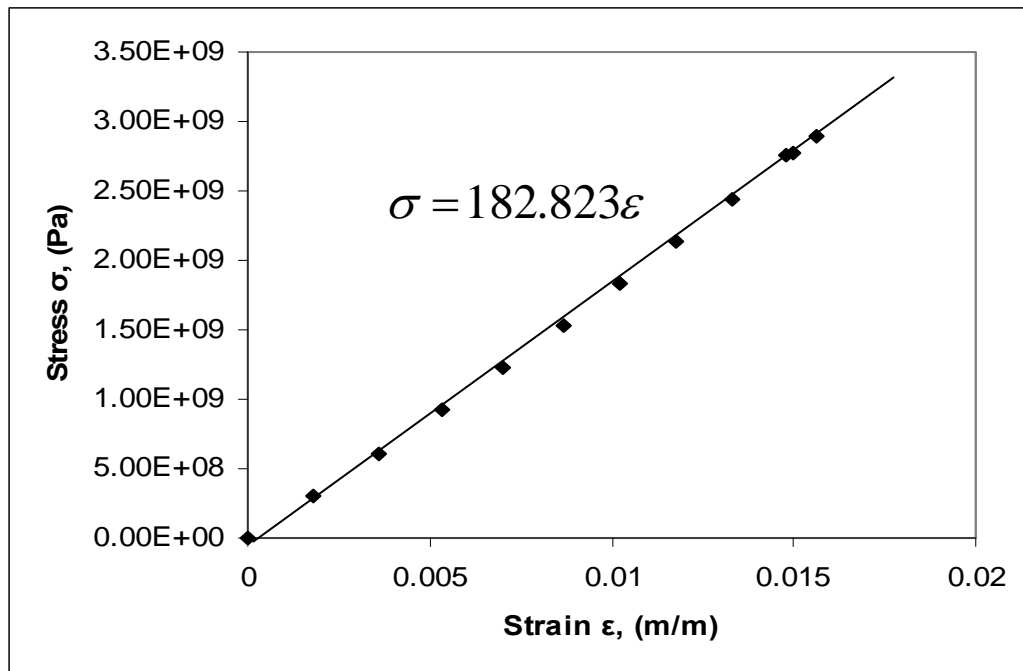


Figure. Linear regression for Stress vs. Strain data