Linear Regression

Civil Engineering Majors

Authors: Autar Kaw, Luke Snyder

http://numericalmethods.eng.usf.edu

Transforming Numerical Methods Education for STEM Undergraduates

Linear Regression

http://numericalmethods.eng.usf.edu

What is Regression?

What is regression? Given n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 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

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

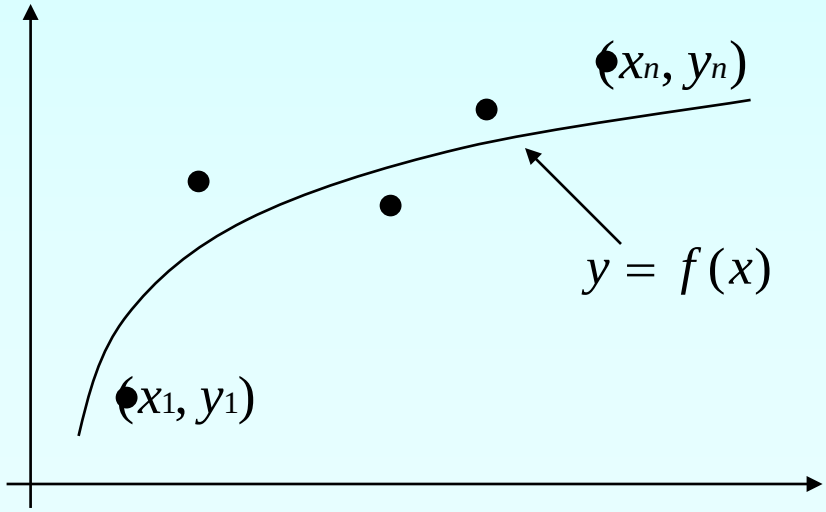

Figure. Basic model for regression

Linear Regression-Criterion#1

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

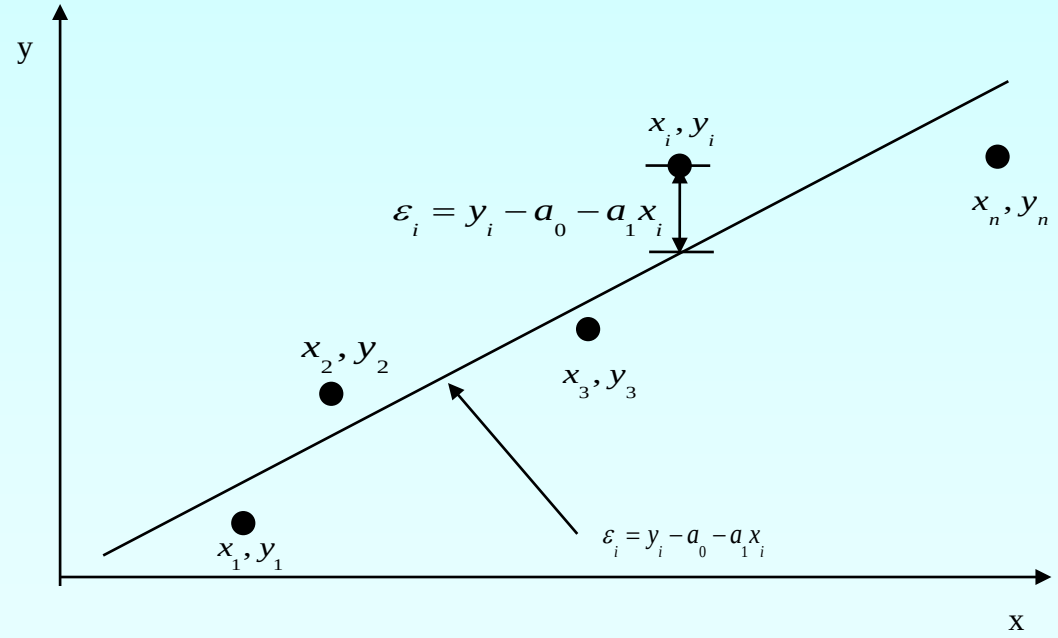

Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i.

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

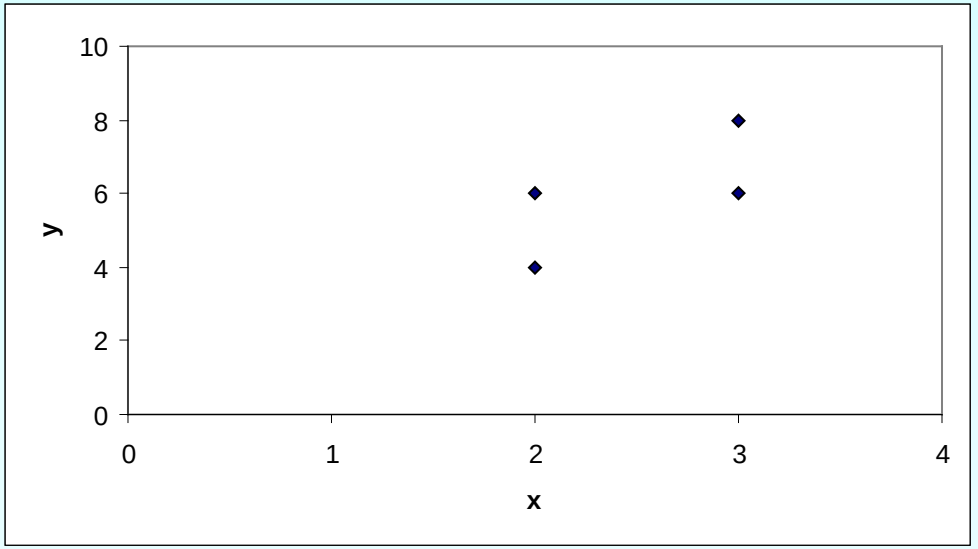

Figure. Data points for y vs. x data.

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	$y_{predicted}$	$\varepsilon = y - y_{predicted}$
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
			$\sum_{i=1}^{4} \varepsilon_i = 0$

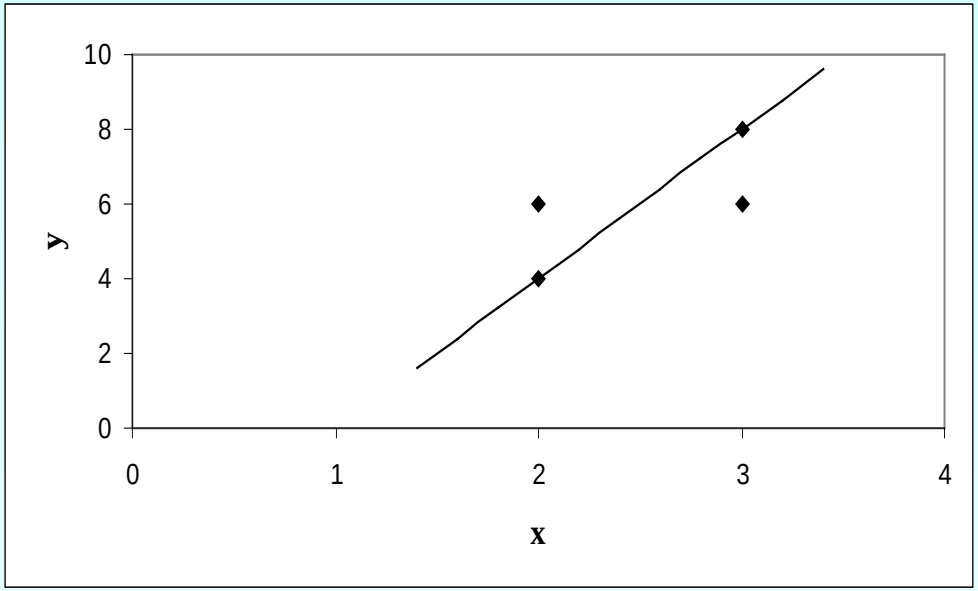

Figure. Regression curve for y=4x-4, y vs. x data

Linear Regression-Criteria#1

Using *y=6* as a regression curve

Table. Residuals at each point for y=6

x	y	$y_{predicted}$	$\varepsilon = y - y_{predicted}$
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
			$\sum_{i=1}^{4} \varepsilon_i = 0$

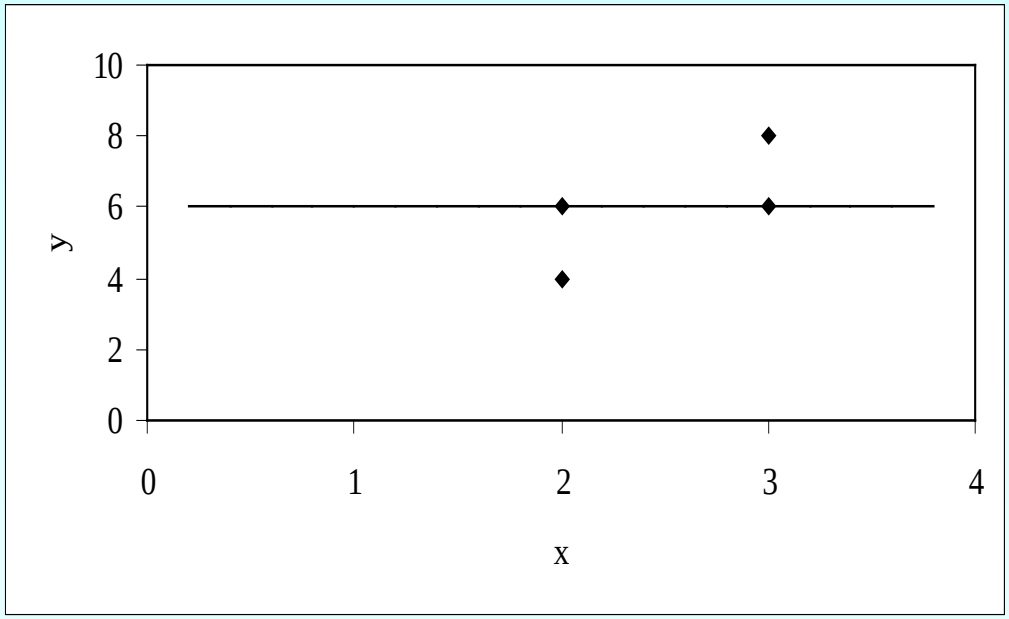

Figure. Regression curve for y=6, y vs. x data

Linear Regression – Criterion #1

$\sum_{i=1}^{4} \varepsilon_i = 0$ for both regression models of y=4x-4 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 $\sum_{i=1}^{n} \left| \varepsilon_i \right|$ work any better?

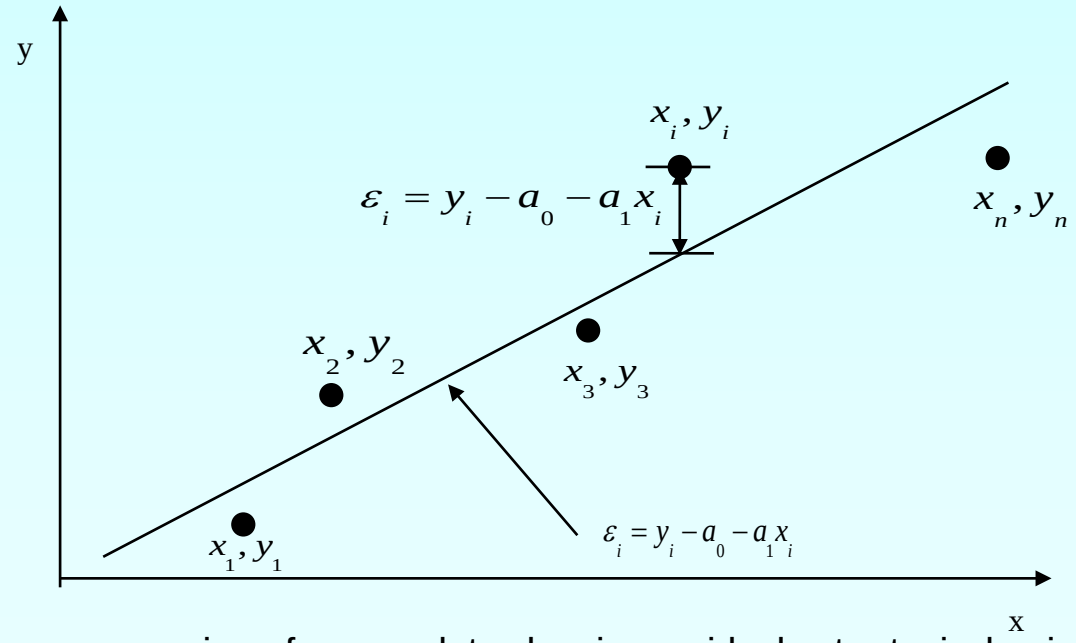

Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i.

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	$y_{predicted}$	$\|\varepsilon\| = \|y - y_{predicted}\|$
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
			$\sum_{i=1}^{4} \left\| \varepsilon_i \right\| = 4$

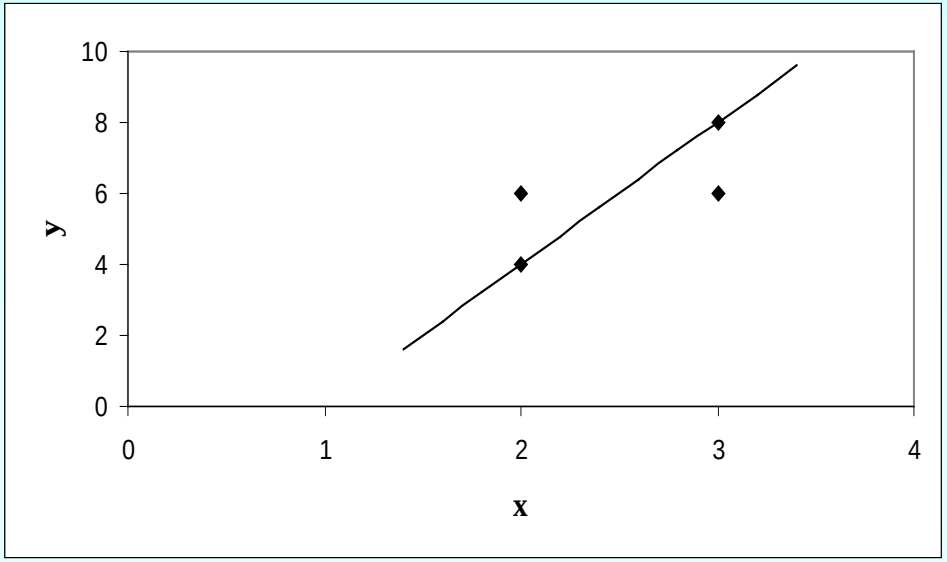

Figure. Regression curve for y=4x-4, y vs. x data

Linear Regression-Criteria#2

Using *y=6* as a regression curve

Table. Absolute residuals employing the y=6 model

x	y	$y_{predicted}$	$\|\varepsilon\| = \|y - y_{predicted}\|$
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
			$\sum_{i=1}^{4} \left\| \varepsilon_i \right\| = 4$

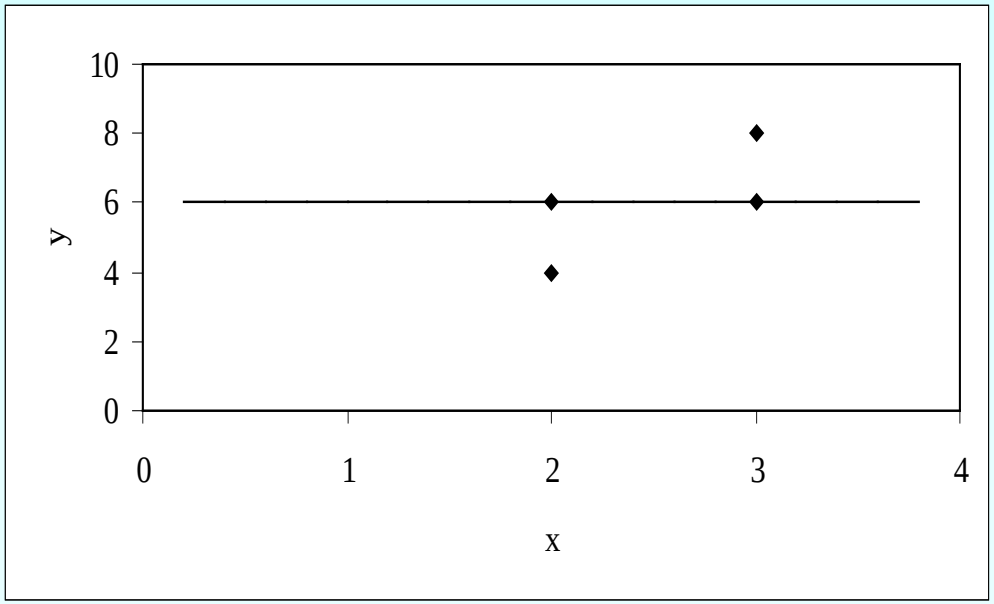

Figure. Regression curve for y=6, y vs. x data

Linear Regression-Criterion#2

$\sum_{i=1}^{4}\left|\varepsilon_i\right| = 4$ for both regression models of y=4x-4 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 $\sum_{i=1}^{4}\left|\varepsilon_i\right| < 4$ 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.

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

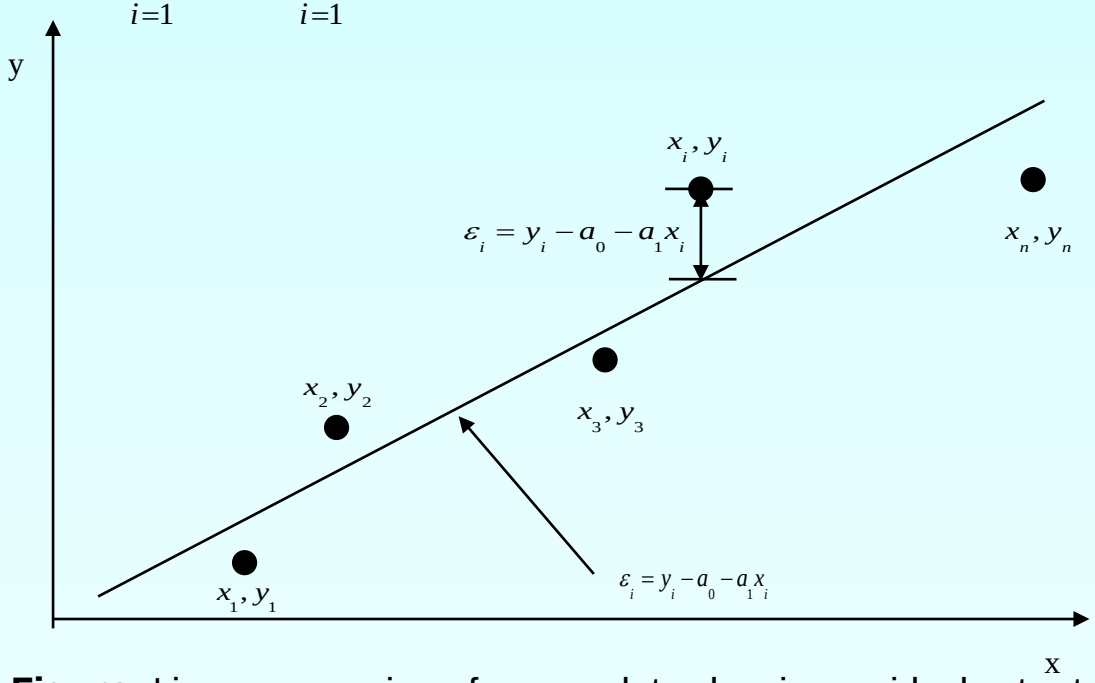

Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i.

Finding Constants of Linear Model

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

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

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

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

giving

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

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

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

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

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

Example 1

The coefficient of thermal expansion of steel is given at discrete values of temperature, as shown in the table.

Temperature, T	Coefficient of Thermal Expansion, α
${}^0F$	in/in^0F
80	6.470×10^{-6}
60	6.360×10^{-6}
40	6.240×10^{-6}
20	6.120×10^{-6}
0	6.000×10^{-6}
−20	5.860×10^{-6}
−40	5.720×10^{-6}
−60	5.580×10^{-6}
−80	5.430×10^{-6}
−100	5.280×10^{-6}
−120	5.090×10^{-6}
−140	4.910×10^{-6}
−160	4.720×10^{-6}
−180	4.520×10^{-6}
−200	4.300×10^{-6}
−220	4.080×10^{-6}
−240	3.830×10^{-6}
−260	3.580×10^{-6}
−280	3.330×10^{-6}
−300	3.070×10^{-6}
−320	2.760×10^{-6}
−340	2.450×10^{-6}

Table. Data points for thermal expansion vs. temperature

If the data is regressed to a first order polynomial,

$$\alpha = k_1 + k_2 T$$

Find the constants of the model.

Example 1 cont.

The necessary summations are calculated as

$$\sum_{i=1}^{22} T_i = -2860\ {}^0F$$

$$\sum_{i=1}^{22} \alpha_i = 1.5070\times10^{-4}\ \ in/in/{}^0F$$

$$\sum_{i=1}^{22} T_i\alpha_i = -1.0416\times10^{-2}\quad in/in$$

$$\sum_{i=1}^{22} T^2 = 726{,}000\ \left({}^0F\right)^2$$

Example 1 cont.

We can now calculate the value of k_2 using

$$k_2 = \frac{n\sum_{i=1}^{22} T_i\alpha_i - \sum_{i=1}^{22} T_i \sum_{i=1}^{22} \alpha_i}{n\sum_{i=1}^{22} T_i^2 - \left(\sum_{i=1}^{22} T_i\right)^2}$$

$$= \frac{22(-1.0416\times10^{-2}) - (-2860)(1.0570\times10^{-4})}{22(726000) - (-2860)^2}$$

$$= 9.3868\times10^{-9}\ in/in/(^0F)^2$$

Example 1 cont.

The value for k_1 can be calculated using $k_1 = \bar{\alpha} - k_2 \bar{T}$ where

$$\bar{\alpha} = \frac{\sum_{i=1}^{22} \alpha_i}{n} = 4.8045 \times 10^{-6}\ in/in^0F \qquad \bar{T} = \frac{\sum_{i=1}^{22} T_i}{n} = -130\ ^0F$$

$$\begin{aligned} k_1 &= \bar{\alpha} - k_2 \bar{T} \\ &= 4.8045 \times 10^{-6} - (9.3868 \times 10^{-9})(-130) \\ &= 6.0248 \times 10^{-6} in/in/^\circ F \end{aligned}$$

The regression model is now given by

$$\alpha = 6.0248 \times 10^{-6} + 9.3868 \times 10^{-9} T$$

Example 1 cont.

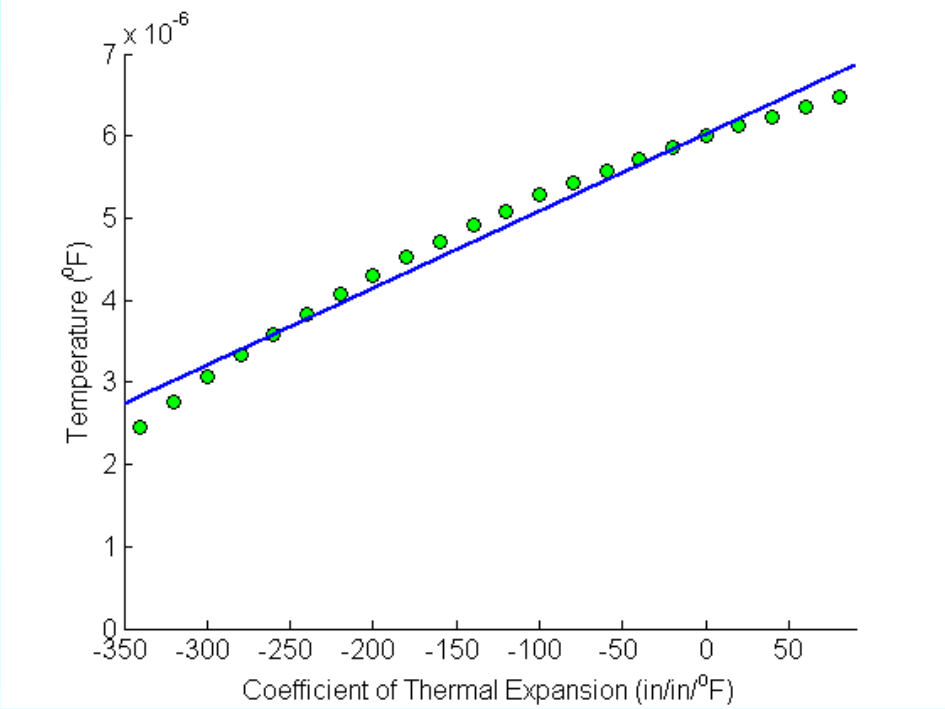

Figure. Linear regression of Coefficient of Thermal expansion vs. Temperature data

Question: Can you find the decrease in the diameter of a solid cylinder of radius 12” if the cylinder is cooled from a room temperature of 80°F to a dry-ice/alcohol bath with temperatures of −108°F?

What would be the error if you used the thermal expansion coefficient at room temperature to find the answer?

Example 2

To find the longitudinal modulus of composite, the following data is collected. Find the longitudinal modulus, E using the regression model $\sigma = E\varepsilon$ and the sum of the square of the residuals.

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

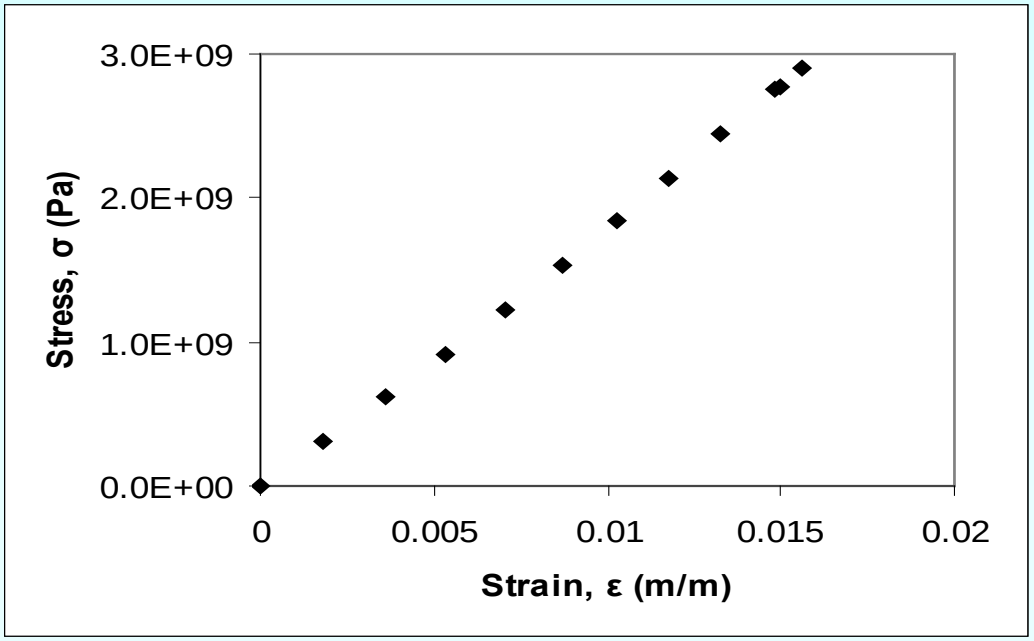

Figure. Data points for Stress vs. Strain data

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

$$S_r = \sum_{i=1}^{n} \gamma_i^2$$

$$= \sum_{i=1}^{n} \left(\sigma_i - E\varepsilon_i\right)^2$$

Differentiate with respect to E

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

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	εσ
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

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

and

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

Using

$$E = \frac{\sum_{i=1}^{12} \sigma_i \varepsilon_i}{\sum_{i=1}^{12} \varepsilon_i^2}$$

$$= \frac{2.3337 \times 10^{8}}{1.2764 \times 10^{-3}}$$

$$= 182.84\ GPa$$

Example 2 Results

The equation $\sigma = 182.84\varepsilon$ describes the data.

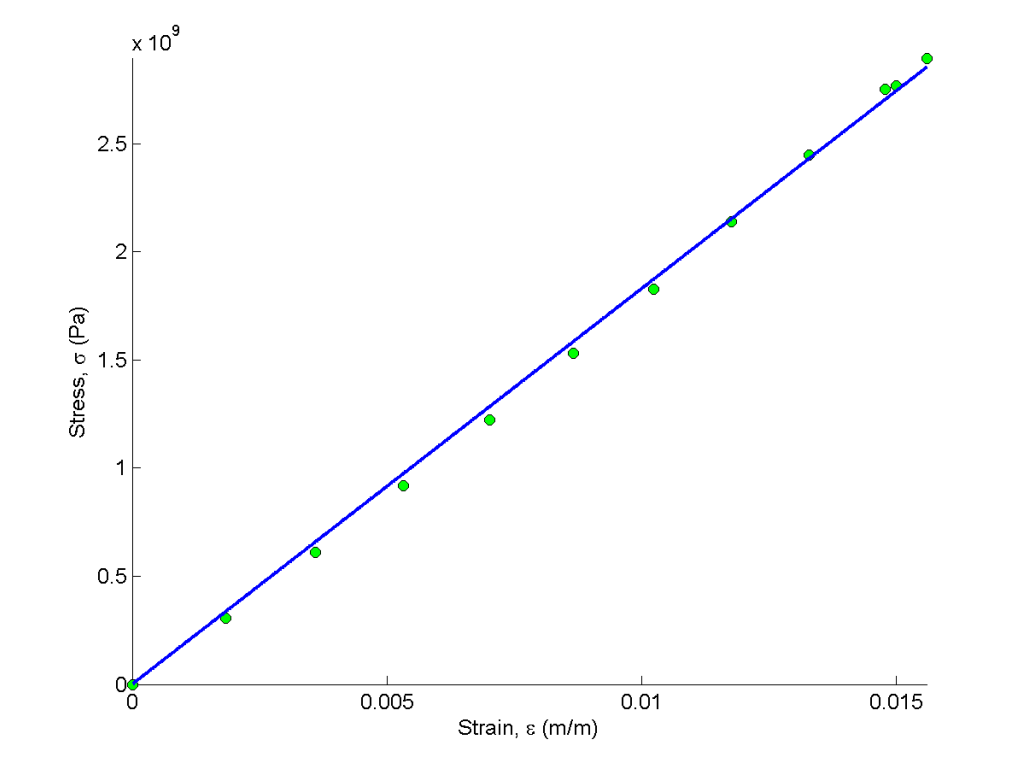

Figure. Linear regression for Stress vs. Strain data

Additional Resources

For all resources on this topic such as digital audiovisual lectures, primers, textbook chapters, multiple-choice tests, worksheets in MATLAB, MATHEMATICA, MathCad and MAPLE, blogs, related physical problems, please visit

http://numericalmethods.eng.usf.edu/topics/linear_regression.html

THE END

http://numericalmethods.eng.usf.edu