Linear Regression

CSE 214: Numerical Analysis Sessional

Line/Curve fitting

Given a set of points:

- experimental data
- tabular data
- etc.

Fit a line or a curve (surface) to the points so that we can easily evaluate f(x) at any x of interest.

If x within data range

➔ **interpolating** (generally safe)

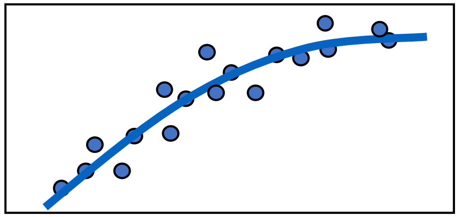
If x outside data range

➔**extrapolating** (often dangerous)

Curve fitting

Two main methods :

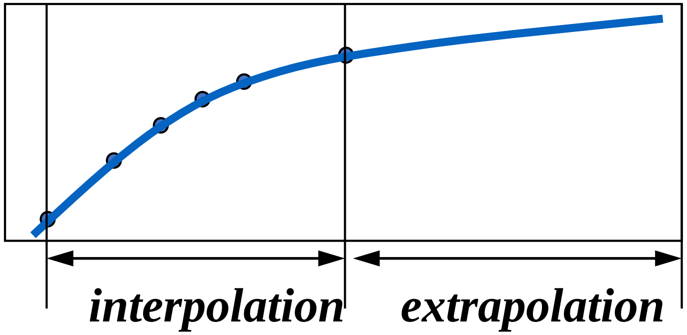
1. **Least-Squares Regression**



- Function is "best fit" to data.
- Does not necessarily pass through points.
- Used for **scattered** data (experimental)
- Can develop models for analysis/design.

2. **Interpolation**

- Function passes through all (or most) points.
- Interpolates values of well-behaved (**precise**) data or for geometric design.



What is Regression?

Regression defines for *n* data points say $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ The best $y = f(x)$ 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

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

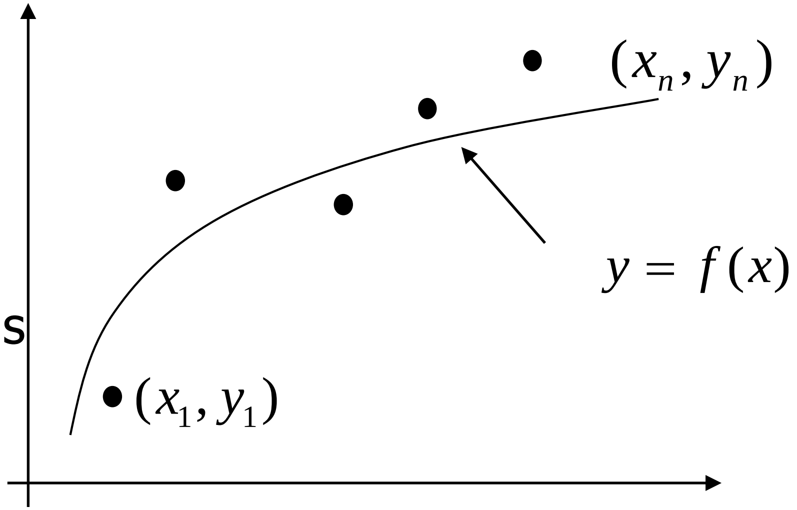


Figure. Basic model for regression

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.

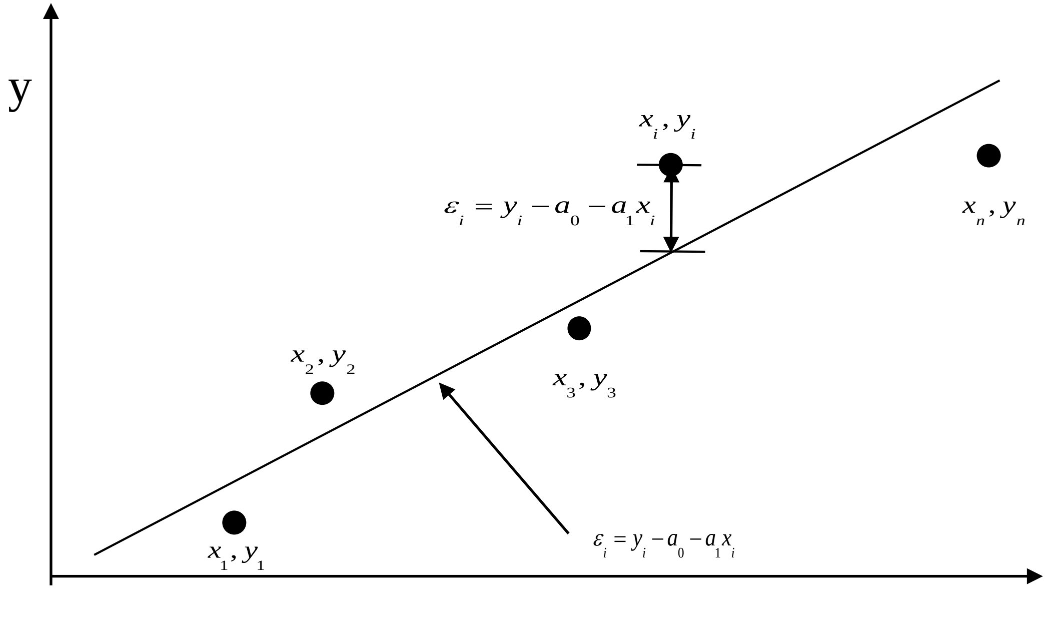

y
x_i, y_i
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x_n, y_n
x_2, y_2
x_3, y_3
x_1, y_1
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x

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

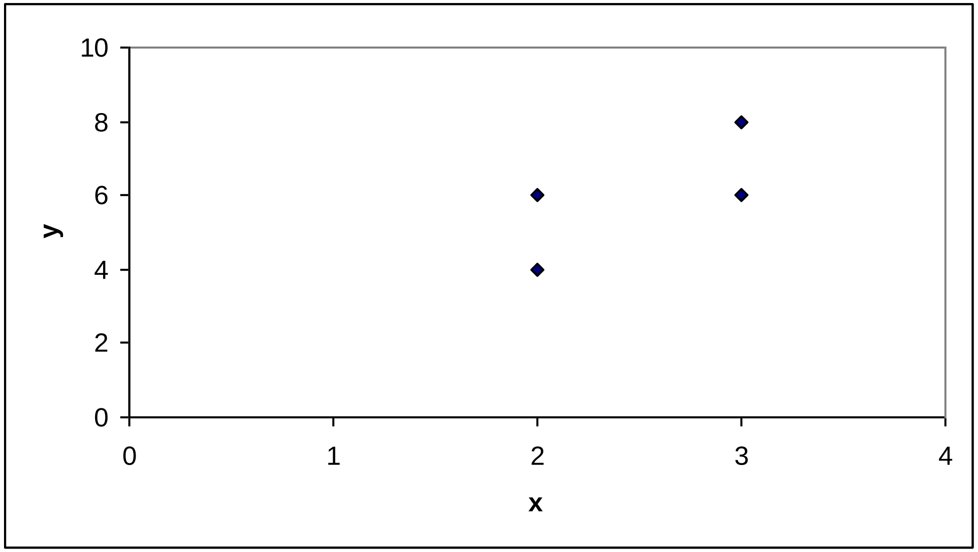


Figure. Data points for y vs. x data.

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
			$\sum_{i=1}^{4} \varepsilon_i = 0$

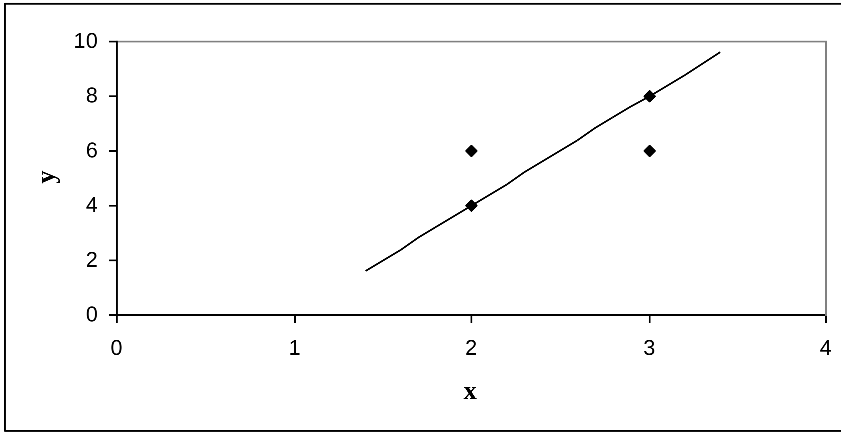


Figure. Regression curve for y=4x-4, y vs. x data

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
			$\sum_{i=1}^{4} \varepsilon_i = 0$

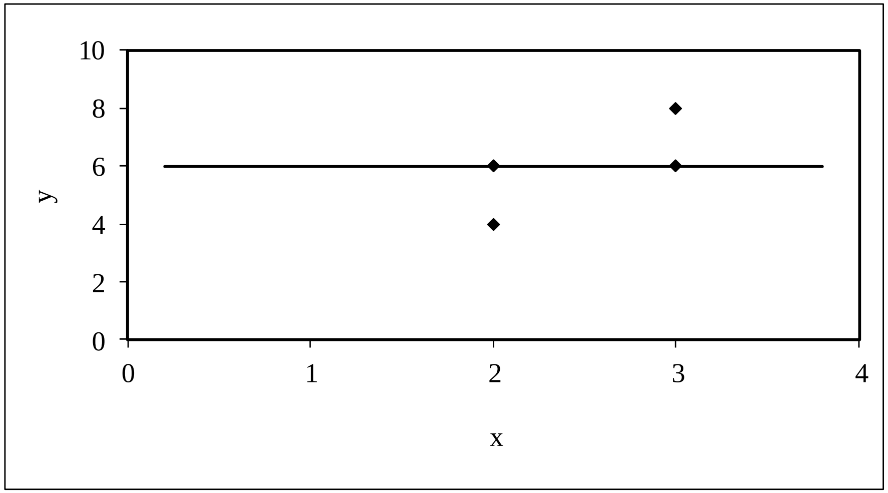


Figure. Regression curve for y=6, y vs. x data

Linear Regression – Criterion #1

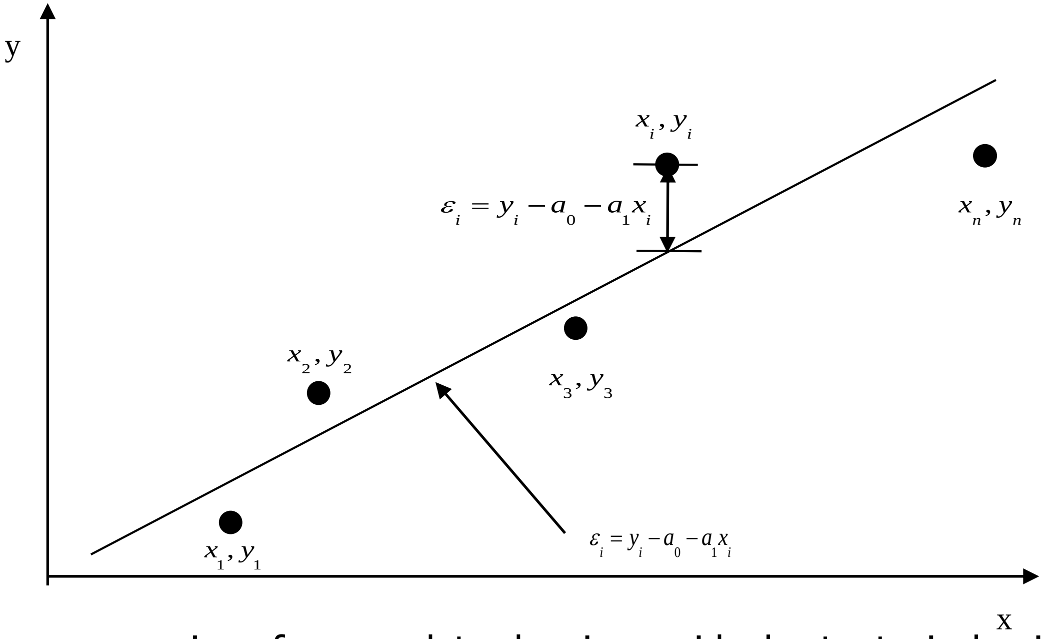
$\sum_{i=1}^{4} \varepsilon_i = 0$ for both regression models of y=4x-4 and y=6.

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?



y
x_i, y_i
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x_n, y_n
x_2, y_2
x_3, y_3
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x_1, y_1
x

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

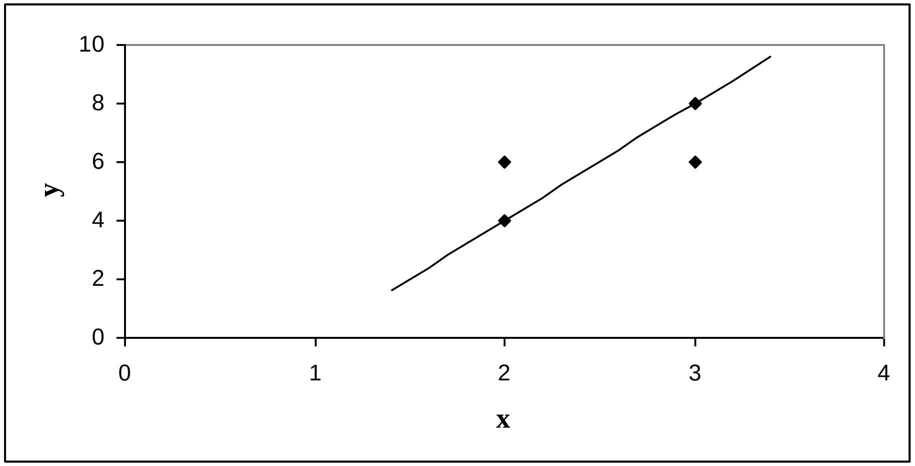


Figure. Regression curve for y=4x-4, y vs. x data

How should we calculate the residual error then?

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.

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

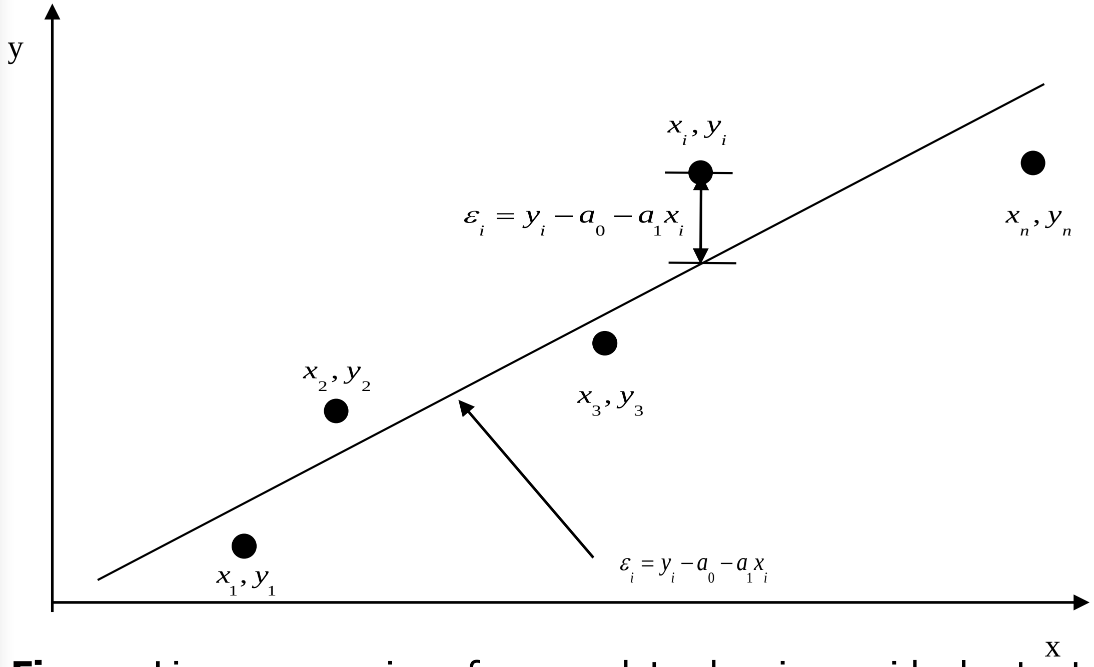

y
x_i, y_i
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x_n, y_n
x_2, y_2
x_3, y_3
x_1, y_1
$\varepsilon_i = y_i - a_0 - a_1 x_i$
x

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

Finding Constants of Linear Model

Minimizing 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.

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

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

giving

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

$$\sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i{}^2 = \sum_{i=1}^{n} y_i x_i \qquad (a_0 = \overline{y} - a_1 \overline{x})$$

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

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

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

Example 1 (2 Const)

The torque, T needed to turn the torsion spring of a mousetrap through an angle, is given below. Find the constants for the model given by
$$T = k_1 + k_2\theta$$

Table: Torque vs Angle for a torsional spring

Angle, θ	Torque, T
Radians	*N-m*
0.698132	0.188224
0.959931	0.209138
1.134464	0.230052
1.570796	0.250965
1.919862	0.313707

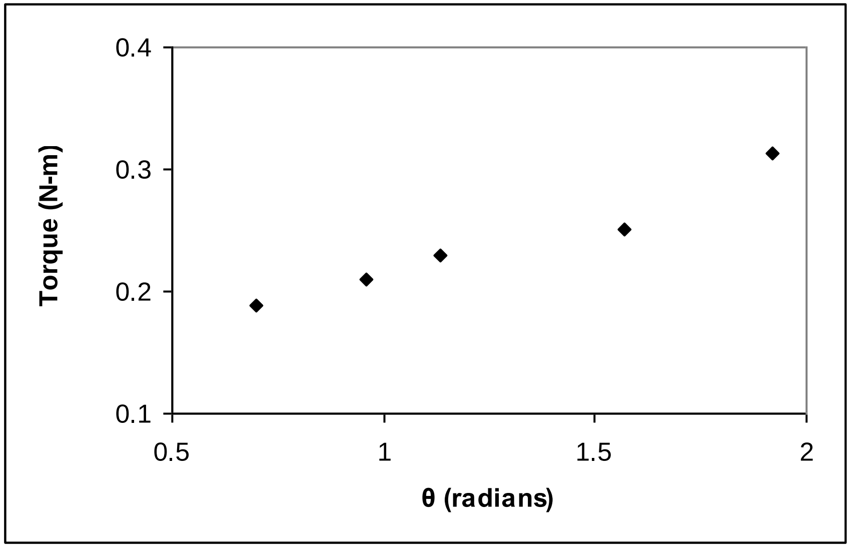


Figure. Data points for Angle vs. Torque data

Example 1 cont.

The following table shows the summations needed for the calculations of the constants in the regression model.

Table. Tabulation of data for calculation of important summations

θ	T	θ^2	$T\theta$
Radians	*N-m*	*Radians²*	*N-m-Radians*
0.698132	0.188224	0.487388	0.131405
0.959931	0.209138	0.921468	0.200758
1.134464	0.230052	1.2870	0.260986
1.570796	0.250965	2.4674	0.394215
1.919862	0.313707	3.6859	0.602274
6.2831	1.1921	8.8491	1.5896

Using equations described for a_0 and a_1 with $n = 5$

$$k_2 = \frac{n\sum_{i=1}^{5}\theta_i T_i - \sum_{i=1}^{5}\theta_i \sum_{i=1}^{5} T_i}{n\sum_{i=1}^{5}\theta_i^2 - \left(\sum_{i=1}^{5}\theta_i\right)^2}$$

$$= \frac{5(1.5896) - (6.2831)(1.1921)}{5(8.8491) - (6.2831)^2}$$

$$= 9.6091 \times 10^{-2}$$ *N-m/rad*

Example 1 cont.

Use the average torque and average angle to calculate k_1

$$\bar{T} = \frac{\sum_{i=1}^{5} T_i}{n} \qquad \bar{\theta} = \frac{\sum_{i=1}^{5} \theta_i}{n}$$

$$= \frac{1.1921}{5} \qquad = \frac{6.2831}{5}$$

$$= 2.3842 \times 10^{-1} \qquad = 1.2566$$

Using,

$$k_1 = \bar{T} - k_2 \, \bar{\theta}$$

$$= 2.3842 \times 10^{-1} - (9.6091 \times 10^{-2})(1.2566)$$

$$= 1.1767 \times 10^{-1} \text{ } \textit{N-m}$$

Example 1 Results

Using linear regression, a trend line is found from the data

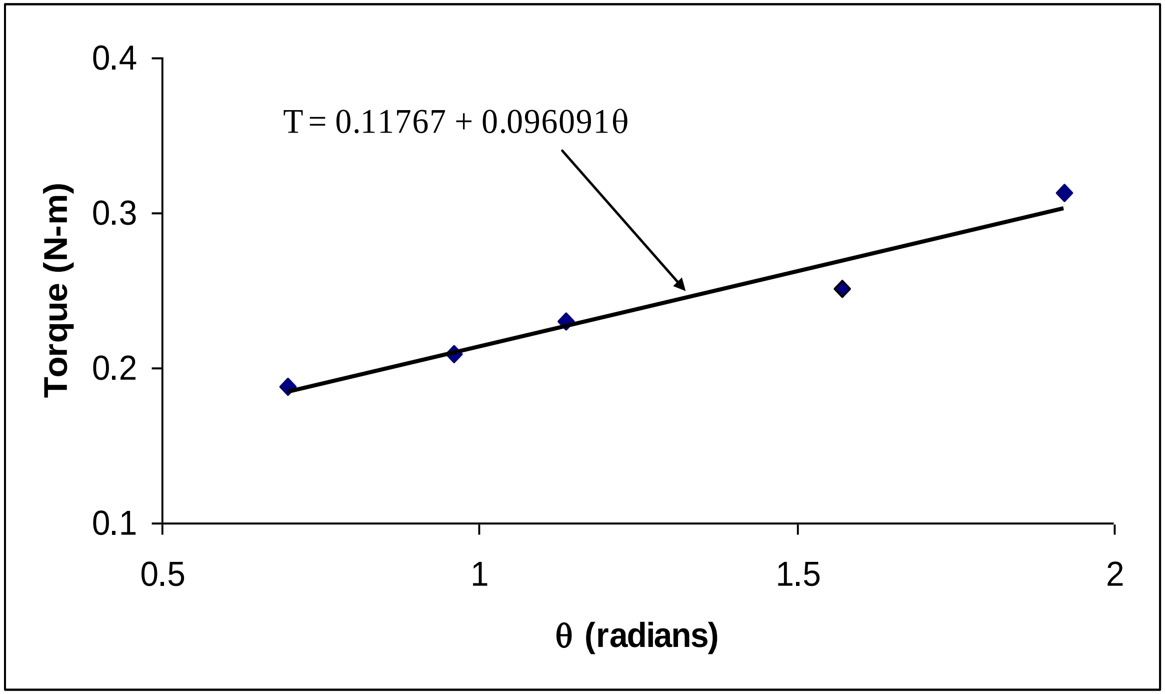


Figure. Linear regression of Torque versus Angle data

Example 2 (1 Const)

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

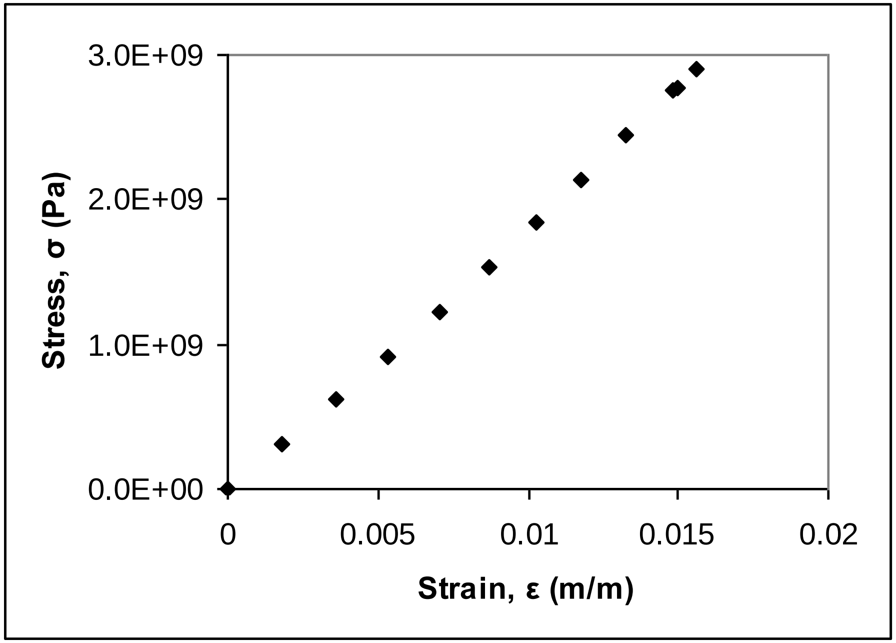


Figure. Data points for Stress vs. Strain data

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

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

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

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	1.8300x10^{-3}	3.0600x10^{8}	3.3489x10^{-6}	5.5998x10^{5}
3	3.6000x10^{-3}	6.1200x10^{8}	1.2960x10^{-5}	2.2032x10^{6}
4	5.3240x10^{-3}	9.1700x10^{8}	2.8345x10^{-5}	4.8821x10^{6}
5	7.0200x10^{-3}	1.2230x10^{9}	4.9280x10^{-5}	8.5855x10^{6}
6	8.6700x10^{-3}	1.5290x10^{9}	7.5169x10^{-5}	1.3256x10^{7}
7	1.0244x10^{-2}	1.8350x10^{9}	1.0494x10^{-4}	1.8798x10^{7}
8	1.1774x10^{-2}	2.1400x10^{9}	1.3863x10^{-4}	2.5196x10^{7}
9	1.3290x10^{-2}	2.4460x10^{9}	1.7662x10^{-4}	3.2507x10^{7}
10	1.4790x10^{-2}	2.7520x10^{9}	2.1874x10^{-4}	4.0702x10^{7}
11	1.5000x10^{-2}	2.7670x10^{9}	2.2500x10^{-4}	4.1505x10^{7}
12	1.5600x10^{-2}	2.8960x10^{9}	2.4336x10^{-4}	4.5178x10^{7}
$\sum\limits_{i=1}^{12}$			1.2764x10^{-3}	2.3337x10^{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

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

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

$$= 182.83 \ GPa$$

Example 2 Results

How well the equation $\sigma = 182.823\varepsilon$ describes the data?

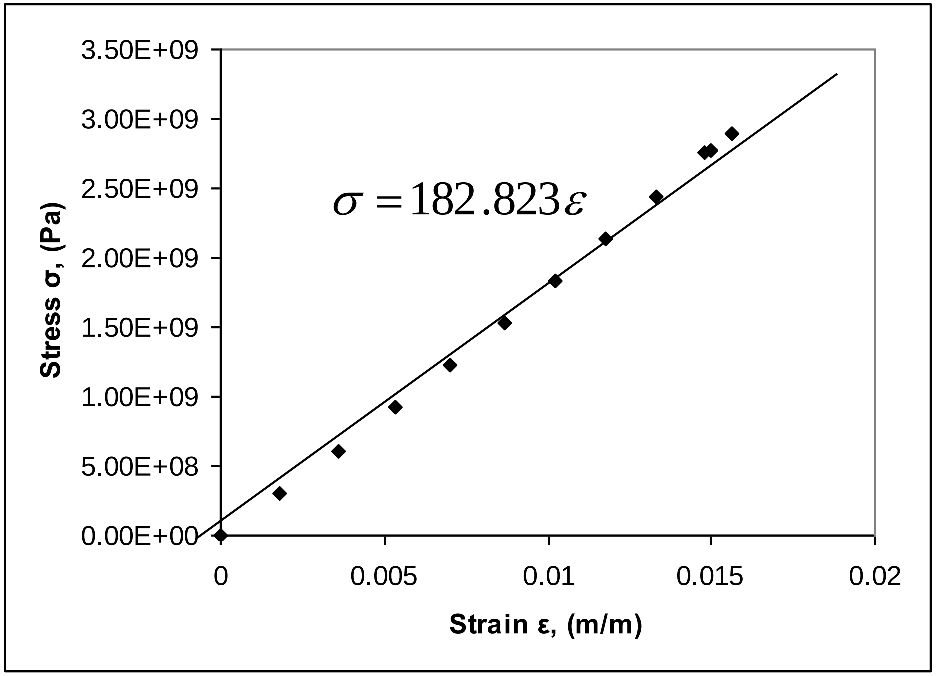


Figure. Linear regression for Stress vs. Strain data

Goodness of fit

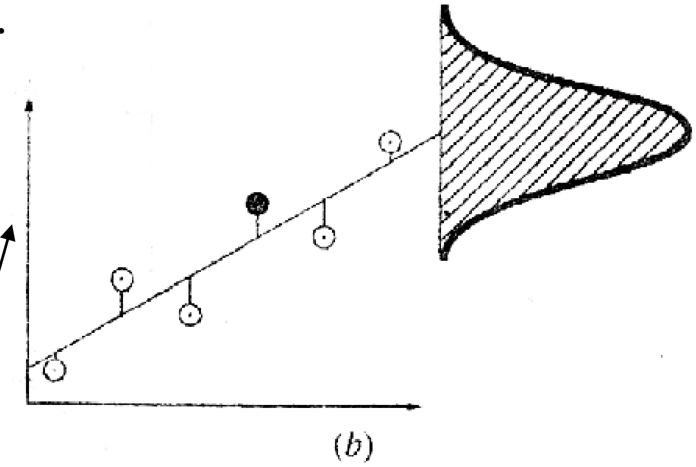
1. In all regression models one is solving an overdetermined system of equations, i.e., more equations than unknowns.

2. How good is the fit?
   Often based on a
   coefficient of determination, r^2

Goodness of fit

***r*²** Compares the average spread of the data about the **regression line** compared to the spread of the data about the **mean**.

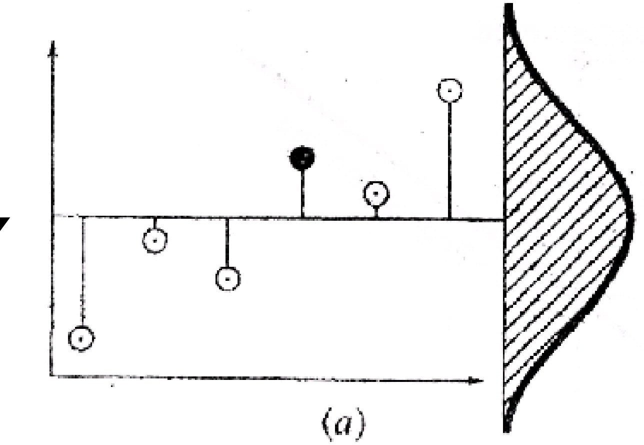
Spread of the data around the regression line:



$$\boxed{S_r = \sum e_i^2 = \sum (y_i - y'_i)^2}$$

Spread of the data around the mean:

$$\boxed{S_t = \sum (y_i - \bar{y})^2}$$



Coefficient of determination

describes how much of variance is “explained” by
the regression equation

$$r^2 = \frac{S_t - S_r}{S_t}$$

- Want ***r*²** close to 1.0; $r^2 \approx 0$ *means no improvement over taking the average value as an estimate.*
- Doesn't work if models have different numbers of parameters.
- Be careful when using different transformations – always do the analysis on the untransformed data.

Precision

If the spread of the points around the line is of similar magnitude along the entire range of the data,

Then one can use

$$s_{x/y} = \sqrt{\frac{S_r}{n-2}}$$

= **standard error of the estimate**
(standard deviation in y)

to describe the precision of the regression estimate