Linear Regression

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

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

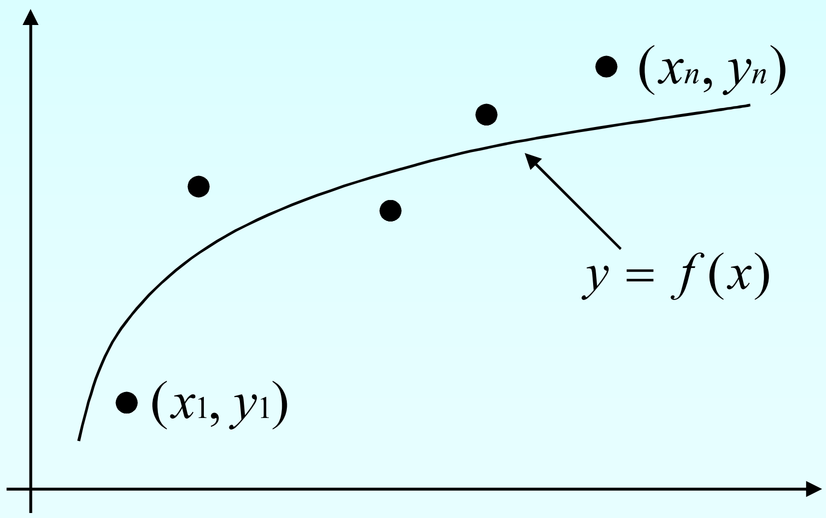


Figure. Basic model for regression

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.

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

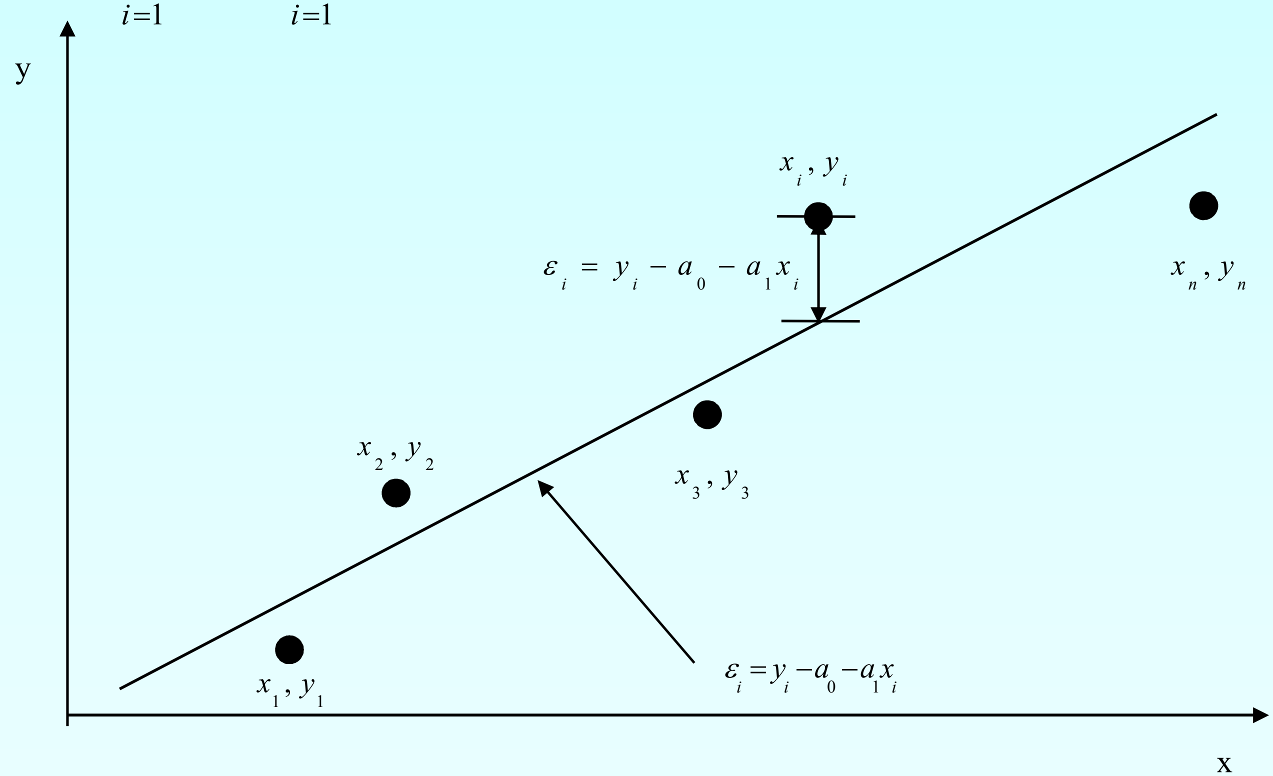


Figure. Linear regression of y vs. x data showing residuals at a typical point, x_i.

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.

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

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

giving

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

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

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

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

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

Example 1

Fit a straight line and find the constants a_0 and a_1 .

x	y
1	2
2	5
4	7
5	10
6	12
8	15
9	19

Example 1 cont.

x	y	x^2	xy
1	2	1	2
2	5	4	10
4	7	16	28
5	10	25	50
6	12	36	72
8	15	64	120
9	19	81	171
35	**70**	**227**	**453**

Example 1 cont.

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

$$a_0 = \frac{70 \times 227 - 35 \times 453}{7 \times 227 - (35)^2} = 0.096$$

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

$$a_1 = \frac{7 \times 453 - 35 \times 70}{7 \times 227 - (35)^2} = 1.98$$

$$y = 0.096 + 1.98x$$

Linear Regression (special case)

Given

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

best fit

$$y = a_1 x$$

to the data.

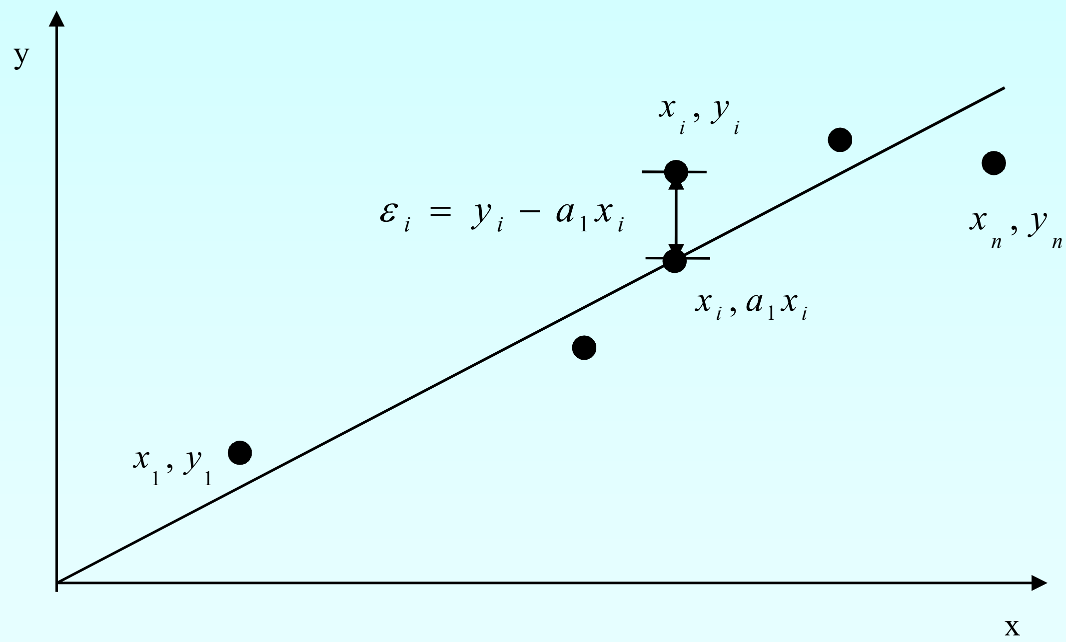
Linear Regression (special case cont.)

$$y = a_1 x$$

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

Is this correct?

Linear Regression (special case cont.)



Linear Regression (special case cont.)

Residual at each data point

$$\varepsilon_i = y_i - a_1 x_i$$

Sum of square of residuals

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

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

Linear Regression (special case cont.)

Differentiate with respect to a_1

$$\frac{dS_r}{da_1} = \sum_{i=1}^{n} 2\left(y_i - a_1 x_i\right)\left(-x_i\right)$$

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

$$\frac{dS_r}{da_1} = 0$$

gives

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

Quadratic polynomial

$$S_r = \sum_{i=1}^{n} \varepsilon_i{}^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i{}^2 \right)^2$$

Differentiating S_r with respect to a_0 a_1 and a_2

$$n\, a_0 + a_1 \sum x_i + a_2 \sum x_i^2 = \sum y_i$$

$$a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 = \sum x_i y_i$$

$$a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 = \sum x_i^2 y_i$$

These equation may be solved by Gauss elimination procedure

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

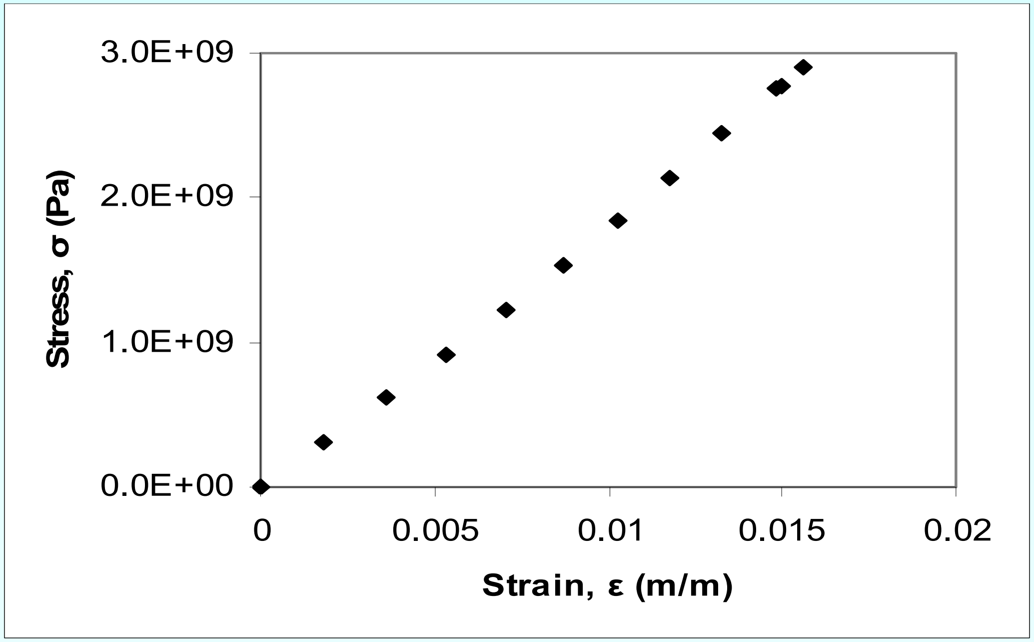


Figure. Data points for Stress vs. Strain data

Example 2 cont.

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}

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

$$\sum_{i=1}^{12}\varepsilon_i^2=1.2764\times10^{-3}$$

$$\sum_{i=1}^{12}\sigma_i\varepsilon_i=2.3337\times10^{8}$$

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

$$=\frac{2.3337\times10^{8}}{1.2764\times10^{-3}}$$

$$=182.84\ GPa$$

Example 2 Results

The equation $\sigma = 182.84 \times 10^9 \varepsilon$ describes the data.

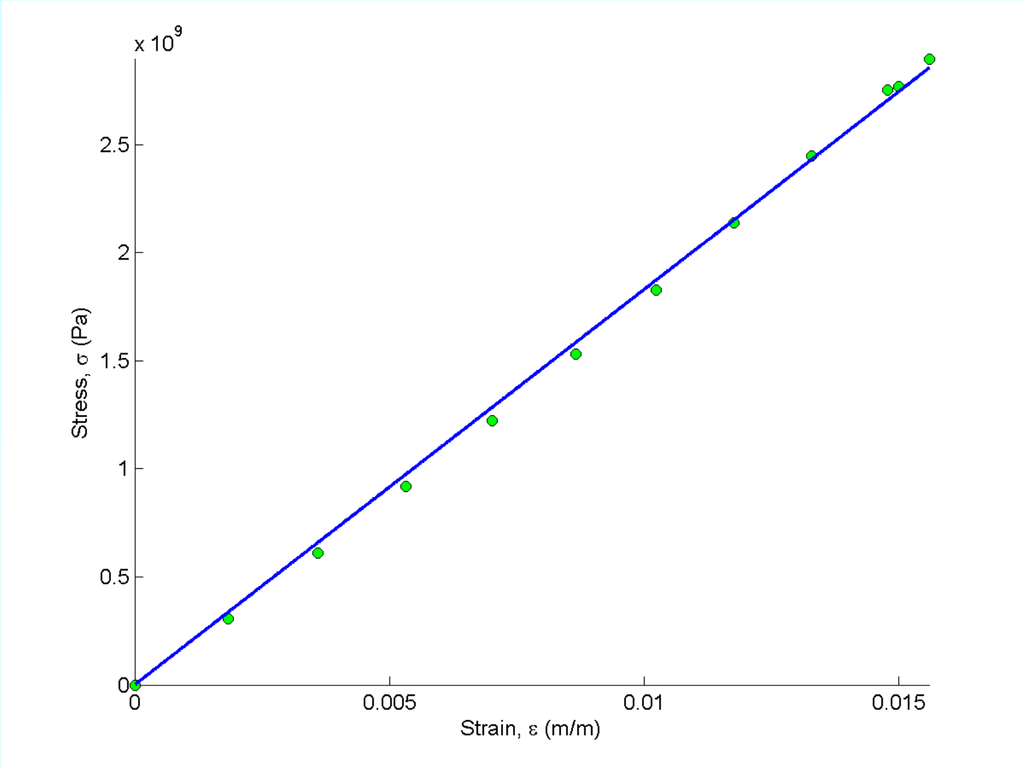


Figure. Linear regression for stress vs. strain data