Chapter 06.03
Linear Regression

After reading this chapter, you should be able to

1. *define regression,*
2. *use several minimizing of residual criteria to choose the right criterion,*
3. *derive the constants of a linear regression model based on least squares method criterion,*
4. *use in examples, the derived formulas for the constants of a linear regression model, and*
5. *prove that the constants of the linear regression model are unique and correspond to a minimum.*

Linear regression is the most popular regression model. In this model, we wish to predict response to n data points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ by a regression model given by

$$y = a_0 + a_1 x \tag{1}$$

where a_0 and a_1 are the constants of the regression model.

A measure of goodness of fit, that is, how well $a_0 + a_1 x$ predicts the response variable y is the magnitude of the residual ε_i at each of the n data points.

$$E_i = y_i - (a_0 + a_1 x_i) \tag{2}$$

Ideally, if all the residuals ε_i are zero, one may have found an equation in which all the points lie on the model. Thus, minimization of the residual is an objective of obtaining regression coefficients.

The most popular method to minimize the residual is the least squares methods, where the estimates of the constants of the models are chosen such that the sum of the squared residuals is minimized, that is minimize $\sum_{i=1}^{n} E_i^{\ 2}$.

Why minimize the sum of the square of the residuals? Why not, for instance, minimize the sum of the residual errors or the sum of the absolute values of the residuals? Alternatively, constants of the model can be chosen such that the average residual is zero without making individual residuals small. Will any of these criteria yield unbiased parameters with the smallest variance? All of these questions will be answered below. Look at the data in Table 1.

Table 1 Data points.

x	y
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0

To explain this data by a straight line regression model,

$$y = a_0 + a_1 x \tag{3}$$

and using minimizing $\sum_{i=1}^{n} E_i$ as a criteria to find a_0 and a_1, we find that for (Figure 1)

$$y = 4x - 4 \tag{4}$$

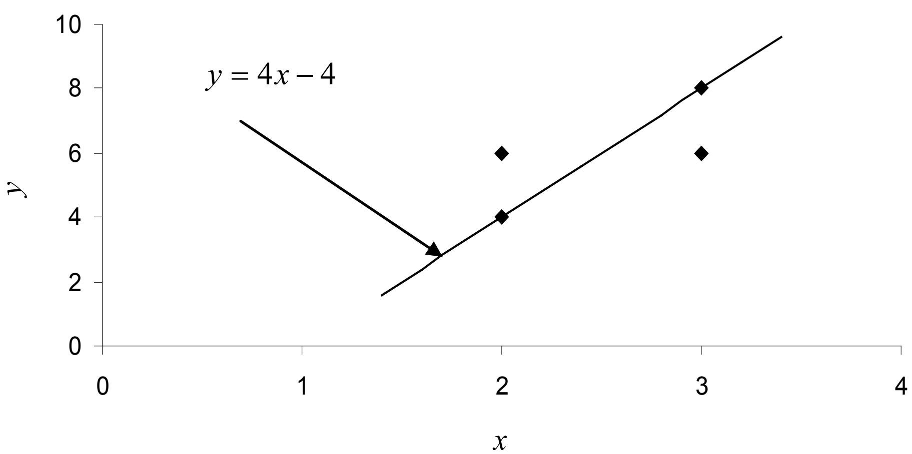

Figure 1 Regression curve $y = 4x - 4$ for y vs. x data.

the sum of the residuals, $\sum_{i=1}^{4} E_i = 0$ as shown in the Table 2.

Table 2 The residuals at each data 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
			$\sum_{i=1}^{4} \varepsilon_i = 0$

So does this give us the smallest error? It does as $\sum_{i=1}^{4} E_i = 0$. But it does not give unique values for the parameters of the model. A straight-line of the model

$$y = 6 \tag{5}$$

also makes $\sum_{i=1}^{4} E_i = 0$ as shown in the Table 3.

Table 3 The residuals at each data point for regression model $y = 6$

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

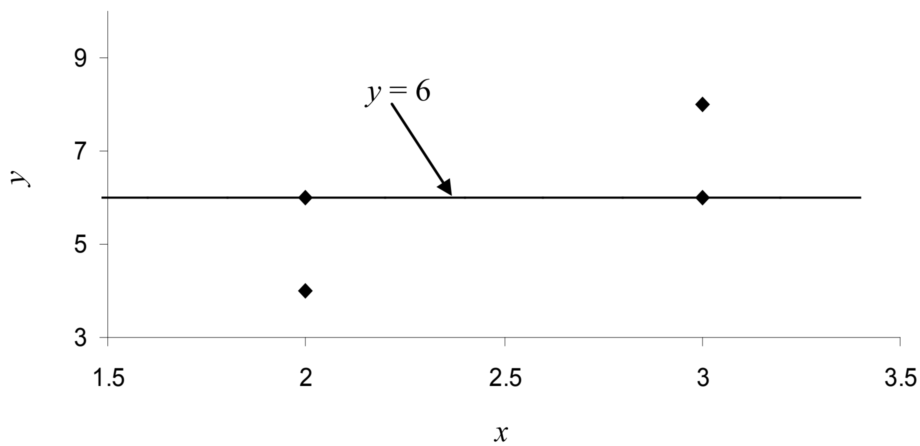

Figure 2 Regression curve $y = 6$ for y vs. x data.

Since this criterion does not give a unique regression model, it cannot be used for finding the regression coefficients. Let us see why we cannot use this criterion for any general data. We want to minimize

$$\sum_{i=1}^{n} E_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) \tag{6}$$

Differentiating Equation (6) with respect to a_0 and a_1, we get

$$\frac{\partial \sum_{i=1}^{n} E_i}{\partial a_0} = -\sum_{i=1}^{n} 1 = -n \tag{7}$$

$$\frac{\partial \sum_{i=1}^{n} E_i}{\partial a_1} = -\sum_{i=1}^{n} x_i = -n\bar{x} \tag{8}$$

Putting these equations to zero, give $n = 0$ but that is not possible. Therefore, unique values of a_0 and a_1 do not exist.

You may think that the reason the minimization criterion $\sum_{i=1}^{n} E_i$ does not work is that negative residuals cancel with positive residuals. So is minimizing $\sum_{i=1}^{n} |E_i|$ better? Let us look at the data given in the Table 2 for equation $y = 4x - 4$. It makes $\sum_{i=1}^{4} |E_i| = 4$ as shown in the following table.

Table 4 The absolute residuals at each data point when employing $y = 4x - 4$.

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
			$\sum_{i=1}^{4} \lvert\varepsilon_i\rvert = 4$

The value of $\sum_{i=1}^{4} |E_i| = 4$ also exists for the straight line model $y = 6$. No other straight line model for this data has $\sum_{i=1}^{4} |E_i| < 4$. Again, we find the regression coefficients are not unique, and hence this criterion also cannot be used for finding the regression model.

Let us use the least squares criterion where we minimize

$$S_r = \sum_{i=1}^{n} E_i^{\ 2} = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2 \tag{9}$$

S_r is called the sum of the square of the residuals.

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

$$\frac{\partial S_r}{\partial a_0} = 2\sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)(-1) = 0 \tag{10}$$

$$\frac{\partial S_r}{\partial a_1} = 2\sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)(-x_i) = 0 \tag{11}$$

giving

$$-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i = 0 \tag{12}$$

$$-\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2 = 0 \quad (13)$$

Noting that $\sum_{i=1}^{n} a_0 = a_0 + a_0 + . . . + a_0 = na_0$

$$na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (14)$$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad (15)$$

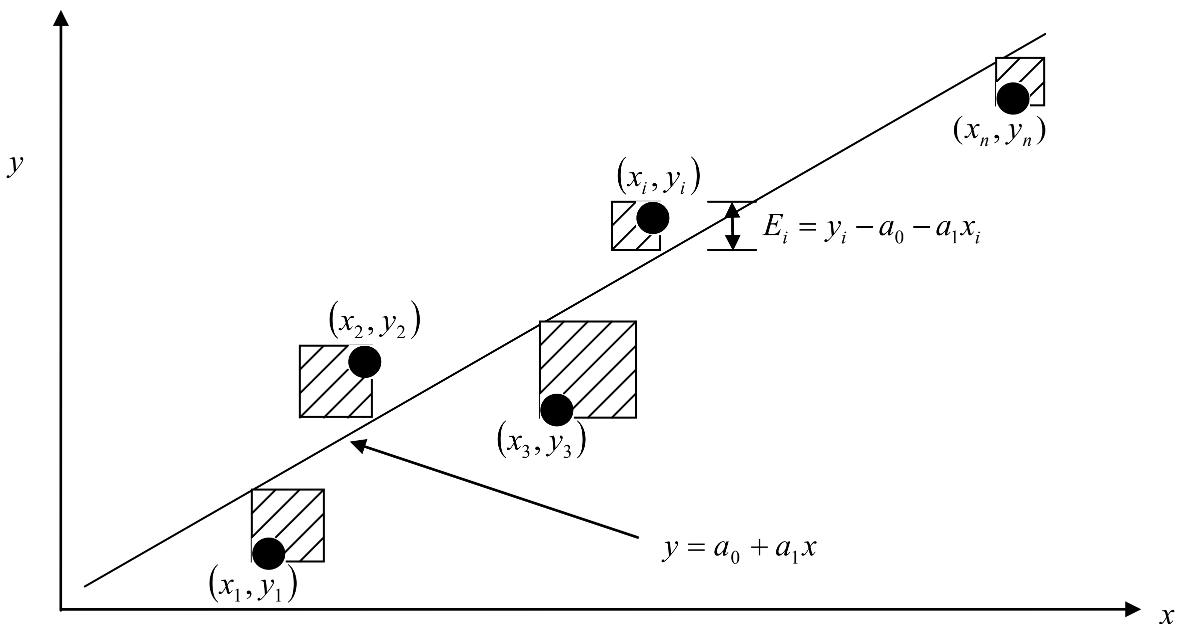

Figure 3 Linear regression of y vs. x data showing residuals and square of residual at a typical point, x_i.

Solving the above Equations (14) and (15) gives

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (16)$$

$$a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (17)$$

Redefining

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \quad (18)$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \quad (19)$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (20)$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad (21)$$

we can rewrite

$$a_1 = \frac{S_{xy}}{S_{xx}} \quad (22)$$

$$a_0 = \bar{y} - a_1\bar{x} \quad (23)$$

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 5 Off target value as a function of machine use.

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

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

Solution

Table 6 shows the summations needed for the calculation of the constants of the regression model.

Table 6 Tabulation of data for calculation of needed summations.

I	t	h	t^2	$t \times h$
$-$	Hours	Millimeters	Hours^2	Millimeter-Hour
1	30	1.10	900	33
2	33	1.21	1089	39.93
3	34	1.25	1156	42.50
4	35	1.23	1225	43.05
5	39	1.30	1521	50.70
6	44	1.40	1936	61.6
7	45	1.42	2025	63.9
$\sum_{i=1}^{7}$	260	8.91	9852	334.68

$$n = 7$$

$$\begin{aligned} a_1 &= \frac{n\sum_{i=1}^{7} t_i h_i - \sum_{i=1}^{7} t_i \sum_{i=1}^{7} h_i}{n\sum_{i=1}^{7} t_i^2 - \left(\sum_{i=1}^{7} t_i\right)^2} \\ &= \frac{7(334.68) - (260)(8.91)}{7(9852) - (260)^2} \\ &= 0.019179 \text{ mm-h} \end{aligned}$$

$$\begin{aligned} \bar{h} &= \frac{\sum_{i=1}^{7} h_i}{n} \\ &= \frac{8.91}{7} \\ &= 1.2729 \text{ mm} \end{aligned}$$

$$\begin{aligned} \bar{t} &= \frac{\sum_{i=1}^{7} t_i}{n} \\ &= \frac{260}{7} \\ &= 37.143 \text{ h} \end{aligned}$$

$$\begin{aligned} a_0 &= \bar{h} - a_1 \bar{t} \\ &= 1.2729 - (0.019179)(37.143) \end{aligned}$$

$$= 0.56050 \text{ mm-h}$$
$$h = 0.56050 + 0.019179t$$

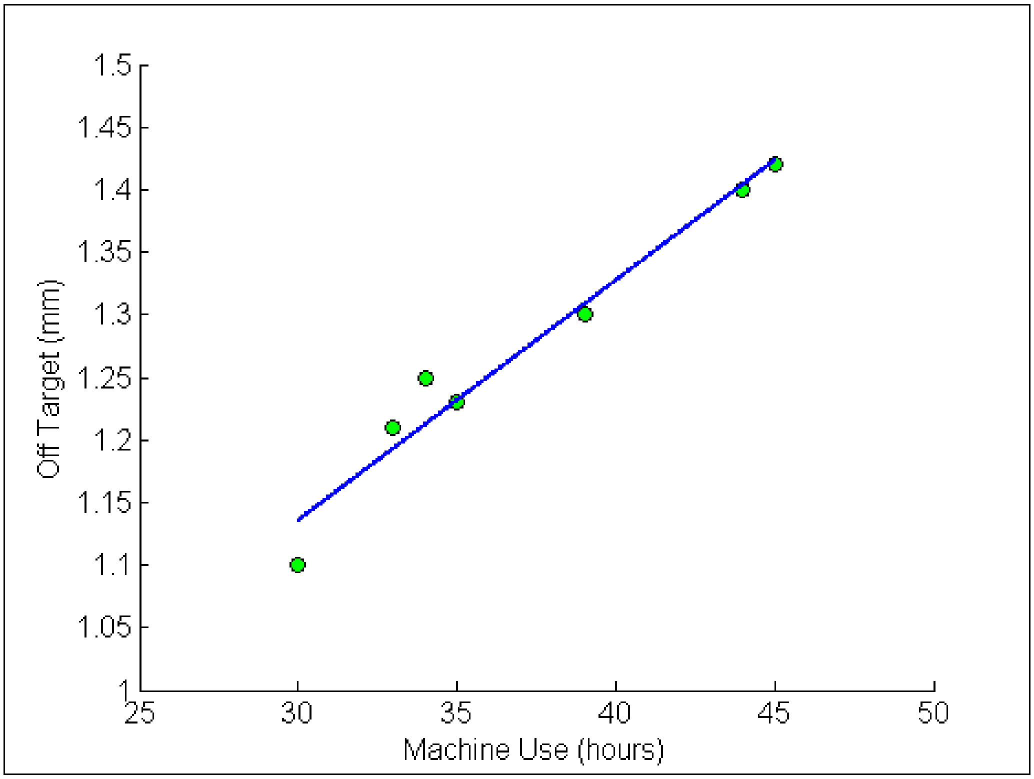

Figure 4 Linear regression of hours of use vs. millimeters off target.

Solving for $h = 2$ mm, the regression model is $h = 0.56050 + 0.019179t$

$$2 = 056050 + 0.019179t$$
$$t = \frac{2 - 0.56050}{0.019179}$$
$$t = 75.056 \text{ hours}$$

Example 2

To find the longitudinal modulus of a composite material, the following data, as given in Table 7, is collected.

Table 7 Stress vs. strain data for a composite material.

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

Find the longitudinal modulus E using the regression model.

$$\sigma = E\varepsilon \tag{24}$$

Solution

Rewriting data from Table 7, stresses versus strain data in Table 8

Table 8 Stress vs strain data for a composite in SI system of units

Strain (m/m)	Stress (Pa)
0.0000	0.0000
1.8300×10^{-3}	3.0600×10^{8}
3.6000×10^{-3}	6.1200×10^{8}
5.3240×10^{-3}	9.1700×10^{8}
7.0200×10^{-3}	1.2230×10^{9}
8.6700×10^{-3}	1.5290×10^{9}
1.0244×10^{-2}	1.8350×10^{9}
1.1774×10^{-2}	2.1400×10^{9}
1.3290×10^{-2}	2.4460×10^{9}
1.4790×10^{-2}	2.7520×10^{9}
1.5000×10^{-2}	2.7670×10^{9}
1.5600×10^{-2}	2.8960×10^{9}

Applying the least square method, the residuals γ_i at each data point is

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

The sum of square of the residuals is

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

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

Again, to find the constant E, we need to minimize S_r by differentiating with respect to E and then equating to zero

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

From there, we obtain

$$E = \frac{\sum_{i=1}^{n} \sigma_i \varepsilon_i}{\sum_{i=1}^{n} \varepsilon_i^{\,2}} \tag{25}$$

Note, Equation (25) only so far has shown that it corresponds to a local minimum or maximum. Can you show that it corresponds to an absolute minimum.
The summations used in Equation (25) are given in the Table 9.

Table 9 Tabulation for Example 2 for needed summations

i	ε	σ	ε^2	$\varepsilon\sigma$
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}

$$n = 12$$

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

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

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

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

$= 182.84 \text{ GPa}$

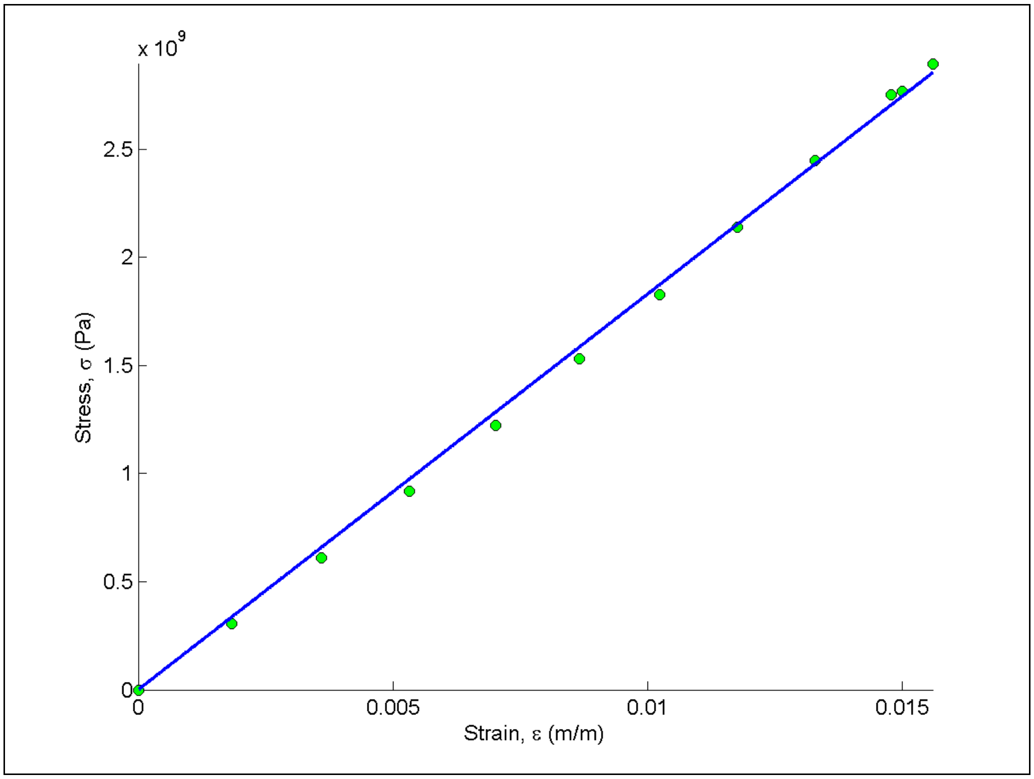

Figure 5 Linear regression model of stress vs. strain for a composite material.

QUESTION:

Given n data pairs, $(x_1, y_1), \ldots, (x_n, y_n)$, do the values of the two constants a_0 and a_1 in the least squares straight-line regression model $y = a_0 + a_1 x$ correspond to the absolute minimum of the sum of the squares of the residuals? Are these constants of regression unique?

ANSWER:

Given n data pairs $(x_1, y_1), \ldots, (x_n, y_n)$, the best fit for the straight-line regression model

$$y = a_0 + a_1 x \quad \text{(A.1)}$$

is found by the method of least squares. Starting with the sum of the squares of the residuals S_r

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \quad \text{(A.2)}$$

and using

$$\frac{\partial S_r}{\partial a_0} = 0 \quad \text{(A.3)}$$

$$\frac{\partial S_r}{\partial a_1} = 0 \quad \text{(A.4)}$$

gives two simultaneous linear equations whose solution is

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{(A.5a)}$$

$$a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{(A.5b)}$$

But do these values of a_0 and a_1 give the absolute minimum of value of S_r (Equation (A.2))? The first derivative analysis only tells us that these values give a local minima or maxima of S_r, and not whether they give an absolute minimum or maximum. So, we still need to figure out if they correspond to an absolute minimum.

We need to first conduct a second derivative test to find out whether the point (a_0, a_1) from Equation (A.5) gives a local minimum or local maximum of S_r. Only then can we proceed to show if this local minimum (or maximum) also corresponds to the absolute minimum (or maximum).

What is the second derivative test for a local minimum of a function of two variables?

If you have a function $f(x, y)$ and we found a critical point (a,b) from the first derivative test, then (a,b) is a minimum point if

$$\frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 > 0 \text{, and} \quad \text{(A.6)}$$

$$\frac{\partial^2 f}{\partial x^2} > 0 \text{ OR } \frac{\partial^2 f}{\partial y^2} > 0 \quad \text{(A.7)}$$

From Equation (A.2)

$$\frac{\partial S_r}{\partial a_0} = \sum_{i=1}^{n} 2(y_i - a_0 - a_1 x_i)(-1)$$
$$= -2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i) \quad \text{(A.8)}$$

$$\frac{\partial S_r}{\partial a_1} = \sum_{i=1}^{n} 2(y_i - a_0 - a_1 x_i)(-x_i)$$
$$= -2\sum_{i=1}^{n}(x_i y_i - a_0 x_i - a_1 x_i^2) \quad \text{(A.9)}$$

then

$$\frac{\partial^2 S_r}{\partial a_0^2} = -2\sum_{i=1}^{n} -1 = 2n \quad \text{(A.10)}$$

$$\frac{\partial^2 S_r}{\partial a_1^2} = 2\sum_{i=1}^{n} x_i^2 \quad \text{(A.11)}$$

$$\frac{\partial^2 S_r}{\partial a_0 \partial a_1} = 2\sum_{i=1}^{n} x_i \quad \text{(A.12)}$$

So, we satisfy condition (A.7) because from Equation (A.10) we see that $2n$ is a positive number. Although not required, from Equation (A.11) we see that $2\sum_{i=1}^{n} x_i^2$ is also a positive number as assuming that all x data points are NOT zero is reasonable.

Is the other condition (Equation (A.6)) for S_r being a minimum met? Yes, we can show (*proof not given that the term is positive*)

$$\begin{aligned}\frac{\partial^2 S_r}{\partial a_0^2}\frac{\partial^2 S_r}{\partial a_1^2} - \left(\frac{\partial^2 S_r}{\partial a_0 \partial a_1}\right)^2 &= (2n)\left(2\sum_{i=1}^{n} x_i^2\right) - \left(2\sum_{i=1}^{n} x_i\right)^2 \\ &= 4\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] \\ &= 4\sum_{\substack{i=1 \\ i<j}}^{n} (x_i - x_j)^2 > 0 \end{aligned} \quad \text{(A.13)}$$

So the values of a_0 and a_1 that we have in Equation (A.5) do correspond to a local minimum of S_r. But, is this local minimum also an absolute minimum. Yes, as given by Equation (A.5), the first derivatives of S_r are zero at *only one* point. This observation also makes the straight-line regression model based on least squares to be unique.

As a side note, the denominator in Equations (A.5) is nonzero as shown by Equation (A.13). This shows that the values of a_0 and a_1 are finite.

LINEAR REGRESSION	
Topic	Linear Regression
Summary	Textbook notes of Linear Regression
Major	Industrial Engineering
Authors	Egwu Kalu, Autar Kaw, Cuong Nguyen
Date	November 15, 2012
Web Site	http://numericalmethods.eng.usf.edu