Holistic Numerical Methods Institute
committed to bringing numerical methods to undergraduates

Multiple-Choice Test
Linear Regression
Regression
COMPLETE SOLUTION SET

1. Given $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, best fitting data to $y = f(x)$ by least squares requires minimization of

(A) $\sum_{i=1}^{n} [y_i - f(x_i)]$

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

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

(D) $\sum_{i=1}^{n} [y_i - \bar{y}]^2$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

Solution

The correct answer is (C).

A measure of goodness of fit, that is, how $f(x)$ predicts the response variable y is the magnitude of the residual E_i at each of the n data points.

$$E_i = y_i - f(x_i),\ i = 1,2,.....,n-1,n$$

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

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

Thus, best fitting data to $y = f(x)$ by least squares requires minimization of

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

2. The following data

x	1	20	30	40
y	1	400	800	1300

is regressed with least squares regression to $y = a_0 + a_1 x$. The value of a_1 most nearly is

A) 27.480
B) 28.956
C) 32.625
D) 40.000

Solution

The correct answer is (C).

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

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

Since

$$n = 4$$

$$\sum_{i=1}^{4} x_i y_i = 1\times 1 + 20\times 400 + 30\times 800 + 40\times 1300 = 84001$$

$$\sum_{i=1}^{4} x_i = 1 + 20 + 30 + 40 = 91$$

$$\sum_{i=1}^{4} y_i = 1 + 400 + 800 + 1300 = 2501$$

$$\sum_{i=1}^{4} x_i^2 = 1^2 + 20^2 + 30^2 + 40^2 = 2901$$

then

$$a_1 = \frac{4\times 84001 - 91\times 2501}{4\times 2901 - (91)^2}$$

$$= \frac{108413}{3323}$$

$$= 32.625$$

3. The following data

x	1	20	30	40
y	1	400	800	1300

is regressed with least squares regression to $y = a_1 x$. The value of a_1 most nearly is

A) 27.480
B) 28.956
C) 32.625
D) 40.000

Solution

The correct answer is (B).

Using the least squares criterion, we minimize

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

To find a_1 we minimize S_r with respect to a_1.

$$\frac{dS_r}{da_1} = 2\sum_{i=1}^{n} \left(y_i - a_1 x_i \right)\left(- x_i \right) = 0$$

giving

$$-\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} a_1 x_i^2 = 0$$

$$-\sum_{i=1}^{n} y_i x_i + a_1 \sum_{i=1}^{n} x_i^2 = 0$$

Solving the above equation for a_1 gives

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

Can you show that this corresponds to a local minimum, and then the absolute minimum.

Since

$$n = 4$$

$$\sum_{i=1}^{4} x_i \, y_i = 1 \times 1 + 20 \times 400 + 30 \times 800 + 40 \times 1300 = 84001$$

$$\sum_{i=1}^{4} x_i^2 = 1^2 + 20^2 + 30^2 + 40^2 = 2901$$

then

$$a_1 = \frac{84001}{2901}$$
$$= 28.956$$

4. An instructor gives the same y vs. x data as given below to four students and asks them to regress the data with least squares regression to $y = a_0 + a_1 x$.

x	1	10	20	30	40
y	1	100	400	600	1200

They each come up with four different answers for the straight-line regression model. Only one is correct. The correct model is

A) $y = 60x - 1200$

B) $y = 30x - 200$

C) $y = -139.43 + 29.684x$

D) $y = 1 + 22.782x$

Solution

The correct answer is (C).

We know for the straight line regression model $y = a_0 + a_1 x$ the values of a_0 and a_1 are

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

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

Since

$$n = 5$$

$$\sum_{i=1}^{5} x_i y_i = 1\times 1 + 10\times 100 + 20\times 400 + 30\times 600 + 40\times 1200 = 75001$$

$$\sum_{i=1}^{5} x_i = 1 + 10 + 20 + 30 + 40 = 101$$

$$\sum_{i=1}^{5} y_i = 1 + 100 + 400 + 600 + 1200 = 2301$$

$$\sum_{i=1}^{5} x_i^2 = 1^2 + 10^2 + 20^2 + 30^2 + 40^2 = 3001$$

then

$$a_1 = \frac{5\times 75001 - 101\times 2301}{5\times 3001 - (101)^2}$$

$$= \frac{142604}{4804}$$

$$= 29.684$$

$$a_0 = \frac{2301}{5} - 29.684 \times \frac{101}{5}$$
$$= -139.42$$

Hence, the linear regression model is

$$y = 29.684x - 139.42$$

Can you think of a different way to do the problem?

5. A torsion spring of a mousetrap is twisted through an angle of 180°. The torque vs. angle data is given below.

Torsion, T (N-m)	0.110	0.189	0.230	0.250
Angle, θ (rad)	0.10	0.50	1.1	1.5

The relationship between the torque and the angle is $T = a_0 + a_1\theta$.

The amount of strain energy stored in the mousetrap spring in Joules is

A) 0.29872
B) 0.41740
C) 0.84208
D) 1561.8

Solution

The correct answer is (C).

The linear regression curve for the torque is $T = a_0 + a_1\theta$. The values of a_0 and a_1 are

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

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

Since

$$n = 4$$

$$\sum_{i=1}^{4}\theta_i T_i = 0.10\times 0.110 + 0.50\times 0.189 + 1.1\times 0.230 + 1.5\times 0.250 = 0.7335$$

$$\sum_{i=1}^{4}\theta_i = 0.10 + 0.50 + 1.1 + 1.5 = 3.2$$

$$\sum_{i=1}^{4} T_i = 0.110 + 0.189 + 0.230 + 0.250 = 0.779$$

$$\sum_{i=1}^{4}\theta_i^2 = 0.10^2 + 0.50^2 + 1.1^2 + 1.5^2 = 3.72$$

then

$$a_1 = \frac{4\times 0.7335 - 3.2\times 0.779}{4\times 3.72 - 3.2^2}$$

$$= \frac{0.4412}{4.64}$$

$$= 0.095086 \text{ N - m/radian}$$

$$a_0 = \frac{0.779}{4} - 0.095086 \times \frac{3.2}{4}$$
$$= 0.11868 \text{ N - m}$$

Hence, the linear regression model is

$$T = 0.095086\theta + 0.11868 \ \text{N} - m$$

The potential energy U stored in the spring from $\theta = 0$ to $\theta = \pi$ is be given by

$$U = \int_0^{\pi} T d\theta$$
$$= \int_0^{\pi} (0.095086\theta + 0.11868) d\theta$$

$$= \left[\frac{0.095086}{2} \theta^2 + 0.11868\theta \right]_0^{\pi}$$
$$= \left[0.047543\theta^2 + 0.11868\theta \right]_0^{\pi}$$
$$= \left(0.047543 \times \pi^2 + 0.11868 \times \pi \right) - \left(0.047543 \times 0^2 + 0.11868 \times 0 \right)$$
$$= 0.84208 \text{ J}$$

6. A scientist finds that regressing the y vs. x data given below to $y = a_0 + a_1 x$ results in the coefficient of determination for the straight-line model, r^2 to being zero.

x	1	3	11	17
y	2	6	22	?

The missing value for y at $x = 17$ most nearly is

A) -2.4444
B) 2.0000
C) 6.8889
D) 34.000

Solution
The correct answer is (A).

Let

$$y(17) = b$$

Since

$$r^2 = 0$$

this means that

$$S_t = S_r$$

where

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and

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

The two being equal is possible if $a_1 = 0$ and $a_0 = \bar{y}$. If $a_1 = 0$, then

$$n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = 0$$

$$4(1\times 2 + 3\times 6 + 11\times 22 + 17\times b) - (1+3+11+17)(2+6+22+b) = 0$$

$$b = -2.4444$$

Extra notes for the student
What is a_0?

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

$$= 6.8889$$

Is a_0 the average value of y?

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$= \frac{\sum_{i=1}^{4} y_i}{4}$$

$$= \frac{(2+6+22-2.4444)}{4}$$

$$= 6.8889$$

Question:

Can you prove that if $r^2 = 0$ for a straight line regression model $y = a_0 + a_1 x$ then the regression model reduces to $y = \bar{y}$?

Hint: Solve the three equations

$$\frac{\partial S_r}{\partial a_0} = 0$$

$$\frac{\partial S_r}{\partial a_1} = 0$$

$$S_t = S_r$$

This will give $a_1 = 0$ and $a_0 = \bar{y}$.

Check the blog entry also: http://numericalmethods.eng.usf.edu/blog_entries.html