LINEAR REGRESSION

Complete the following problems to reinforce your understanding of the concept covered in this module.

Problem 1:

The following data represents the high school versus college GPA for a selected number of students. Determine the least squares regression line.

Student	HS GPA	College GPA
1	2.0	1.6
2	2.2	2.0
3	2.6	1.8
4	2.7	2.8
5	2.8	2.1
6	3.1	2.0
7	2.9	2.6
8	3.2	2.2
9	3.3	2.6
10	3.6	3.0

Problem 2:

Determine the correlation coefficient for the data given in the previous problem.

Problem 3:

A certain experiment yielded the data points (-3,70), (1,21), (-7,110), and (5,-35). Determine the least squares regression line and the goodness of fit of the line.

LINEAR REGRESSION

Solution 1:

The least squares regression line is given in the standard form $y = mx + b$ and can be found using the following three step process:

Step 1 – Calculate:

$$\sum x_i \qquad \sum x_i^2 \qquad (\sum x_i)^2 \qquad \overline{x} = \frac{1}{n}(\sum x_i) \qquad \sum x_i y_i$$

$$\sum y_i \qquad \sum y_i^2 \qquad (\sum y_i)^2 \qquad \overline{y} = \frac{1}{n}(\sum y_i)$$

Expanding the date table given, we find the values to be:

x_i	x_i^2	y_i	y_i^2	$x_i y_i$
2.0	4	1.6	2.56	3.2
2.2	4.84	2.0	4	4.4
2.6	6.76	1.8	3.24	4.68
2.7	7.29	2.8	7.84	7.56
2.8	7.84	2.1	4.41	5.88
3.1	9.61	2.0	4	6.20
2.9	8.41	2.6	6.76	7.54
3.2	10.24	2.2	4.84	7.04
3.3	10.89	2.6	6.76	8.58
3.6	12.96	3.0	9	10.80
$\sum x_i = 28.4$	$\sum x_i^2 = 82.84$	$\sum y_i = 22.7$	$\sum y_i^2 = 53.41$	$\sum x_i y_i = 65.88$

$$\sum (x_i)^2 = 806.56 \qquad \overline{x} = \frac{1}{10}(28.4) = 2.84$$

$$\sum (y_i)^2 = 515.29 \qquad \overline{y} = \frac{1}{10}(22.7) = 2.27$$

Step 2 - Calculate the slope of the line given by:

$$m=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{n\sum x_i^2-(\sum x_i)^2}=\frac{10(65.88)-(28.4)(22.7)}{10(82.84)-806.56}=.65$$

Step 3 - Determine the y-intercept:

$$b=\overline{y}-m\overline{x}$$

$$b=2.27-.65(2.84)=.434$$

Therefore, the least regression line that represents this date is given by the equation:

$$y=.65x+.434$$

Solution 2:

The correlation coefficient, r, also known as the goodness of fit, is calculated using the following formula:

$$r=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{(n\sum x_i^2-(\sum x_i)^2)(n\sum y_i^2-(\sum y_i)^2)}}$$

The data again is:

x_i	x_i^2	y_i	y_i^2	$x_i y_i$
2.0	4	1.6	2.56	3.2
2.2	4.84	2.0	4	4.4
2.6	6.76	1.8	3.24	4.68
2.7	7.29	2.8	7.84	7.56
2.8	7.84	2.1	4.41	5.88
3.1	9.61	2.0	4	6.20
2.9	8.41	2.6	6.76	7.54
3.2	10.24	2.2	4.84	7.04
3.3	10.89	2.6	6.76	8.58
3.6	12.96	3.0	9	10.80
$\sum x_i=28.4$	$\sum x_i^2=82.84$	$\sum y_i=22.7$	$\sum y_i^2=53.41$	$\sum x_i y_i=65.88$

$$\sum(x_i)^2 = 806.56 \qquad \bar{x} = \frac{1}{10}(28.4) = 2.84$$

$$\sum(y_i)^2 = 515.29 \qquad \bar{y} = \frac{1}{10}(22.7) = 2.27$$

So the correlation coefficient is:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n\sum x_i^2 - (\sum x_i)^2)(n\sum y_i^2 - (\sum y_i)^2)}} = \frac{10(65.88) - (28.4)(22.7)}{\sqrt{(10(82.4) - 806.56)(10(53.41) - 515.29)}} = .69$$

The line has a correlation coefficient of .69. Typically, if r exceeds .85, the fit is good, otherwise, the fit is poor. It would not be accurate to conclude on any correlations between the set of data points.

Solution 3:

The least squares regression line is given in the standard form $y = mx + b$ and can be found using the following three step process:

Step 1 – Calculate:

$$\sum x_i \qquad \sum x_i^2 \qquad (\sum x_i)^2 \qquad \bar{x} = \frac{1}{n}(\sum x_i) \qquad \sum x_i y_i$$

$$\sum y_i \qquad \sum y_i^2 \qquad (\sum y_i)^2 \qquad \bar{y} = \frac{1}{n}(\sum y_i)$$

Expanding the date table given, we find the values to be:

x_i	x_i^2	y_i	y_i^2	$x_i y_i$
-3	9	70	4800	-210
1	1	21	441	21
-7	49	110	12100	-770
5	25	-35	1225	-175
$\sum x_i = -4$	$\sum x_i^2 = 84$	$\sum y_i = 166$	$\sum y_i^2 = 18566$	$\sum x_i y_i = -1134$

$$\sum(x_i)^2 = 16 \qquad \bar{x} = \frac{1}{4}(-4) = -1$$

$$\sum(y_i)^2 = 27556 \qquad \bar{y} = \frac{1}{4}(166) = 41.5$$

Step 2 - Calculate the slope of the line given by:

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{4(-1134) - (-4)(166)}{4(84) - 16} = -12.1$$

Step 3 - Determine the y-intercept:

$$b = \bar{y} - m\bar{x}$$

$$b = 41.5 + 12.1(-1) = 29.4$$

Therefore, the least regression line that represents this date is given by the equation:

$$y = -12.1x + 29.4$$

To determine the goodness of fit we calculate the correlation coefficient, r:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n\sum x_i^2 - (\sum x_i)^2)(n\sum y_i^2 - (\sum y_i)^2)}} = \frac{4(-1134) - (-4)(166)}{\sqrt{(4(84) - 16)(4(18566) - 27556)}} = -1.0$$

The line has a correlation coefficient of .69 which means the fit is a perfect straight line.