LINEAR REGRESSION

Often times it may be necessary to draw a straight line through a set of data points. Using the method of squares, one can define the least squares regression line that best represents the given data.

The regression line is given in the standard form $y = mx + b$ and can be defined using a four step process.

Step 1 - Calculate:

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

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

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

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

Step 3 - Determine the y-intercept:

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

Step 4 - Determine the goodness of fit by calculating the correlation coefficient, r:

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

Typically, if r exceeds .85, the fit is good, otherwise, the fit is poor. If r equals 1, the fit is a perfect straight line.

Concept Example:

The following problem introduces the concept reviewed within this module. Use this content as a primer for the subsequent material.

Given the following data points, determine the least squares regression line:

i	x_i	y_i
1	1.2	1.1
2	2.3	2.1
3	3.0	3.1
4	3.8	4.0
5	4.7	4.9
6	5.9	5.9

Solution:

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	$x_i y_i$
1.2	1.4	1.1	1.2	1.3
2.3	5.3	2.1	4.4	4.8
3.0	9.0	3.1	9.6	9.3
3.8	14.3	4.0	16.0	15.2
4.7	22.1	4.9	24.0	23.0
5.9	34.8	5.9	34.8	34.8
$\sum x_i = 20.9$	$\sum x_i^2 = 86.9$	$\sum y_i = 21.1$	$\sum y_i^2 = 90$	$\sum x_i y_i = 88.4$

$$\sum\left(x_i\right)^2 = 436.8 \qquad \overline{x} = \frac{1}{6}\left(20.9\right) = 3.48$$

$$\sum\left(y_i\right)^2 = 445.2 \qquad \overline{y} = \frac{1}{6}\left(21.1\right) = 3.52$$

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

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{6(88.4) - (20.9)(21.1)}{6(86.9) - 436.8} = 1.06$$

Step 3 - Determine the y-intercept:

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

$$b = 3.52 - 1.06\left(3.48\right) = -.17$$

Therefore, the least regression line is given by the equation:

$$y = 1.06x + .17$$