EX: An engineer wishes to quickly determine whether a straight line adequately describes the current-versus-voltage characteristics of a new semiconductor device. The engineer measures the four data points listed below:

(1 V, 5 μA) (2 V, 3 μA) (3 V, 5 μA) (4 V, 11 μA)

a) Use simple linear regression to fit a straight line to the data.

b) Calculate the coefficient of determination, R^2, for the straight line fit.

c) Based on (b), is the device linear?

Find the sample mean of the data.

SOL'N: a) We use (x_i, y_i) to denote data points given in the problem, and we first compute the sample mean of the x values and y values:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}x_1 + \ldots + \frac{1}{n}x_n$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}y_1 + \ldots + \frac{1}{n}y_n$$

Here, the number of data points, n, is four.

We obtain the following sample-mean values:

$$\bar{x} = \frac{1+2+3+4}{4}\text{V} = 2.5\text{V} \quad \text{and} \quad \bar{y} = \frac{5+3+5+11}{4}\mu\text{A} = 6\mu\text{A}$$

NOTE: The x_i need *not* be evenly spaced for linear regression.

Now we compute the straight-line fit:

$$\hat{y} = a_1 x_i + a_2$$

where the coefficients are given by the following formulae (see Ref) derived by minimizing the total squared error for the estimated y values.

$$a_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad \text{and} \quad a_2 = \bar{y} - a_1\bar{x}$$

The calculation is follows:

$$a_1 = \frac{(1-2.5)(5-6)+(2-2.5)(3-6)+(3-2.5)(5-6)+(4-2.5)(11-6)\text{V}\mu\text{A}}{(1-2.5)^2+(2-2.5)^2+(3-2.5)^2+(4-2.5)^2\text{V}^2}$$

or

$$a_1 = 2\mu\text{A}/\text{V}$$

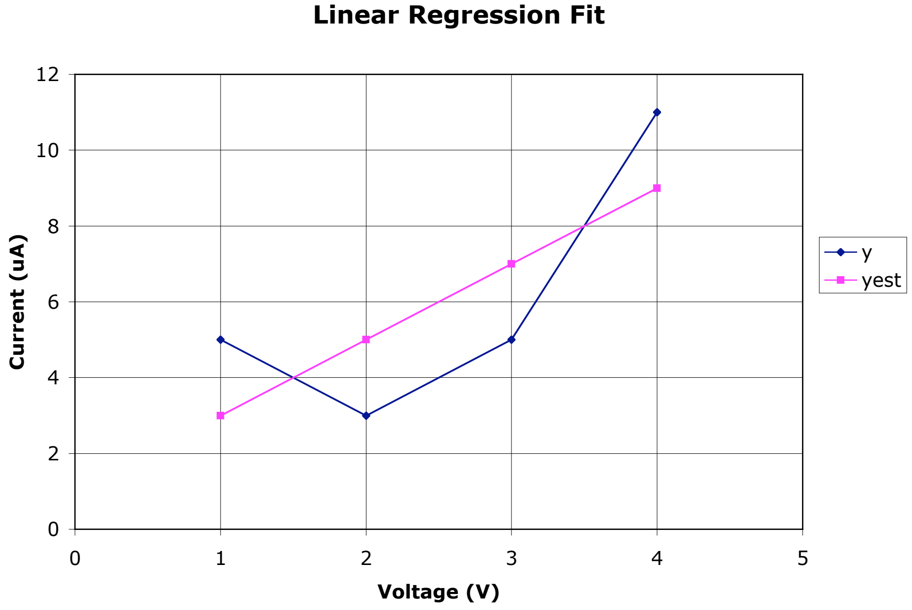
and

$$a_2 = 6\mu\text{A} - 2\mu\text{A}/\text{V}\cdot 2.5\text{V} = 1\mu\text{A}.$$

Using the x_i and the line fit, $\hat{y} = a_1 x_i + a_2 = 2x_i + 1$, we obtain the following estimated points:

(1 V, 3 μA)   (2 V, 5 μA)   (3 V, 7 μA)   (4 V, 9 μA)

The data and these estimates are shown in the plot, below.



b) To determine the quality of the line fit, we calculate the coefficient of determination, R^2:

$$r \equiv 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

The coefficient of determination measures what fraction of the squared error our linear fit removes compared with the squared error a horizontal line (or constant value) removes. In other words, R^2 tells us how much better the line fit is than just assuming the y values are constant, (with a value equal to the sample mean).

Using the data values from the problem, the value of R^2 is as follows:

$$R^2 = 1 - \frac{2^2 + (-2)^2 + (-2)^2 + 2^2}{(-1)^2 + (-3)^2 + (-1)^2 + 5^2} = 1 - \frac{16}{36} = \frac{5}{9} \approx 0.56$$

c) The value of R^2 is quite low, indicating that the linear fit may be unwarranted. We may have "overfit" the data by choosing a function that has more parameters than the data supports. This is always a danger. If we wish to use linear regression to predict future results, we want to be confident that our line fit has actually captured the nature of the device we are modeling. Based on the results presented above, the evidence is weak that the device under test is linear.

REF: Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, and Keying Ye, *Probability and Statistics for Engineers and Scientists,* 7th Ed., Upper Saddle River, NJ: Prentice Hall, 2002.