2.1 Introduction

An engineering curve fitting plays an important role in the analysis and interpretation of experimental data and in it's correlation with mathematical model formulated from fundamental engineering principles.

In the most general sense, curve fitting involves the determination of a continuous function.

$$\boldsymbol{y = f(x)}$$

Which results in the most "reasonable" or "best" fit of experimentally measured values $(x_1,y_1),(x_2,y_2)$.

In **curve fitting**, we are given ***n*** points (pairs of numbers) $(x_1, y_1),\ldots,(x_n, y_n)$ and we want to determine a function $f(x)$ such that:

$$\boldsymbol{f(x_1) \approx y_1}$$

For instance, we have four points:

(1.3, 0.103), (0.1, 1.099), (0.2, 0.808), (1.3, 1.897)

These points correspond to the interpolation polynomial

$$\boldsymbol{f(x) = x^3 - x + 1}$$

As in figure (2-1). However, if we graph the points, we see that they lie nearly on a straight line. Hence if these values are obtained in an experiment and thus involve an experimental error, and if the nature of the experiment suggests a linear relation, we better fit a straight line through the points.

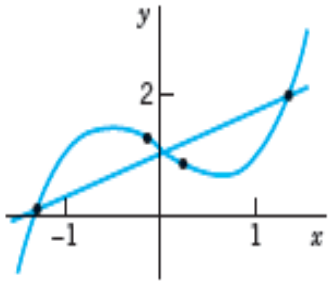

Figure (2-1) Approximate fitting of a straight line

Such a line may be useful for predicting values to be expected for other values of x. A widely used principle for fitting straight lines is the **method of least squares.**

2.2 Least Squares Approximation

This section discusses the need for approximating functions for sets of discrete data (e.g., for interpolation, differentiation, and integration), the desirable properties of approximating functions and the benefits of using polynomials for approximating functions.

2.2.1 The Straight Line Approximation

The simplest polynomial is a linear polynomial, the straight line. Least squares straight-line approximations are an extremely useful and common approximate fit. The least squares straight line fit is determined as follows. Given N data points, *(xi, Yi),* fit the best straight line through the set of data. The approximating function is

$$y = a + bx$$

Where **a, b** are coefficient representing the intercept and slope, respectively

The criteria of least square approximation for linear form states that **the sum of the squares of the distances of those points from the straight line is minimum,** where the distance is measured in the vertical direction (the y-direction). See figure (2.2).

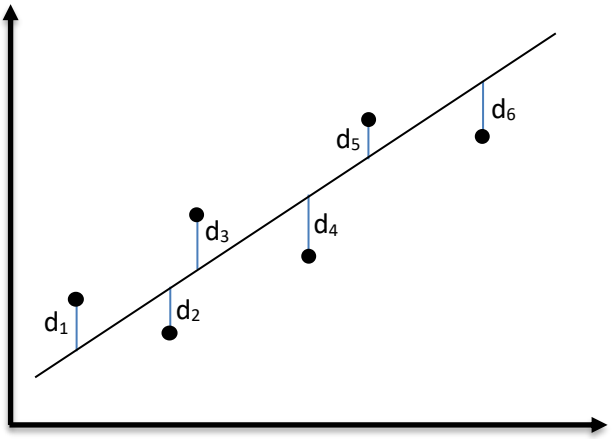

Figure (2.2) best line through data points

$$\boldsymbol{D = d_1^2 + d_2^2 + d_3^2 + \cdots + d_6^2}$$

The best linear model will have the smallest value for (D)

Now let see the figure (2.3), the point on the line with the abscissa x_j has the ordinate $a + bx_j$. Hence its distance from (x_j, y_j) is $|y_j - a - bx_j|$. The sum of the square is

$$q = \sum_{j=1}^{n} (y_j - a - bx_j)^2$$

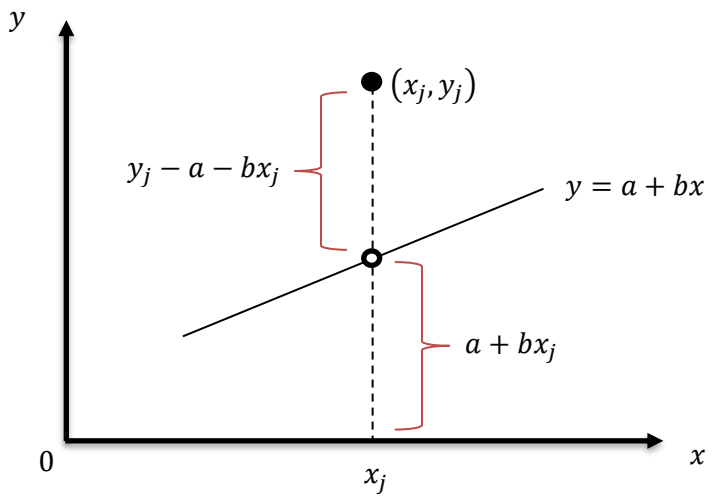

Figure (2.3) vertical distance of a point (x_j, y_j) from a straight line $y = a + bx$

A necessary condition for q to be minimum is

$$\frac{\partial q}{\partial a} = -2\sum(y_j - a - bx_j) = 0$$

$$\frac{\partial q}{\partial b} = -2\sum x_j(y_j - a - bx_j) = 0$$

(where we sum over j from 1 to n). Dividing by 2, writing each sum as three sums, and taking one of them to the right, we obtain the result

$$an + b\sum x_j = \sum y_j$$

$$a\sum x_j + b\sum x_j^2 = \sum x_j y_j$$

Solving the above equation simultaneously:

$$b = \frac{n\sum x_j y_j - \sum x_j \sum y_j}{n\sum x_j^2 - \left(\sum x_j\right)^2}$$

$$a = \bar{Y} - b\bar{X}$$

Where:

$$\bar{Y} = \frac{\sum y}{n} \qquad \bar{X} = \frac{\sum x}{n}$$

Example (2-1)

Using the method of least squares, fit a straight line to the x, y values in the following data

X_i	Y_i
1	**0.5**
2	**2.5**
3	**2.0**
4	**4.0**
5	**3.5**
6	**6.0**
7	**5.5**

Solution:

$n = 7, \ \sum x_i = 28, \ \sum x_i^2 = 140, \ \sum y_i = 24, \ \sum x_i y_i = 119.5$

$$\bar{X} = \frac{\sum x}{n} = \frac{28}{7} = 4, \qquad \bar{Y} = \frac{\sum y}{n} = \frac{24}{7} = 3.4286$$

$$b = \frac{n\sum x_j y_j - \sum x_j \sum y_j}{n\sum x_j^2 - \left(\sum x_j\right)^2} = \frac{7 * (119.5) - 28 * (24)}{7 * (140) - (28)^2} = 0.8393$$

$$a = \bar{Y} - b\bar{X} = 0.0714$$

$\therefore y = 0.0714 + 0.8393x$ → **The general solution**

H.W (1)

The percent reduction in area for normalized medium-Cast Steel specimens in a series of tension tests was found to vary a follows with the Carbon content of the Steel.

% Carbon in Steel	Reduction of area%
0.2	54.0
0.25	48.8
0.3	45.3
0.35	40.1
0.4	35.2
0.45	32.2
0.5	27.5

Determine a linear relationship between these variables in the form $y = a + bx$ which best fit the given data

H.W (2)

From Hooke's law ***F=Ks***, estimate the spring modulus ***(K)*** from the force ***(F)*** and elongation ***(s)*** where (F,s) =(1,0.3),(2,0.7),(4,1.3),(6,1.9),(10,3.2),(20,6.3)

2.3 Quantification of Error of Linear Regression

The approximation method of fitting several points in a line definitely contains an error can be expressed as the residuals. To evaluate this residual, recall the sum of the squares

$$q = \sum_{j=1}^{n} e_i^2 = \sum_{j=1}^{n} \left(y_j - a - bx_j\right)^2$$

The square of the residual represents the square of the vertical distance between the data and the estimated line through the data points. The standard error of the estimate can be computed from

$$s_{y/x} = \sqrt{\frac{q}{n-2}}$$

This represents the standard error of estimate in the case of the linear regression. The subscript (y/x) represents that the error for a predicted value of (y) corresponding to a particular value of (x). (n) Is the number of the points in the problem. The dividing by $(n-2)$ resulting from two data derived estimates $(a\ and\ b)$ which are used for computing the error (q). Another important factor is the standard deviation, which is a common measure of spread in general. It can be formulated as

$$s_y = \sqrt{\frac{q_T}{n-1}}$$

where q_T is the total sum of the squares of the residuals between the data points and the mean, or

$$q_T = \sum (y_i - \bar{Y})^2$$

Example (2.2)

Fit a straight line to the *x* and *y* values in the table below. Also, compute the total standard deviation and the standard error of the estimation.

Solution

$$n = 7, \sum x_i y_i = 119.5, \sum x_i^2 = 140, \sum x_i = 28\ , \sum y_i = 3.428571$$

$$\bar{X} = \frac{\sum x}{n} = \frac{28}{7} = 4, \qquad \bar{Y} = \frac{\sum y}{n} = \frac{24}{7} = 3.4286$$

$$b = \frac{n\sum x_j y_j - \sum x_j \sum y_j}{n\sum x_j^2 - \left(\sum x_j\right)^2} = \frac{7 * (119.5) - 28 * (24)}{7 * (140) - (28)^2} = 0.8393$$

$$a = \bar{Y} - b\bar{X} = 0.0714$$

$$\therefore y = 0.0714 + 0.8393x$$

$$s_y = \sqrt{\frac{q_T}{n-1}} = \sqrt{\frac{22.7143}{6}} = 1.9457$$

$$s_{y/x} = \sqrt{\frac{q}{n-2}} = \sqrt{\frac{2.9911}{5}} = 0.7735$$

x_i	y_i	$(y_i - \bar{Y})^2$	$(y_i - a - bx_i)^2$
1	0.5	8.5765	0.1687
2	2.5	0.8622	0.5625
3	2.0	2.0408	0.3473
4	4.0	0.3265	0.3273
5	3.5	0.0051	0.5896
6	6.0	6.6122	0.7972
7	5.5	4.2908	0.1993
Σ	24.0	22.7143	2.9911

2.4 Curve Fitting by Polynomials of Degree m

Section (2.2) exhibits deriving an equation of a straight line by using the least-square method. For many cases, using the polynomial regression gives better results and more accuracy than the linear regression. Our method of curve fitting can be generalized from a polynomial $\boldsymbol{y = a + bx}$ to a polynomial of degree m

$$P(x) = b_0 + b_1 x + b_2 x^2 + \cdots + b_m x^m$$

Where $m \leq n - 1$, then $\boldsymbol{q}$ takes the form

$$q = \sum_{j=1}^{n} \left(y_j - P(x_j)\right)^2$$

And depends on $(\boldsymbol{m + 1})$ parameters $b_0 \ldots . b_m$

$$\frac{\partial q}{\partial b_0} = 0, \qquad \frac{\partial q}{\partial b_m} = 0$$

Which give a system of $(\boldsymbol{m+1})$ normal equations. In the case of a quadratic polynomial

$$P(x) = b_0 + b_1 x + b_2 x^2$$

The normal equations are (summation from 1 to n)

$$b_0 n + b_1 \sum x_j + b_2 \sum x_j^2 = \sum y_j$$

$$b_0 \sum x_j + b_1 \sum x_j^2 + b_2 \sum x_j^3 = \sum x_i y_j$$

$$b_0 \sum x_j^2 + b_1 \sum x_j^3 + b_2 \sum x_j^4 = \sum x_j^2 y_j$$

Example (2-2)

Fit a parabola through the data (0,5),(2,4),(4,1),(6,6),(8,7).

Solution:

$$n = 5, \sum x_j = 20, \sum x_j^2 = 120, \sum x_j^3 = 800, \sum x_j^4 = 5664$$

$$\sum y_j = 23, \sum x_j y_j = 104, \sum x_j^2 y_j = 696$$

Hence:

$$5b_0 + 20b_1 + 120b_2 = 23$$

$$20b_0 + 120b_1 + 800b_2 = 104$$

$$120b_0 + 800b_1 + 566b_2 = 696$$

Solving them we obtain the quadratic least squares parabola

$$y = 5.11429 - 1.41429x + 0.21429x^2$$

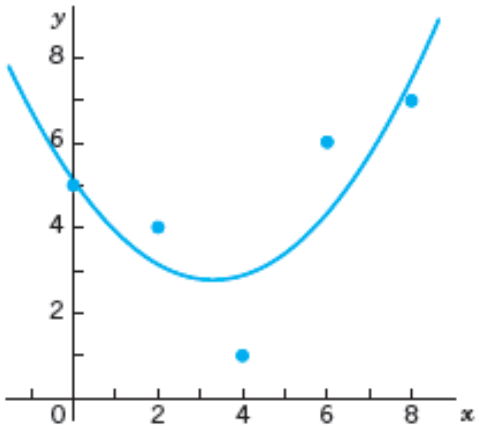

Figure (3-3) least square method of example (2-2)

H.W (3)

t [hr] = worker's time on duty, y [sec] = His/her, (her is reaction time), find a quadratic relation between (t,y), using the following points (t,y) =(1,2.0),(2,1.78),(3,1.9),(4,2.35),(5,2.7).

H.W (4)

Derive the formula for the normal equations of a cubic least squares parabola.