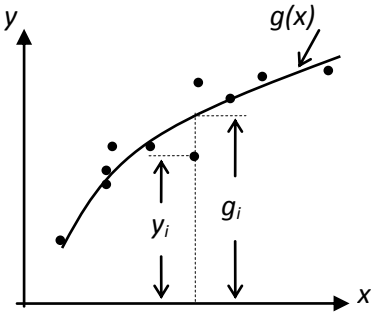
8- Curve Fitting

Least-squares criterion (linear regression)

Let (x_1, y_1), (x_2, y_2), ….., (x_n, y_n) a set of observations to be modeled, $g(x)$ is the approximating model, and e is the local error (residual) between the observations and the model, that is $e_i = g_i - y_i$. In the least squares method, to get a good approximating model, the total error (which is the sum of the squares of the local errors around the regression line) $E = \sum_{i=1}^{n} e_i^2$ must be minimized.

Let $g(x) = a_o + a_1 x$ (1st order polynomial, i.e. a straight line),



$$E = \sum_{i=1}^{n} (g_i - y_i)^2 \quad \Rightarrow \quad E = \sum_{i=1}^{n} (a_o + a_1 x_i - y_i)^2 ,$$

The total error E is minimized if $\frac{\partial E}{\partial a_o} = 0$ and $\frac{\partial E}{\partial a_1} = 0$.

$$\frac{\partial E}{\partial a_o} = 2\sum_{i=1}^{n} (a_o + a_1 x_i - y_i) \quad \Rightarrow \quad 2\sum_{i=1}^{n} (a_o + a_1 x_i - y_i) = 0 ,$$

$$\sum_{i=1}^{n} a_o + \sum_{i=1}^{n} a_1 x_i - \sum_{i=1}^{n} y_i = 0 . \qquad \text{But} \quad \sum_{i=1}^{n} a_o = n.a_o ,$$

$$\therefore \quad n.a_o + \sum_{i=1}^{n} x_i a_1 = \sum_{i=1}^{n} y_i . \qquad \text{…….. (1)}$$

Similarly $\frac{\partial E}{\partial a_1} = 2\sum_{i=1}^{n} (a_o + a_1 x_i - y_i) x_i \quad \Rightarrow \quad 2\sum_{i=1}^{n} (a_o x_i + a_1 x_i^2 - x_i y_i) = 0 ,$

$$\sum_{i=1}^{n} x_i a_o + \sum_{i=1}^{n} x_i^2 a_1 - \sum_{i=1}^{n} x_i y_i = 0 , \quad \Rightarrow \quad \sum_{i=1}^{n} x_i a_o + \sum_{i=1}^{n} x_i^2 a_1 = \sum_{i=1}^{n} x_i y_i \qquad \text{…..... (2)}$$

In matrix form:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{Bmatrix} a_o \\ a_1 \end{Bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} .$$

Generally, if $g(x)=a_o+a_1x+a_2x^2+.....+a_kx^k$ (kth order polynomial), we will have

$$\begin{bmatrix} n & \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 & .... & \sum_{i=1}^{n}x_i^k \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i^3 & .... & \sum_{i=1}^{n}x_i^{k+1} \\ \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i^3 & \sum_{i=1}^{n}x_i^4 & .... & \sum_{i=1}^{n}x_i^{k+2} \\ .... & .... & .... & .... & .... \\ \sum_{i=1}^{n}x_i^k & \sum_{i=1}^{n}x_i^{k+1} & \sum_{i=1}^{n}x_i^{k+2} & .... & \sum_{i=1}^{n}x_i^{2k} \end{bmatrix} \begin{Bmatrix} a_o \\ a_1 \\ a_2 \\ ... \\ a_k \end{Bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}y_i \\ \sum_{i=1}^{n}x_iy_i \\ \sum_{i=1}^{n}x_i^2y_i \\ .... \\ \sum_{i=1}^{n}x_i^ky_i \end{bmatrix}.$$

Statistical definitions

$\bar{y}$ is the mean of y.

E_m is the total sum of the squares around the mean of y, that is $E_m=\sum_{i=1}^{n}(y_i-\bar{y})^2$.

r^2 is the determination coefficient which is given by $r^2=\dfrac{E_m-E}{E_m}$.

r is the correlation coefficient which is given by $r=\sqrt{r^2}$.

For a perfect fit $(E=0) \quad \Rightarrow \quad r=r^2=1$, signifying that the approximating model $g(x)$ explains 100% of the variability of the data (observations).

Example 1: Given the following data:

x	0	1	2	3	4	5
$f(x)$	2.1	7.7	13.6	27.2	40.9	61.6

Using the least squares criterion:

1- Fit a 1st order polynomial (straight line) to this data.

2- Fit a 2nd order polynomial (quadratic equation) to this data.

Solution:

1- Let the straight line is $g(x)=a_o+a_1x$, then we have

$$\begin{bmatrix} n & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 \end{bmatrix} \begin{Bmatrix} a_o \\ a_1 \end{Bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}y_i \\ \sum_{i=1}^{n}x_iy_i \end{bmatrix},$$

$$n=6, \quad \sum_{i=1}^{n}x_i=0+1+2+3+4+5=15, \quad \sum_{i=1}^{n}x_i^2=0^2+1^2+2^2+3^2+4^2+5^2=55,$$

$$\sum_{i=1}^{n} y_i = 2.1 + 7.7 + 13.6 + 27.2 + 40.9 + 61.6 = 152.6,$$

$$\sum_{i=1}^{n} x_i y_i = 0(2.1) + 1(7.7) + 2(13.6) + 3(27.2) + 4(40.9) + 5(61.6) = 585.6.$$

$$\therefore \begin{bmatrix} 6 & 15 \\ 15 & 55 \end{bmatrix} \begin{Bmatrix} a_o \\ a_1 \end{Bmatrix} = \begin{bmatrix} 152.6 \\ 585.6 \end{bmatrix}.$$

Use Cramer's rule:

$$a_o = \frac{\begin{bmatrix} 152.6 & 15 \\ 585.6 & 55 \end{bmatrix}}{\begin{bmatrix} 6 & 15 \\ 15 & 55 \end{bmatrix}} = \frac{152.6(55) - 15(585.6)}{6(55) - 15(15)} = \frac{-391}{105} = -3.72381,$$

$$a_1 = \frac{\begin{bmatrix} 6 & 152.6 \\ 15 & 585.6 \end{bmatrix}}{\begin{bmatrix} 6 & 15 \\ 15 & 55 \end{bmatrix}} = \frac{6(585.6) - 152.6(15)}{6(55) - 15(15)} = \frac{1224.6}{105} = 11.66286.$$

$$\therefore \; g(x) = -3.72381 + 11.66286x.$$

2- Let the 2nd order polynomial is $q(x) = b_o + b_1 x + b_2 x^2$, then we have

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \end{bmatrix} \begin{Bmatrix} b_o \\ b_1 \\ b_2 \end{Bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \end{bmatrix},$$

$$\sum_{i=1}^{n} x_i^3 = 0^3 + 1^3 + 2^3 + 3^3 + 4^3 + 5^3 = 225, \quad \sum_{i=1}^{n} x_i^4 = 0^4 + 1^4 + 2^4 + 3^4 + 4^4 + 5^4 = 979,$$

$$\sum_{i=1}^{n} x_i^2 y_i = 0^2(2.1) + 1^2(7.7) + 2^2(13.6) + 3^2(27.2) + 4^2(40.9) + 5^2(61.6) = 2488.8.$$

$$\therefore \begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{Bmatrix} b_o \\ b_1 \\ b_2 \end{Bmatrix} = \begin{bmatrix} 152.6 \\ 585.6 \\ 2488.8 \end{bmatrix}.$$

Use Cramer's rule:

$$b_0=\frac{\begin{vmatrix}152.6 & 15 & 55\\ 585.6 & 55 & 225\\ 2488.8 & 225 & 979\end{vmatrix}}{\begin{vmatrix}6 & 15 & 55\\ 15 & 55 & 225\\ 55 & 225 & 979\end{vmatrix}}=\frac{152.6\begin{vmatrix}55 & 225\\ 225 & 979\end{vmatrix}+(-1)(585.6)\begin{vmatrix}15 & 55\\ 225 & 979\end{vmatrix}+2488.8\begin{vmatrix}15 & 55\\ 55 & 225\end{vmatrix}}{6\begin{vmatrix}55 & 225\\ 225 & 979\end{vmatrix}+(-1)(15)\begin{vmatrix}15 & 55\\ 225 & 979\end{vmatrix}+55\begin{vmatrix}15 & 55\\ 55 & 225\end{vmatrix}}=2.47857$$

$$b_1=\frac{\begin{vmatrix}6 & 152.6 & 55\\ 15 & 585.6 & 225\\ 55 & 2488.8 & 979\end{vmatrix}}{\begin{vmatrix}6 & 15 & 55\\ 15 & 55 & 225\\ 55 & 225 & 979\end{vmatrix}}=\frac{-152.6\begin{vmatrix}15 & 225\\ 55 & 979\end{vmatrix}+585.6\begin{vmatrix}6 & 55\\ 55 & 979\end{vmatrix}+(-1)(2488.8)\begin{vmatrix}6 & 55\\ 15 & 225\end{vmatrix}}{6\begin{vmatrix}55 & 225\\ 225 & 979\end{vmatrix}+(-1)(15)\begin{vmatrix}15 & 55\\ 225 & 979\end{vmatrix}+55\begin{vmatrix}15 & 55\\ 55 & 225\end{vmatrix}}=2.35929$$

$$b_2=\frac{\begin{vmatrix}6 & 15 & 152.6\\ 15 & 55 & 585.6\\ 55 & 225 & 2488.8\end{vmatrix}}{\begin{vmatrix}6 & 15 & 55\\ 15 & 55 & 225\\ 55 & 225 & 979\end{vmatrix}}=\frac{152.6\begin{vmatrix}15 & 55\\ 55 & 225\end{vmatrix}+(-1)(585.6)\begin{vmatrix}6 & 15\\ 55 & 225\end{vmatrix}+2488.8\begin{vmatrix}6 & 15\\ 15 & 55\end{vmatrix}}{6\begin{vmatrix}55 & 225\\ 225 & 979\end{vmatrix}+(-1)(15)\begin{vmatrix}15 & 55\\ 225 & 979\end{vmatrix}+55\begin{vmatrix}15 & 55\\ 55 & 225\end{vmatrix}}=1.86071$$

$$\therefore\ q(x)=2.47857+2.35929x+1.86071x^2.$$

Statistical comparison

x_i	y_i	$E_{m_i}=(y_i-\bar{y})^2$	For $g(x)$ $E_i=(g(x_i)-y_i)^2$	For $q(x)$ $E_i=(q(x_i)-y_i)^2$
0	2.1	548.340278	33.9167629	0.14331524
1	7.7	317.433611	0.0571449	1.00286204
2	13.6	142.006944	36.0229236	1.0815792
3	27.2	2.83361111	16.5223552	0.80491401
4	40.9	236.646944	4.11128342	0.61951067
5	61.6	1302.00694	49.1332304	0.65162027
$\sum$	153.1	2549.3	139.8	4.3
$\bar{y}=\frac{\sum y_i}{n}$	$\bar{y}=\frac{153.1}{6}=25.51667$	$r^2=\frac{E_m-E}{E_m}$	$r^2=\frac{2549.3-139.8}{2549.3}=0.9452$	$r^2=\frac{2549.3-4.3}{2549.3}=0.9983$

Since r^2, for $q(x)$, is closer to one, thus the quadratic equation $q(x)$ is better than the linear equation $g(x)$ in representing the given data.

Example 2: (Final 2014) The volume of water pumped by a pump is measured as a function of time as tabulated below:

Time, t, sec	0	1	5	8
Volume, V, m^3	2.1	7.7	13.6	27.2

Fit the equation $V = at + bt^3$ (where a and b are constants) to the above data using the least squares method.

Solution:

Since the required equation $V = at + bt^3$ is a 3rd order polynomial, thus, to make use of the general least squares matrix, we compare it with the general form of a 3rd order polynomial $g(t) = a_o + a_1 t + a_2 t^2 + a_3 t^3$. It is obvious that the first and third constants do not exist in the required equation, thus we cancel the first and third row and column of the general least squares (4×4) matrix,

$$\begin{bmatrix} n & \sum_{i=1}^{n} t_i & \sum_{i=1}^{n} t_i^2 & \sum_{i=1}^{n} t_i^3 \\ \sum_{i=1}^{n} t_i & \sum_{i=1}^{n} t_i^2 & \sum_{i=1}^{n} t_i^3 & \sum_{i=1}^{n} t_i^4 \\ \sum_{i=1}^{n} t_i^2 & \sum_{i=1}^{n} t_i^3 & \sum_{i=1}^{n} t_i^4 & \sum_{i=1}^{n} t_i^5 \\ \sum_{i=1}^{n} t_i^3 & \sum_{i=1}^{n} t_i^4 & \sum_{i=1}^{n} t_i^5 & \sum_{i=1}^{n} t_i^6 \end{bmatrix} \begin{Bmatrix} a_o \\ a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} V_i \\ \sum_{i=1}^{n} t_i V_i \\ \sum_{i=1}^{n} t_i^2 V_i \\ \sum_{i=1}^{n} t_i^3 V_i \end{bmatrix},$$

to get,

$$\begin{bmatrix} \sum_{i=1}^{n} t_i^2 & \sum_{i=1}^{n} t_i^4 \\ \sum_{i=1}^{n} t_i^4 & \sum_{i=1}^{n} t_i^6 \end{bmatrix} \begin{Bmatrix} a \\ b \end{Bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} t_i V_i \\ \sum_{i=1}^{n} t_i^3 V_i \end{bmatrix},$$

$$\sum_{i=1}^{n} t_i^2 = 0^2 + 1^2 + 5^2 + 8^2 = 90, \qquad \sum_{i=1}^{n} t_i^4 = 4722, \qquad \sum_{i=1}^{n} t_i^6 = 277770,$$

$$\sum_{i=1}^{n} t_i V_i = 232.2, \quad \text{and} \qquad \sum_{i=1}^{n} t_i^3 V_i = 13081.8.$$

$$\therefore \begin{bmatrix} 90 & 4722 \\ 4722 & 277770 \end{bmatrix} \begin{Bmatrix} a \\ b \end{Bmatrix} = \begin{bmatrix} 232.2 \\ 13081.8 \end{bmatrix}.$$

Solving the above matrix, we get: $a = 1.008852 \approx 1$ and $b = 0.029946 \approx 0.03$.

$\therefore$ The required equation is $V = t + 0.03t^3$.