8.1 Discrete Least Squares Approximation

Table 8.1

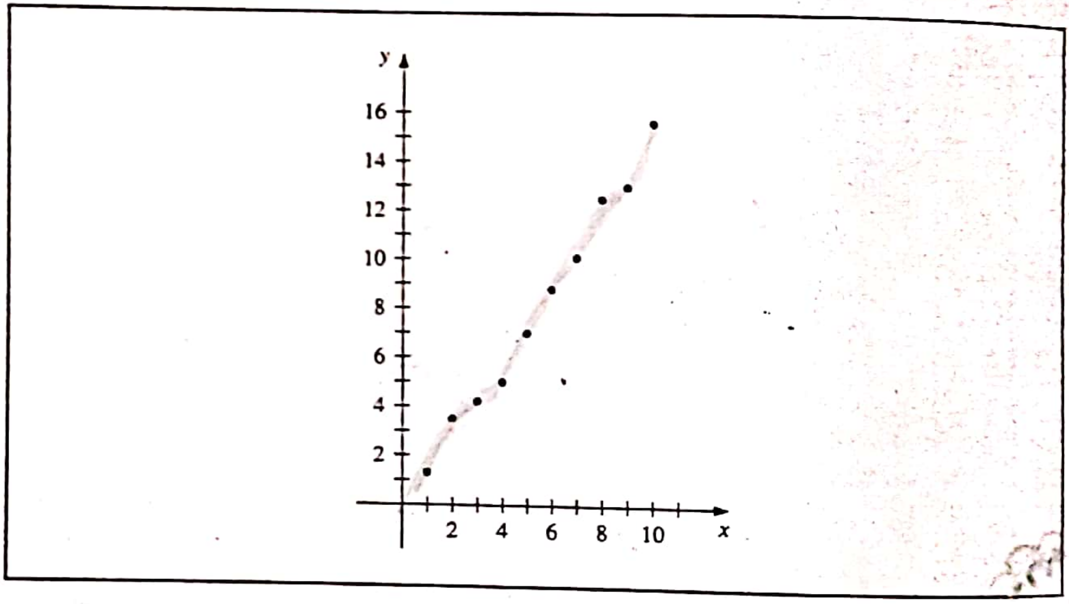
x_i	y_i	x_i	y_i
1	1.3	6	8.8
2	3.5	7	10.1
3	4.2	8	12.5
4	5.0	9	13.0
5	7.0	10	15.6

Consider the problem of estimating the values of a function at nontabulated points, given the experimental data in Table 8.1.

Figure 8.1 shows a graph of the values in Table 8.1. From this graph, it appears that the actual relationship between x and y is linear. The likely reason that no line precisely fits the data is because of errors in the data. So it is unreasonable to require that the approximating function agree exactly with the data. In fact, such a function would introduce oscillations that were not originally present. For example, the graph of the ninth-degree interpolating polynomial shown in unconstrained mode for the data in Table 8.1 is obtained in Maple using the commands

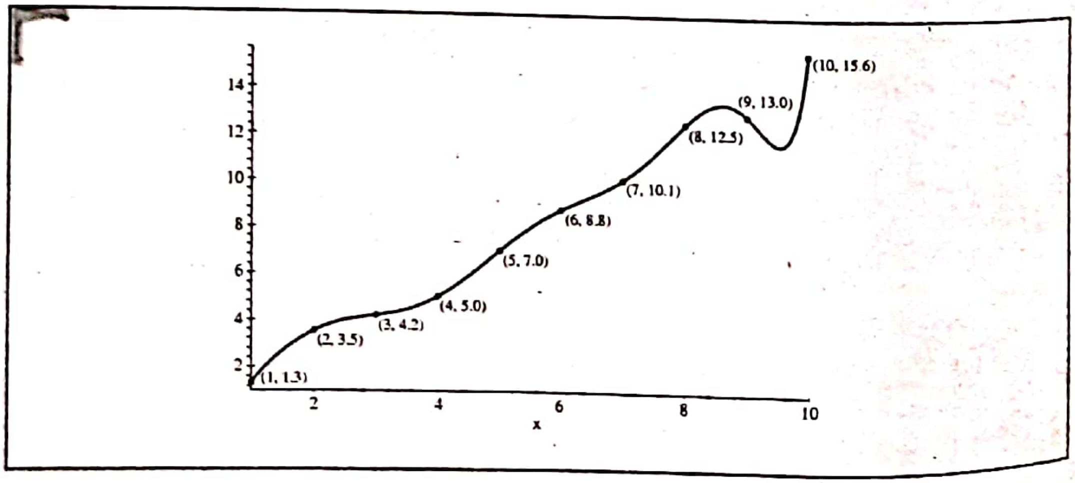
$p := interp([1,2,3,4,5,6,7,8,9,10],[1.3,3.5,4.2,5.0,7.0,8.8,10.1,12.5,13.0,15.6],x):$
$plot(p,x=1..10)$

Figure 8.1



The plot obtained (with the data points added) is shown in Figure 8.2.

Figure 8.2



This polynomial is clearly a poor predictor of information between a number of the data points. A better approach would be to find the "best" (in some sense) approximating line, even if it does not agree precisely with the data at any point.

Let $a_1x_i + a_0$ denote the ith value on the approximating line and y_i be the ith given y-value. We assume throughout that the independent variables, the x_i, are exact, it is the dependent variables, the y_i, that are suspect. This is a reasonable assumption in most experimental situations.

The problem of finding the equation of the best linear approximation in the absolute sense requires that values of a_0 and a_1 be found to minimize

$$E_\infty(a_0, a_1) = \max_{1\le i\le 10} \{|y_i - (a_1x_i + a_0)|\}.$$

This is commonly called a **minimax** problem and cannot be handled by elementary techniques.

Another approach to determining the best linear approximation involves finding values of a_0 and a_1 to minimize

(a+bxi))

$$E_1(a_0, a_1) = \sum_{i=1}^{10} |y_i - (a_1x_i + a_0)|.$$

This quantity is called the **absolute deviation.** To minimize a function of two variables, we need to set its partial derivatives to zero and simultaneously solve the resulting equations. In the case of the absolute deviation, we need to find a_0 and a_1 with

$$0 = \frac{\partial}{\partial a_0}\sum_{i=1}^{10} |y_i - (a_1x_i + a_0)| \quad \text{and} \quad 0 = \frac{\partial}{\partial a_1}\sum_{i=1}^{10} |y_i - (a_1x_i + a_0)|.$$

The problem is that the absolute-value function is not differentiable at zero, and we might not be able to find solutions to this pair of equations.

Linear Least Squares

The **least squares** approach to this problem involves determining the best approximating line when the error involved is the sum of the squares of the differences between the y-values on the approximating line and the given y-values. Hence, constants a_0 and a_1 must be found that minimize the least squares error:

$$E_2(a_0, a_1) = \sum_{i=1}^{10} \left[y_i - (a_1x_i + a_0)\right]^2.$$

The least squares method is the most convenient procedure for determining best linear approximations, but there are also important theoretical considerations that favor it. The minimax approach generally assigns too much weight to a bit of data that is badly in error, whereas the absolute deviation method does not give sufficient weight to a point that is considerably out of line with the approximation. The least squares approach puts substantially more weight on a point that is out of line with the rest of the data, but will not permit that point to completely dominate the approximation. An additional reason for considering the least squares approach involves the study of the statistical distribution of error. (See [Lar], pp. 463–481.) ?

The general problem of fitting the best least squares line to a collection of data $\{(x_i, y_i)\}_{i=1}^m$ involves minimizing the total error,

$y_i - (a+bx_i)]^2$

$$E \equiv E_2(a_0, a_1) = \sum_{i=1}^{m} \left[y_i - (a_1x_i + a_0)\right]^2,$$

with respect to the parameters a_0 and a_1. For a minimum to occur, we need both

$$\frac{\partial E}{\partial a_0} = 0 \quad \text{and} \quad \frac{\partial E}{\partial a_1} = 0,$$

that is,

$$0 = \frac{\partial}{\partial a_0} \sum_{i=1}^{m} [(y_i - (a_1 x_i - a_0)]^2 = 2 \sum_{i=1}^{m} (y_i - a_1 x_i - a_0)(-1)$$

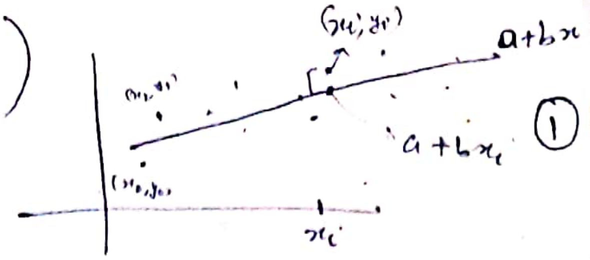
and

$$0 = \frac{\partial}{\partial a_1} \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)]^2 = 2 \sum_{i=1}^{m} (y_i - a_1 x_i - a_0)(-x_i).$$

These equations simplify to the **normal equations**:

$$a_0 \cdot m + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \quad \text{and} \quad a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i.$$

Curve fitting (I) (Straight line $y = a + bx$)



$(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$

$e_i = y_i - (a + bx_i)$

Least square Method

Minimize the sum of the squares of the errors ~~is~~

i.e $E = \sum_{i=1}^{n} e_i^2$

$$\text{Min } E = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

For minimize, $\frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b} = 0$

$$\frac{\partial E}{\partial a} = 0 \Rightarrow \sum_{i=1}^{n} 2[y_i - (a + bx_i)](-1) = 0 \Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a + b \sum_{i=1}^{n} x_i$$

$$\Rightarrow \sum_i y_i = na + b \sum x_i$$

$$\frac{\partial E}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} 2[y_i - (a + bx_i)](-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = 0$$

Normal equations

$$\left.\begin{aligned} \sum_i y_i &= na + b \sum_i x_i \\ \sum_i x_i y_i &= a \sum_i x_i + b \sum_i x_i^2 \end{aligned}\right\} \text{Find the values of } a \,\&\, b$$

then $y = a + bx$ is the best fit

~~Ex~~.

Q

x	0.2	0.4	0.6	0.8	1
y	0.447	0.632	0.775	0.895	1

$n = 5$ pt. Find Σx, Σy, Σx^2, Σxy

x	y	xy	x^2
0.2	0.447	0.0894	0.04
0.4	0.632	0.2528	0.16
0.6	0.775	0.465	0.36
0.8	0.894	0.7152	0.64
1	1	1	1
$\Sigma x = 3$	$\Sigma y = 3.748$	$\Sigma xy = 2.5224$	$\Sigma x^2 = 2.2$ ~~2.0064~~

Normal eq.

$3.748 = 5a + 3b$

$2.5224 = 3a + \overset{2.2}{\cancel{2.0064}}\,b$

find $a \,\&\, b$

$a = 0.3392$

$b = 0.684$

(II) Quadratic ~~form~~ (second degree poly)

$$y = a + bx + cx^2$$

$$e_i = y_i - (a + bx_i + cx_i^2)$$

$$\text{Min } E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (a + bx_i + cx_i^2)\right]^2$$

For minimization, $\frac{\partial E}{\partial a} = 0$, $\frac{\partial E}{\partial b} = 0$, $\frac{\partial E}{\partial c} = 0$

$$\frac{\partial E}{\partial a} = 0 \Rightarrow \sum_{i=1}^{n} 2\left[y_i - (a + bx_i + cx_i^2)\right](-1) \Rightarrow \sum_i y_i = na + b\sum x_i + c\sum x_i^2$$

$$\frac{\partial E}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} 2\left[y_i - (a + bx_i + cx_i^2)\right](-x_i) \Rightarrow \sum x_i y_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3$$

$$\frac{\partial E}{\partial c} = 0 \Rightarrow \sum_{i=1}^{n} 2\left[y_i - (a + bx_i + cx_i^2)\right](-x_i^2) \Rightarrow \sum x_i^2 y_i = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4$$

↓ Normal equations

Find the value of a, b & c

and then $y = a + bx + cx^2$ is second degree poly.

Q

x	-2	-1	0	1	2
y	15	1	1	3	19

$y = a + bx + cx^2$

$n = 5$

$\sum x$, $\sum y$, $\sum xy$, $\sum x^2 y$, $\sum x^2$, $\sum x^3$, $\sum x^4$

x	y	xy	x^2y	x^2	x^3	x^4
-2	15	-30	60	4	-8	16
-1	1	-1	1	1	-1	1
0	1	0	0	0	0	0
1	3	3	3	1	1	1
2	19	38	76	4	8	16
$\sum x = 0$	$\sum y = 39$	$\sum xy = 10$	$\sum x^2y = 140$	$\sum x^2 = 10$	$\sum x^3 = 0$	$\sum x^4 = 34$

Normal eq. $39 = 5a + (0)b + (10)c \Rightarrow$ ~~$a = 39$~~ $5a + 10c = 39$

$10 = (0)a + 10b + c(0) \Rightarrow b = 1$

$140 = 10a + b(0) + 34(c) \Rightarrow 10a + 34c = 140$

$$\Rightarrow a = \frac{-37}{35},\ b = 1,\ c = \frac{31}{7}$$

$$\therefore y = \frac{1}{35}(-37 + 35x + 155x^2)$$

Q Use the method of least square to fit the curve $y = ax + \frac{b}{\sqrt{x}}$. ③

x	0.2	0.3	0.5	1	2
y	16	14	11	6	3

Sol. $e_i = y_i - \left(ax_i + \frac{b}{\sqrt{x_i}}\right)$, $n = 5$

$$\text{Min } E = \sum_{i=1}^{5} e_i^2 = \sum_{i=1}^{5}\left[y_i - \left(ax_i + \frac{b}{\sqrt{x_i}}\right)\right]^2$$

For minimization,

$$\frac{\partial E}{\partial a} = 0 \Rightarrow \sum_{i=1}^{5} 2\left[y_i - \left(ax_i + \frac{b}{\sqrt{x_i}}\right)\right](-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{5} x_i y_i = a\sum_{i=1}^{5} x_i^2 + b\sum_{i=1}^{5}\sqrt{x_i}$$

$$\frac{\partial E}{\partial b} = 0 \Rightarrow \sum_{i=1}^{5} 2\left[y_i - \left(ax_i + \frac{b}{\sqrt{x_i}}\right)\right]\left(-\frac{1}{\sqrt{x_i}}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{5}\frac{y_i}{\sqrt{x_i}} = a\sum\sqrt{x_i} + b\sum_{i=1}^{5}\frac{1}{x_i}.$$

We need, ~~$\sum x_i$~~, ~~$\sum y_i$~~, $\sum x_i y_i$, $\sum x_i^2$, $\sum\sqrt{x_i}$, $\sum\frac{1}{x_i}$, $\sum\frac{y_i}{\sqrt{x_i}}$.

x	y	xy	x^2	$\sqrt{x_i}$	$\frac{1}{x_i}$	$\frac{y_i}{\sqrt{x_i}}$
0.2	16	3.2	0.04	0.4472	5	35.7782
0.3	14	4.2	0.09	0.5477	3.3333	25.5614
0.5	11	5.5	0.25	0.7071	2	15.55565
1	6	6	1	1	1	6
2	3	6	4	1.4142	0.5	2.1213
		$\sum xy =$ 24.9	$\sum x^2 =$ 5.2536	$\sum\sqrt{x_i} =$ 4.1162	$\sum\frac{1}{x_i} =$ 11.8333	$\sum\frac{y_i}{\sqrt{x_i}} =$ 85.0174

Normal equations, $24.9 = 5.2536a + 4.1162b$

$85.0174 = 4.1162a + 11.8333b$

$\Rightarrow a = -1.1836,\ b = 7.5961$

$\therefore$ Least square fit, $y = ax + \frac{b}{\sqrt{x}}$

$\Rightarrow y = -1.1836x + \frac{7.5961}{\sqrt{x}}$

Least square Error

$$E = \sum_{i=1}^{5} e_i^2$$

$$= \sum_{i=1}^{5}\left[y_i - \left(ax_i + \frac{b}{\sqrt{x_i}}\right)\right]^2$$

$$= \sum_{i=1}^{5}\left[y_i - \left(-1.1836x_i + \frac{7.5596}{\sqrt{x_i}}\right)\right]^2$$

$= 1.6887$

Q Obtain least square fit of the form $y = ab^x$. to the following data (4)

x	1	2	3	4	5	6	7	8
y	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

Sol. $y = ab^x$

$\Rightarrow \log y = \log a + x \log b \quad -(1)$

Let ~~log~~ $Y = \log y$, $A = \log a$, $B = \log b$

$\therefore$ (1) becomes, $Y = A + Bx$.

$\therefore e_i = Y_i - (A + Bx_i)$

Min $E = \sum_{i=1}^{8} e_i^2 = \sum_{i=1}^{8} [Y_i - (A+Bx_i)]^2$

For Min, $\frac{\partial E}{\partial A} = 0$, $\frac{\partial E}{\partial B} = 0$

$$\Rightarrow \left.\begin{aligned} \Sigma Y_i &= nA + B\,\Sigma x_i \\ \Sigma x_i Y_i &= A\,\Sigma x_i + B\,\Sigma x_i^2 \end{aligned}\right\}$$ Normal equation.

x	y	$Y = \log y$	xY	x^2
1	1.0	0	0	1
2	1.2	0.0792	0.1584	4
3	1.8	0.2553	0.7659	9
4	2.5	0.3979	1.5916	16
5	3.6	0.5563	2.7815	25
6	4.7	0.6721	4.0326	36
7	6.6	0.8195	5.7365	49
8	9.1	0.9590	7.6720	64
$\Sigma x = 36$	$\Sigma y = 30.5$	$\Sigma Y_i = 3.7393$	$\Sigma xY = 22.7385$	$\Sigma x^2 = 204$

Normal equations

$3.7393 = 8A + 36B$

$22.7385 = 36A + 204B$

$\Rightarrow A = 0.1656,\ B = 0.1407$

$\Rightarrow \log a = 0.1656,\ \log b = 0.1407$

$\boxed{a = 1.464}$ $\boxed{b = 1.3826}$

~~134.6148 =~~

$y = (1.464)(1.3826)^x$

~~14.9572 = 36A~~

$134.6148 = 288A + 1296B$

$181.908 = 288A + 1632B$

$-47.2932 = -336B$

$A = \frac{3.7393 - 36(0.14075)}{8}$

$= 0.1659$

$\boxed{B = 0.14075}$

Example 14. *We are given the values of a function of the variable t. Obtain a least square fit of the form $f = ae^{-3t} + be^{-2t}$.*

t	0.1	0.2	0.3	0.4
$f(t)$	0.76	0.58	0.44	0.35

Sol. Using the method of least square, we minimize the error

$$E = \sum_{i=1}^{4}[f_i - (ae^{-3t_i} + be^{-2t_i})]^2$$

and obtain the normal equations

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{4}(f_i - ae^{-3t_i} - be^{-2t_i})e^{-3t_i} = 0$$

$$\frac{\partial E}{\partial b} = \sum_{i=1}^{4}(f_i - ae^{-3t_i} - be^{-2t_i})e^{-2t_i} = 0$$

$$a\sum_{i=1}^{4}e^{-6t_i} + \sum_{i=1}^{4}e^{-5t_i} - \sum_{i=1}^{4}f_ie^{-2t_i} = 0$$

$$a\sum_{i=1}^{4}e^{-5t_i} + b\sum_{i=1}^{4}e^{-4t_i} - \sum_{i=1}^{4}f_ie^{-2t_i} = 0$$

Using the table values, we obtain the system of equations

$$1.106023a + 1.332876b - 1.16542 = 0$$

$$1.332876a + 1.7622740b - 1.409764 = 0,$$

which have the solution $a = 0.6853$, $b = 0.3058$.
Therefore the least square fit is given by

$$f(t) = 0.6853e^{-3t} + 0.3058e^{-2t}.$$

Remark 5.1. *If data is quite large then we can make it small by changing the origin and appropriating scaling.*

Example 15. *Show that the line of fit to the following data is given by* $y = 0.7x + 11.28$.

x	0	5	10	15	20	25
y	12	15	17	22	24	30

Sol. Here $n = 6$. We fit a line of the form $y = A + Bx$.
Let $u = \dfrac{x - 15}{5}$, $v = y - 20$ and line of the form $v = a + bu$.

x	y	u	v	uv	u^2
0	12	−3	−8	24	9
5	15	−2	−5	10	4
10	17	−1	−3	3	1
15	22	0	2	0	0
20	24	1	4	4	1
25	30	2	10	20	4
Σ		−3	0	61	19

The normal equations are,

$$0 = 6a - 3b$$

$$61 = -3a + 19b.$$

By solving $a = 1.7428$ and $b = 3.4857$.
Therefore equation of the line is $v = 1.7428 + 3.4857u$.
Changing in to original variable, we obtain

$$y - 20 = 1.7428 + 3.4857\left(\frac{x - 15}{5}\right)$$

$$\implies y = 11.2857 + 0.6971x.$$