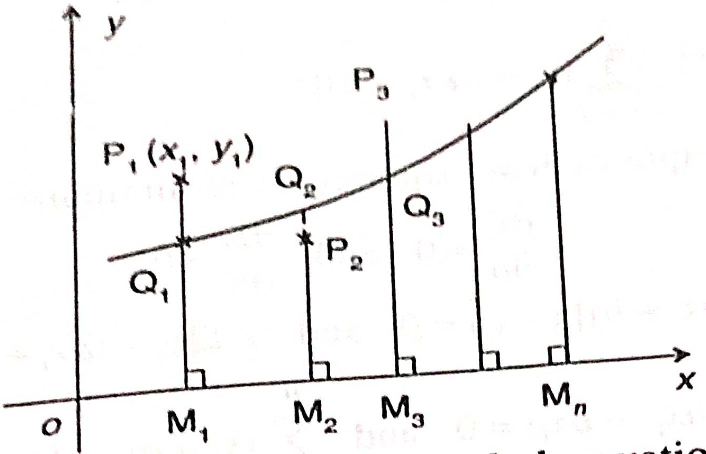
x	3	17	53	147
y				

1.6 The principle of least squares

In the previous articles, we have seen two methods of fitting curve, *viz.*, (*i*) the graphical method and (*ii*) the method of group averages. The first one is a rough method and in the second method, the evaluation of constants vary from one grouping to another grouping of data. So, we adopt another method, called *the method of least squares* which gives a unique set of values to the constants in the equation of the fitting curve.



Let (x_i, y_i), $i = 1, 2, ..., n$ be the n sets of observations and let

$$y = f(x) \qquad ...(1)$$

be the relation suggested between x and y.

Let (x_i, y_i) be represented by the point P_i. Let the ordinate at P_i meet $y = f(x)$ at Q_i and the x-axis at M_i.

$$M_iQ_i = f(x_i) \quad \text{and} \quad M_iP_i = y_i$$

$$Q_iP_i = M_iP_i - M_iQ_i = y_i - f(x_i), \quad i = 1, 2, ..., n$$

$d_i = y_i - f(x_i)$ is called the residual at $x = x_i$. Some of the d_i's may be positive and some may be negative.

$E = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2$ is the sum of the squares of the residuals.

If $E = 0$, *i.e.*, each $d_i = 0$, then all the n points P_i will lie on $y = f(x)$.

If not, we will choose $f(x)$ such that E is minimum. That is, the best fitting curve to the set of points is that for which E is minimum. This principle is known as the *principle of least squares* or the *least square criterion*.

This principle *does not suggest* to determine the *form of the curve* $y = f(x)$ but it *determines the values of the parameters* or constants of the equation of the curve.

We will consider some of the best fitting curves of the type :

(*i*) a straight line (*ii*) a second degree curve (*ii*) the exponential curve $y = ae^{bx}$ (*iv*) the curve $y = ax^n$.

1.7. Fitting a straight line by the method of least squares

Let (x_i, y_i), $i = 1, 2, ..., n$ be the n sets of observations and let the related relation by $y = ax + b$. Now we have to select a and b so that the straight line is the best fit to the data.

As explained earlier, the residual at $x = x_i$ is

$$d_i = y_i - f(x_i) = y_i - (ax_i + b),\ i = 1, 2, ..., n$$

$$E = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [y_i - (a\,x_i + b)]^2$$

By the principle of least squares, E is minimum.

$$\therefore \qquad \frac{\partial E}{\partial a} = 0 \text{ and } \frac{\partial E}{\partial b} = 0$$

i.e., $$2\Sigma[y_i - (ax_i + b)]\,(-x_i) = 0 \text{ and } 2\,\Sigma[y_i - (ax_i + b)](-1) = 0$$

i.e., $$\sum_{i=1}^{n} (x_i y_i - ax_i^2 - bx_i) = 0 \text{ and } \sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

i.e., $$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \qquad ...(1)$$

and $$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i \qquad ...(2)$$

Since, x_i, y_i are known, equations (1) and (2) give two equations in a and b. Solve for a and b from (1) and (2) and obtain the best fit $y = ax + b$.

Note 1. Equations (1) and (2) are called *normal equations*.

2. Dropping suffix i from (1) and (2), the normal equations are

$$a\Sigma x + nb = \Sigma y \text{ and } a\Sigma x^2 + b\Sigma x = \Sigma xy$$

which are got by taking Σ on both sides of $y = ax + b$ and also taking Σ on both sides after multiplying by x both sides of $y = ax + b$.

3. Transformations like $X = \dfrac{x-a}{h}$, $Y = \dfrac{y-b}{k}$ reduce the linear equation $y = \alpha x + \beta$ to the form $Y = AX + B$. Hence, a linear fit is another linear fit in both systems of coordinates.

Example 14. *By the method of least squares find the best fitting straight line to the data given below :*

x :	5	10	15	20	25
y :	15 (16	19	23	26	30

Solution. Let the straight line be $y = ax + b$

The normal equations are $a\Sigma x + 5b = \Sigma y$ ...(1)

$$a\,\Sigma x^2 + b\Sigma x = \Sigma xy \quad ...(2)$$

To calculate Σx, Σx^2, Σy, Σxy we form below the table.

	x	y	x^2	xy
	5	16	25	80
	10	19	100	190
	15	23	225	345
	20	26	400	520
	25	30	625	750
Total	75	114	1375	1885

The normal equations are $75a + 5b = 114$ ...(1)

$$1375a + 75b = 1885 \quad ...(2)$$

Eliminate b; multiply (1) by 15

$$1125a + 75b = 1710 \quad ...(3)$$

(2) – (3) gives, $250a = 175$ or $a = 0{\cdot}7$

Hence $b = 12{\cdot}3$

Hence, the best fitting line is $\mathbf{y = 0{\cdot}7x + 12{\cdot}3}$

Aliter. Let $X = \dfrac{x - 15}{5}$, $Y = y - 23$

Let the line in the new variable be $Y = AX + B$ ...(1)

	x	y	X	X^2	Y	XY
	5	16	– 2	4	– 7	14
	10	19	– 1	1	– 4	4
	15	23	0	0	0	0
	20	26	1	1	3	3
	25	30	2	4	7	14
Total : Σ			0	10	– 1	35

The normal equations are $A\Sigma X + 5B = \Sigma Y$ ...(4)

$$A\Sigma X^2 + B\Sigma X = \Sigma XY \quad ...(5)$$

Therefore, $5B = -1$ $\therefore$ $B = -0{\cdot}2$

$\therefore$ $10A = 35$ $A = 3{\cdot}5$

The equations $Y = 3{\cdot}5X - 0{\cdot}2$

i.e., $$y - 23 = 3 \cdot 5\left(\frac{x - 15}{5}\right) - 0 \cdot 2 = 0 \cdot 7x - 10 \cdot 5 - 0 \cdot 2$$

i.e., $$Y = 0 \cdot 7x + 12 \cdot 3$$

which is the same equation as seen before.

Example 15. *Fit a straight line to the data given below. Also estimate the value of y at $x = 2 \cdot 5$.*

x :	*0*	*1*	*2*	*3*	*4*
y :	*1*	*1·8*	*3·3*	*4·5*	*6·3*

Solution. Let the best fit be $y = ax + b$ ...(1)

The normal equations are

$$a\Sigma x + 5b = \Sigma y \quad ...(2)$$

$$a\Sigma x^2 + b\Sigma x = \Sigma xy \quad ...(3)$$

We prepare the table for easy use.

	x	y	x^2	xy
	0	1·0	0	0
	1	1·8	1	1·8
	2	3·3	4	6·6
	3	4·5	9	13·5
	4	6·3	16	25·2
Total	10	16·9	30	47·1

Substituting in (2) and (3), we get,

$$10a + 5b = 16 \cdot 9$$

$$30a + 10b = 47 \cdot 1$$

Solving, we get, $a = 1 \cdot 33$, $b = 0 \cdot 72$

Hence, the equation is $y = 1 \cdot 33x + 0 \cdot 72$

$$y\ (at\ x = 2 \cdot 5) = 1 \cdot 33 \times 2 \cdot 5 + 0 \cdot 72 = 4 \cdot 045$$

Example 16. *Fit a straight line to the following data. Also estimate the value of y at $x = 70$.*

x :	*71*	*68*	*73*	*69*	*67*	*65*	*66*	*67*
y :	*69*	*72*	*70*	*70*	*68*	*67*	*68*	*64*

Solution. Since the values of x and y are larger, we choose the origins for x and y at 69 and 67 respectively. In other words, we transform x and y.

Let $X = x - 69$, and $Y = y - 67$

Let $Y = aX + b$ be the best fit. ...(1)

The normal equations are

$$a\Sigma X + 8b = \Sigma Y \quad ...(2)$$

$$a\Sigma X^2 + b\Sigma X = \Sigma XY \quad ...(3)$$

Calculations :

x	y	X	Y	X^2	XY
71	69	2	2	4	4
68	72	− 1	5	1	− 5
73	70	4	3	16	12
69	70	0	3	0	0
67	68	− 2	1	4	− 2
65	67	− 4	0	16	0
66	68	− 3	1	9	− 3
67	64	− 2	− 3	4	6
Total:		− 6	12	54	12

Substituting in (2) and (3),

$$-6a + 8b = 12 \qquad \text{...(4)}$$

$$54a - 6b = 12$$

Solving, we get, $a = 0{\cdot}424242$, $b = 0{\cdot}181818$

Therefore, $\quad Y = 0{\cdot}4242X + 0{\cdot}1818$

i.e., $\quad y - 67 = 0{\cdot}4242\,(x - 69) + 0{\cdot}1818$

$$y = 0{\cdot}4242x + 37{\cdot}909$$

$$y\,(x = 70) = 0{\cdot}4242 \times 70 + 37{\cdot}909 = \mathbf{67{\cdot}6030}$$

Example 17. *By proper transformation, convert the relation $y = a + bxy$ to a linear form and find the equation to fit the data.*

x :	− 4	1	2	3
y :	4	6	10	8

[*MS. BE 1972*]

Solution. Let $X = xy$ $\therefore$ The equation becomes $y = a + bx$.

The normal equations are

$$a\Sigma X + b\Sigma X^2 = \Sigma Xy \qquad \text{...(1)}$$

and

$$4a + b\Sigma X = \Sigma y \qquad \text{...(2)}$$

x	y	X	X^2	XY
− 4	4	− 16	256	− 64
1	6	6	36	36
2	10	20	400	200
3	8	24	576	192
Total:	28	34	1268	364

Using (1) and (2),

$$34a + 1268b = 364$$

$$4a + 34b = 28$$

Solving we get, $b = 0{\cdot}1286$, $a = 5{\cdot}9069$

Therefore, the equation is $y = 5{\cdot}9069 + 0{\cdot}1286\,X$

i.e., $\quad y = 5{\cdot}9069 + 0{\cdot}1286\,xy.$

Using this equation we get $y\,(1 - 0{\cdot}1286x) = 5{\cdot}9069$

i.e., $\quad y = \dfrac{5{\cdot}9069}{1 - 0{\cdot}1286x}$

We tabulate the values to verify:

x :	-4	1	2	3
y :	5·23	6·78	7·95	9·62

Note. If we take $u = \frac{1}{xy}$, $v = \frac{1}{x}$ we get

$v = au + b$. Taking this as linear, we get

$a = 10{\cdot}5,\ b = -0{\cdot}13$

That is $y = 10{\cdot}5 - 0{\cdot}13xy$

i.e., $$y = \frac{10{\cdot}5}{1 + 0{\cdot}13x}$$

Now tabulating, we get

x :	-4	1	2	3
y :	21·87	92·92	0·12	7·55

The values of y are far away from the given values. Perhaps, the selection of the form is not correct.

1.8. Fitting a parabola or fitting a second degree curve (by the method of least squares)

Let (x_i, y_i), $i = 1, 2, \ldots, n$ be n sets of observations of two related variables x and y. Let $y = ax^2 + bx + c$ be the equation which fits them best.

Now, we have to find the constants a, b, c.

For any $x = x_i$, the expected value of y is $ax_i^2 + bx_i + c$ and the corresponding observed value is y_i.

The residual $d_i = y_i - (ax_i^2 + bx_i + c)$

Let E denote the sum of the squares of the residuals.

That is, $$E = \sum_{i=1}^{n} [y_i - (ax_i^2 + bx_i + c)]^2$$

By the principle of least squares, E is minimum for best values a, b, c.

$$\therefore \quad \frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b} = 0 \text{ and } \frac{\partial E}{\partial c} = 0$$

$\therefore$ Differentiating E, partially w.r.t. a, b, c and equating to zero, we get,

$$\sum_{i=1}^{n} 2\,[y_i - (ax_i^2 + bx_i + c)]\,(-x_i^2) = 0$$

$$\sum_{i=1}^{n} 2\,[y_i - (ax_i^2 + bx_i + c)]\,(-x_i) = 0$$

$$\sum_{i=1}^{n} 2\,[y_i - (a\,x_i + bx_i + c)]\,(-1) = 0$$

Simplifying, we get

$$a\Sigma\, x_i^4 + b\Sigma x_i^3 + c\Sigma x_i^2 = \Sigma x_i^2 y_i$$

$$a\Sigma x_i^3 + b\Sigma x_i^2 + c\Sigma x_i = \Sigma x_i y_i$$

$$a\Sigma x_i^2 + b\Sigma x_i + nc = \Sigma y_i$$

Dropping the suffices, the *normal equations* are

$$a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2 = \Sigma\, x^2 y \quad ...(1)$$

$$a\Sigma x^3 + b\Sigma x^2 + c\Sigma x = \Sigma xy \quad ...(2)$$

$$a\Sigma x^2 + b\Sigma x + nc = \Sigma y \quad ...(3)$$

The three equations (1), (2), (3) give the values of a, b, c. Substituting these values of a, b, c in $y = ax^2 + bx + c$, we get the result.

Note. To obtain the normal equations, we remember the following :

(*i*) In $y = ax^2 + bx + c$, take Σ on both sides.

(*ii*) Multiply by x both sides and then take Σ on both sides.

(*iii*) Multiply both sides by x^2 and then take Σ on both sides.

Example 18. *Fit a parabola, by the method of least squares, to the following data; also estimate y at x = 6.*

x :	1	2	3	4	5
y :	5	12	26	60	97

Solution. Let $y = ax^2 + bx + c$ be the best fit.

Then, the normal equations are

$$a\Sigma x^2 + b\Sigma x + 5c = \Sigma y \quad ...(1)$$

$$a\Sigma x^3 + b\Sigma x^2 + c\Sigma x = \Sigma xy \quad ...(2)$$

$$a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2 = \Sigma x^2 y \quad ...(3)$$

We form the table.

	x	y	x^2	x^3	x^4	xy	x^2y
	1	5	1	1	1	5	5
	2	12	4	8	16	24	48
	3	26	9	27	81	78	234
	4	60	16	64	256	240	960
	5	97	25	125	625	485	2425
Total:	15	200	55	225	979	832	3672

Hence the equations (1), (2), (3) become,

$$55a + 15b + 5c = 200 \quad ...(4)$$

$$225a + 55b + 15c = 832 \quad ...(5)$$

$$979a + 225b + 55c = 3672 \quad ...(6)$$

Solving we get, $a = 5.7143$, $b = -11.0858$ and $c = 10.4001$

Hence, the parabola is, $y = 5.7143x^2 - 11.0858x + 10.4001$

$$y(x = 6) = 149.6001$$

Example 19. *Fit a second degree parabola to the data.*

x :	1929	1930	1931	1932	1933	1934	1935
y :	352	356	357	358	360	361	361

Let $\quad X = x - 1932, \quad Y = y - 357$

Let $\quad Y = aX^2 + bX + C$ be the best fit.

The normals equations are

$$a\Sigma x^2 + b\Sigma X + 7c = \Sigma Y \quad \ldots(1)$$

$$a\Sigma X^3 + b\Sigma X^2 + c\Sigma X = \Sigma XY \quad \ldots(2)$$

$$a\Sigma X^4 + b\Sigma X^3 + c\Sigma X^2 = \Sigma X^2 Y \quad \ldots(3)$$

Calculation Table :

	x	y	X	Y	X^2	X^3	X^4	XY	X^2Y
	1929	352	−3	−5	9	−27	81	15	−45
	1930	356	−2	−1	4	−8	16	2	−4
	1931	357	−1	0	1	−1	1	0	0
	1932	358	0	1	0	0	0	0	0
	1933	360	1	3	1	1	1	3	3
	1934	361	2	4	4	8	16	8	16
	1935	361	3	4	9	27	81	12	36
Total			0	6	28	0	196	40	6

Hence the normal equations become,

$$28a + 7c = 6 \quad \ldots(4)$$

$$28b = 40 \quad \ldots(5)$$

$$196a + 28c = 6 \quad \ldots(6)$$

$\therefore \quad b = 1.4286, \quad a = -0.21429, \quad c = 1.7143$

$\therefore \quad$ The equation is $Y = -0.21429\, X^2 + 1.4286\, X + 1.7143$

i.e., $\quad y - 357 = -0.21429\,(x - 1932)^2 + 1.7143 + 1.4286\,(x - 1932)$

i.e., $\quad y = -0.21429\, x^2 + 829.445\, x - 802265.33$

Example 20. *Fit a second degree parabola to the following data, taking y as dependent variable.*

x :	1	2	3	4	5	6	7	8	9
y :	2	6	7	8	10	11	11	10	9

Solution. Let $X = x - \bar{x} = x - 5$ and $Y = y - 7$

Let $Y = aX^2 + bX + c$ be the best fit.

x	y	X	Y	X^2	X^3	X^4	XY	X^2Y
1	2	−4	−5	16	−64	256	20	−80
2	6	−3	−1	9	−27	81	3	−9
3	7	−2	0	4	−8	16	0	0

	4	8	−1	1	1	−1	1	−1	1
	5	10	0	3	0	0	0	0	0
	6	11	1	4	1	1	1	4	4
	7	11	2	4	4	8	16	8	16
	8	10	3	3	9	27	81	9	27
	9	9	4	2	16	64	256	8	32
Total			0	11	60	0	708	51	−9

The normal equations are

$$708a + 60c = -9$$

$$60b = 51$$

$$60a + 9c = 11$$

Solving we get, $b = 0{\cdot}85$, $a = -0{\cdot}2673$, $c = 3{\cdot}0042$

Hence the equation is

$$Y = -0{\cdot}2673\,X^2 + 0{\cdot}85\,X + 3{\cdot}0042$$

$$\therefore \quad y - 7 = -0{\cdot}2673\,(x-5)^2 + 0{\cdot}85\,(x-5) + 3{\cdot}0042$$

i.e., $$\mathbf{y = -0{\cdot}2673\,x^2 + 3{\cdot}523\,x - 0{\cdot}9283}$$

1.9. Fitting an exponential curve

Let (x_i, y_i), $i = 1, 2, \ldots, n$ be the n sets of observations of related data and let $y = ab^x$ be the best fit for the data. Then taking logarithm on both sides,

$$\log_{10} y = \log_{10} a + x \log_{10} b$$

i.e., $$Y = A + Bx \text{ where } Y = \log_{10} y,\ A = \log_{10} a,\ B = \log_{10} b$$

This being linear in x and Y, we can find A, B since x and $Y = \log_{10} y$ are known. From A, B, we can get a, b and hence $y = ab^x$ is found out.

1.10. Fitting a curve of the form $y = ax^b$

Since $y = ax^b$, $\log_{10} y = \log_{10} a + b \log_{10} x$

i.e., $$Y = A + bX \text{ where } Y = \log_{10} y,\ X = \log_{10} x \text{ and } A = \log_{10} a$$

Again using this linear fit, we find A, b.

Hence a, b are known. Thus $y = ab^x$ is found out.

Example 21. *From the table given below, find the best values of a and b in the law $y = ae^{bx}$ by the method of least squares.*

x :	0	5	8	12	20
y :	3·0	1·5	1·0	0·55	0·18

Solution. $y = ae^{bx}$

$$\therefore \quad \log_{10} y = \log_{10} a + bx \log_{10} e \qquad \ldots(1)$$

i.e., $$Y = A + Bx$$

The normal equations are ...(2)

$$B\Sigma x + 5A = \Sigma Y$$

$B\Sigma x^2 + A\Sigma x = \Sigma xY$

x	y	Y	x^2	xY
0	3.0	0.4771	0	0
5	1.5	0.1761	25	0.8805
8	1.0	0	64	0
12	0.55	−0.2596	144	−3.1152
20	0.18	−0.7447	400	−14.894
Total 45		−0.3511	633	−17.1287

Using equations (2) and (3),

$$5A + 45B = -0.3511$$

$$45A + 633B = -17.1287$$

Solving we get $A = 0.4815,\ B = -0.0613$

$$a = 10^A = 3.0304$$

$$b \log_{10} e = B = -0.0613$$

$$b = -0.0613 \times \log_e 10 = -0.1411$$

∴ Hence, the curve is $\mathbf{y = 3.0304\, e^{-0.1411x}}$

Example 22. *Fit a curve of the form $y = ab^x$ to the data*

x :	*1*	*2*	*3*	*4*	*5*	*6*
y :	*151*	*100*	*61*	*50*	*20*	*8*

Solution. $y = ab^x$

∴ $\log_{10} y = \log_{10} a + x \log_{10} b$

i.e., $Y = A + Bx$ ...(1)

The normal equations are

$$B\Sigma x + 6A = \Sigma Y \quad ...(2)$$

$$B\Sigma x^2 + A\Sigma x = \Sigma xY \quad ...(3)$$

	x	y	Y	x^2	xY
	1	151	2.1790	1	2.1790
	2	100	2.0	4	4.0
	3	61	1.7853	9	5.3559
	4	50	1.6990	16	6.7960
	5	20	1.3010	25	6.5050
	6	8	0.9031	36	5.4186
Total	21		9.8674	91	30.2545

Using (2) and (3), we get

$$6A + 21B = 9.8674 \quad ...(4)$$

$$21A + 91B = 30.2545 \quad ...(5)$$

Solving, $A = 2.5010,\ B = -0.2447$

Since $\log_{10} a = A,\ a = 10^A = 316.9568$

$$b = 10^B = 0.5692$$

∴ the equation is $\mathbf{y = 316.9568\,(0.5692)^x}$

Example 23. *It is known that the curve $y = ax^b$ fits in the data given below. Find the best values of a and b.*

x :	1	2	3	4	5	6
y :	1200	900	600	200	110	50

Solution. $y = ax^b$

Taking logarithm, $\log_{10} y = \log_{10} a + b \log_{10} x$

i.e.,
$$Y = A + bX \quad ...(1)$$

where $Y = \log_{10} y$, $X = \log_{10} x$, $A = \log_{10} a$

The normal equations are

$$b\Sigma X + 6A = \Sigma Y \quad ...(2)$$

$$b\Sigma X^2 + A\Sigma X = \Sigma XY \quad ...(3)$$

	x	y	X	Y	X^2	XY
	1	1200	0·0	3·0792	0·0	0·0
	2	900	0·3010	2·9542	0·0906	0·8892
	3	600	0·4771	2·7782	0·2276	1·3255
	4	200	0·6021	2·3010	0·3625	1·3854
	5	110	0·6990	2·0414	0·4886	1·4269
	6	50	0·7781	1·6990	0·6054	1·3220
Total			2·8573	14·8530	1·7747	6·3490

Using (2) and (3), we have,

$$6A + 2{\cdot}8573b = 14{\cdot}8530$$

$$2{\cdot}8573A + 1{\cdot}7747b = 6{\cdot}3490$$

Solving, $A = 3{\cdot}3086$, $b = -1{\cdot}7494$

$\therefore$ $a = 10^A = 2035$

Hence, the equation is $\mathbf{y = 2035\, x^{-1{\cdot}7494}}$

1.11. Calculation of the sum of the squares of the residuals in the case of straight line fit

In fitting a straight line, we have seen that the sum of the squares of the residuals E is given by

$$E = \Sigma[y - (ax + b)]^2$$
$$= \Sigma[y - (ax + b)]\,[y - (ax + b)]$$
$$= \Sigma\{y\,[y - (ax + b)] - ax\,[y - (ax + b)] - b\,[y - (ax + b)]\}$$
$$= \Sigma y\,[y - (ax + b)] - a\Sigma x\,[y - (ax + b)] - b\Sigma[y - (ax + b)]$$
$$= \Sigma y\,[y - (ax + b)]$$ since the last two sums vanish due to normal equations $\Sigma x\,[y - (ax + b)] = 0$; $\Sigma[y - (ax + b)] = 0$

$$\therefore \quad E = \Sigma y^2 - a\Sigma xy - b\Sigma y$$

when we fit a straight line by the method of least squares, the error committed (which is minimum) is given by

$$\boxed{E = \Sigma y^2 - a\Sigma xy - b\Sigma y}$$

1.12. Calculation of the sum of the squares of the residuals in the case of parabola fit

In fitting a parabola, we have seen that the sum of the squares of residuals E is given by

$$E = \Sigma[y - (ax^2 + bx + c)]^2$$
$$= \Sigma[y - (ax^2 + bx + c)]\,[y - (ax^2 + bx + c)]$$
$$= \Sigma y\,[y - (ax^2 + bx + c)] - a\Sigma x^2[y - (ax^2 + bx + c)]$$
$$- b\Sigma x\,[y - (ax^2 + bx + c)] - c\Sigma[y - (ax^2 + bx + c)]$$

$= \Sigma y\,[y - (ax^2 + bx + c)]$ since the last three summations vanish due to normal equations.

$$= \Sigma y^2 - a\Sigma x^2 y - b\Sigma xy - c\Sigma y$$

When we fit a parabola by the method of least squares, the error committed (which is minimum) is given by

$$\boxed{E = \Sigma y^2 - a\Sigma x^2 y - b\Sigma xy - c\Sigma y}$$

Note. To remember this formula, multiply $y - (ax^2 + bx + c)$ by y and take Σ.

Example 24. *Fit a straight line and a parabola to the following data and find out which one is most appropriate. Reason out for your conclusion.*

x :	*0*	*1*	*2*	*3*	*4*
y :	*1*	*1·8*	*1·3*	*2·5*	*6·3*

Solution. *Stage 1* : Let $y = ax + b$ be the best fit.

Then the normal equations are

$$a\Sigma x + 5b = \Sigma y \quad ...(1)$$

and

$$a\Sigma x^2 + b\Sigma x = \Sigma xy \quad ...(2)$$

	x	y	x^2	x^3	x^4	xy	x^2y	y^2
	0	1	0	0	0	0	0	1
	1	1·8	1	1	1	1·8	1·8	3·24
	2	1·3	4	8	16	2·6	5·2	1·69
	3	2·5	9	27	81	7·5	22·5	6·25
	4	6·3	16	64	256	25·2	100·8	39·69
Total	10	12·9	30	100	354	37·1	130·3	51·87

Using (1) and (2), $10a + 5b = 12{\cdot}9$

and $30a + 10b = 37{\cdot}1$

Solving, we get, $b = 0{\cdot}32$, $a = 1{\cdot}33$

The best straight line fit is $\mathbf{y = 1{\cdot}33x + 0{\cdot}32}$

Error in this case $= E_1 = \Sigma y^2 - a\Sigma xy - b\Sigma y$

$$= 51{\cdot}87 - 1{\cdot}33\,(37{\cdot}1) - 0{\cdot}32(12{\cdot}9)$$
$$= \mathbf{-1{\cdot}60}$$

Stage 2 : Let $y = Ax^2 + Bx + c$ be the parabola fit.

The normal equations are

$$A\Sigma x^2 + B\Sigma x + 5c = \Sigma y$$
$$A\Sigma x^3 + B\Sigma x^2 + c\Sigma x = \Sigma xy$$
$$A\Sigma x^4 + B\Sigma x^3 + c\Sigma x^2 = \Sigma x^2 y$$

i.e.,

$$30A + 10B + 5c = 12{\cdot}9 \quad \ldots(3)$$
$$100A + 30B + 10c = 37{\cdot}1 \quad \ldots(4)$$
$$354A + 100B + 30c = 130{\cdot}3 \quad \ldots(5)$$

Solving, $A = 0{\cdot}55$, $B = -1{\cdot}07$, $c = 1{\cdot}42$

The best fitting curve is $\mathbf{y = 0{\cdot}55x^2 - 1{\cdot}07x + 1{\cdot}42}$

Error in this case is $E_2 = \Sigma y^2 - A\Sigma x^2 y - B\Sigma xy - c\Sigma y$

$$= 51{\cdot}87 - (0{\cdot}55)(130{\cdot}3) + 1{\cdot}07\,(37{\cdot}1) - 1{\cdot}42\,(12{\cdot}9)$$
$$= \mathbf{1{\cdot}584}$$

The error in the case of straight line fit is $= |E_1| = 1{\cdot}60$

The error in the case of parabola fit is $|E_2| = 1{\cdot}584$

$$|E_2| = 1{\cdot}584 < 1{\cdot}60 = |E_1|$$

∴ the parabola fit is better.

But both errors are more or less equal since $1{\cdot}584 \approx 1{\cdot}6$

Hence, we may even say both are of same status.

Though total errors are same, there may be different deviations as the table below exhibits.

x		0	1	2	3	4
y by data given	:	1	1·8	1·3	2·5	6·3
y by st. line fit	:	0·32	1·65	2·98	4·31	5·64
y by parabola fit	:	1·42	0·90	1·48	3·16	5·94

EXERCISE 1.3

Use the *method of least squares* and do the following problems :

1. Fit a straight line to the following data. Hence find $y(x = 25)$.

x	:	0	5	10	15	20
y	:	7	11	16	20	26

2. Fit a straight line to the following data :

x	:	0·0	0·2	0·4	0·6	0·8	1·0
y	:	−1·85	−1·20	−0·55	0·15	0·80	1·35

3. The weights of a calf taken at weekly intervals are supplied below. Fit a straight line and calculate the average rate of growth per week.

Age x	:	1	2	3	4	5	6	7	8	9	10
Weight y	:	52·5	58·7	65·0	70·2	75·4	81·1	87·2	95·5	102·2	106·4

4. Fit a straight line to the data given below :

x :	1	2	3	4	5	6
y :	1200	900	600	200	110	50

5. Fit a straight line to the following data :

x :	6	8	10	12	14	16	18	20	22	24
y :	3·8	3·7	4·0	3·9	4·3	4·2	4·2	4·4	4·5	4·5

6. Fit a st. line to the data.

x :	−3	−1	1	4	5	7	10
y :	−2	−1	0	1·5	2	3	4·5

7. Find the best fit of straight line to the data.

t :	6	7	7	8	8	8	9	9	10
θ :	5	5	4	5	4	3	4	3	3

8. Obtain a st. line fit to the data below :

x :	1	2	3	4	5
y :	4	3	6	7	11

9. Find the best fitting straight line to the data.

(a)

x :	75	80	93	65	87	71	98	68	84	77
y :	82	78	86	72	91	80	95	72	89	74

[MKU]

(b)

x :	0	5	10	15	20	25	30
y :	10	14	19	25	31	36	39

[MS.BE]

(c)

x :	1	2	3	4	5
y :	16	19	23	26	30

[MS. BE 72]

(d)

x :	1	2	3	4	5	6
y :	14	27	41	56	68	75

[MS. M.Sc.]

10. Fit a curve of the form $y = ax + b$ to the data.

x :	1	2	3	4	5
y :	14	27	40	55	68

[MS. BE]

11. A rubber band stretched under a force F is found to increase in length l. The following observations were got.

F :	2	3	4	5	6	7	8	9	10	11	12
l :	10	17	21	28	36	43	51	60	70	83	93

Fit a straight line. [MS.BE]

12. Fit a straight line to the data.

x :	0·5	1·0	1·5	2·0	2·5	3·0
y :	0·31	0·82	1·29	1·85	2·51	3·02

13. Fit a parabola to the data.

x :	0	0·1	0·2	0·3	04	0·5	0·6	0·7	0·8	0·9
y :	3·1950	3·2299	3·2532	3·2611	3·2516	3·2282	3·1807	3·1266	3·0594	2·9759

14. Fit a parabola to the data.

x :	1	2	3	4	5
y :	2	3	5	8	10

15. Fit a st. line to the data :

x :	3	4	5	6	7
y :	6	9	10	11	12

16. The variables x and y are related by $y = Ax^n$. Find the curve from the following data :

x :	1	2	3	4	5
y :	7·1	27·8	62·1	110	161

17. Fit a parabola to the data given below :

x :	1	2	3	4	5
y :	90	220	390	625	915

18. The following table gives the levels of prices in certain years. Fit a second degree parabola to the data.

Year :	1875	76	77	78	79	80	81	82	83	84	85
Price :	88	87	81	78	74	79	85	84	90	92	100

[*MS. BE 1970*]

19. Fit a second degree curve to the data.

x :	1929	1930	1931	1932	1933	1934	1935	1936	1937
y :	352	356	357	358	360	361	361	360	359

20. Fit a curve of the form $y = ax + bx^2$ to the data.

x :	1	2	3	4	5
y :	1·8	5·1	8·9	14·1	19·8

21. Fit a parabola to the data given below :

x :	1	2	3	4	5
y :	1250	1400	1650	1950	2300

22. Fit a parabola to the data.

x :	1·0	1·5	2·0	2·5	3·0	3·5	4·0
y :	1·1	1·3	1·6	2·6	2·7	3·4	4·1

23. Fit a curve of the form $y = ax^2 + b$ for the following data.

x :	1	2	3	4	5
y :	20·9	24·1	28·9	36·1	44·7

24. Fit a curve of the form $y = ax^2 + bx + c$ given the table.

x :	10	20	30	40	50	60
y :	157	179	210	252	302	361

25. Fit a law of the type $y = ae^{bx}$ to the data.

x :	0	1	2	3
y :	1·05	2·10	3·85	8·30

26. Fit a curve of the form $y = ae^{bx}$ to the data given below

x :	0	2	4
y :	5·012	10	31·62

[*BE. 1971*]

27. Fit a curve of the form $y = ae^{bx}$ to the data.

x :	1	2	3	4	5	6	7	8
y :	15·3	20·5	27·4	36·6	49·1	65·6	87·8	117·6

28. The following table gives the number of petals on a certain species of flower. Fit an exponential curve to the data.

No. of petals	x :	5	6	7	8	9	10
No. of flowers having a specified no. of petals	y :	133	55	23	7	2	2

29. Fit a curve of the form $y = ab^x$ to the data given below :

x :	1	2	3	4	5	6
y :	151	100	61	50	20	8

30. Fit the curve $pv^{\gamma} = k$ to the following data.

p :	0·5	1·0	1·5	2·0	2·5	3·0
v :	1·62	1·00	0·75	0·62	0·52	0·46

31. Fit a curve of the form $y = ab^x$ to the data.

x :	2	3	4	5	6
y :	144	172·8	207·4	248·8	298·5

32. Fit the curve of the form $y = ae^{bx}$ to the data given below:

x :	1	2	3	4
y :	1·65	2·70	4·50	7·35

33. An experiment on the life of a cutting tool at different cutting speeds are given below :

Speed v units :	350	400	500	600
Life T in min. :	61	26	7	2·6

Fit a relation of the form $v = aT^b$ [BE'67]

34. The horse power I, required to drive a ship of displacement D tons at a ten-knob speed is given by the table below. Find a formula of the form $I = aD^n$ to fit the data.

D :	1720	2300	3200	4100
I :	655	789	1000	1164

35. For the following data, fit a straight line and a parabola and show that the parabola fits the data significantly better than the straight line

x :	0·00	0·25	0·50	0·75	1·00
y :	0·00	0·06	0·20	0·60	0·90

36. Given the following data:

x :	0	1	2	3	4
y :	1	5	10	22	38

find the straight line and the parabola of best fit and calculate the sum of the squares of the residuals in both cases. Which curve is more appropriate and why ?

37. Fit the parabola and the straight line of best fit to the following data and out of them which is more reliable ?

x :	– 2	1	4	7	10
y :	10·8	1·9	29·2	92	190

38. It is known that x, y are related by $y = \frac{a}{x} + bx$ and the experimental values are given below.

x :	1	2	4	6	8
y :	5·43	6·28	10·32	14·86	19·5

Obtain the best values of a and b.

39. Growth of bacteria (y) in a culture after x hours is given below. Estimate the growth when $x = 7$ hours.

Hours	x :	0	1	2	3	4	5	6
Growth (Number)	y :	32	47	65	92	132	190	275

40. Fit the straight and the parabola of best fit to the following data and explain which is more preferable.

x :	0	1	2	3	4	5
y :	14	18	22	28	35	39

41. Fit the straight line and the parabola of best fit to the following data and explain which is more reliable.

x :	1	2	3	4	5
y :	10	12	8	10	14

42. Find the most plausible values of x, y, z from the equations.
$3x + 2y - 5z = 5$, $x - y + 2z = 3$, $4x + y + 4z = 21$ and $-x + 3y + 3z = 14$

[Hint : Let $S = (3x + 2y - 5z - 5)^2 + (x - y + 2z - 3)^2 + (4x + y + 4z - 21)^2 + (-x + 3y + 3z - 14)^2$

$\frac{\partial S}{\partial x} = 0$, $\frac{\partial S}{\partial y} = 0$, $\frac{\partial S}{\partial z} = 0$ give respectively $27x + 6y = 88$, $6x + 15y + z = 70$, $y + 54z = 107$

Solving $x = 2{\cdot}47$, $y = 3{\cdot}55$, $z = 1{\cdot}92$ are the most plausible values of x, y, z).

43. Fit a st. line to the following data showing the production of a commodity in different areas in Coimbatore.

Year	x :	1911	1912	1913	1914	1915
Production in (1000 tons)	y :	10	12	8	10	14

1.13. Method of moments

We will see below another method of curve fitting called *the method of moments.*

Let (x_i, y_i), $i = 1, 2, ..., n$ be n sets of observations of related data so that the x's are equally spaced. That is, $x_i - x_{i-1} = \Delta x = h$ (constant), $i = 2, 3, ... n$. For such a set of n points, we define,

the first moment $= \mu_1 = \Sigma y \Delta x = \Delta x \Sigma y$

the second moment $= \mu_2 = \Sigma xy \Delta x = \Delta x \Sigma xy$

the third moment $= \mu_3 = \Sigma x^2 y \Delta x = \Delta x \Sigma x^2 y$

and so on. These are known as the moments of the *observed* values of y

Let $y = f(x)$ be a curve which fits the data best.