3.16 CURVE FITTING

Let there be two variables x and y which give us a set of n pairs of numerical values (x_1, y_1), (x_2, y_2).....(x_n, y_n). In order to have an approximate idea about the relationship of these two variables, we plot these n paired points on a graph thus, we get a diagram showing the simultaneous variation in values of both the variables called *scatter or dot diagram*. From scatter diagram, we get only an approximate non-mathematical relation between two variables. Curve fitting means an exact relationship between two variables by algebraic equations, infact this relationship is the equation of the curve. Therefore, curve fitting means to form an equation of the curve from the given data. Curve fitting is considered of immense importance both from the point of view of theoretical and practical statistics.

Theoretically, it is useful in the study of correlation and regression. Practically, it enables us to represent the relationship between two variables by simple algebraic expressions *e.g.*, polynomials, exponential or logarithmic functions.

It is also used to estimate the values of one variable corresponding to the specified values of the other variable.

The constants occurring in the equation of approximate curve can be found by following methods:

(*i*) Graphical method (*ii*) Method of group averages

(*iii*) Method of least squares (*iv*) Method of moments.

Out of the above four methods, we will only discuss and study here *method of least squares.*

3.17 METHOD OF LEAST SQUARES

[U.P.T.U. MCA (C.O.) 2008; U.P.T.U. (C.O.) 2008 ; U.P.T.U. 2008]

Method of least squares provides a unique set of values to the constants and hence suggests a curve of best fit to the given data.

Suppose we have m-paired observations (x_1, y_1), (x_2, y_2), ..., (x_m, y_m) of two variables x and y. It is required to fit a polynomial of degree n of the type

$$y = a + bx + cx^2 + ... + kx^n \qquad ...(1)$$

of these values. We have to determine the constants a, b, c, ..., k such that it represents the curve of best fit of that degree.

In case $m = n$, we get in general a unique set of values satisfying the given system of equations.

But if $m > n$ then, we get m equations by putting different values of x and y in equation (1) and we want to find only the values of n constants. Thus, there may be no such solution to satisfy all m equations.

Therefore, we try to find out those values of a, b, c, ..., k which satisfy all the equations as nearly as possible. We apply the method of least squares in such cases.

Putting $x_1, x_2, ..., x_m$ for x in (1), we get

$$\begin{aligned} y_1' &= a + bx_1 + cx_1^2 + ... + kx_1^n \\ y_2' &= a + bx_2 + cx_2^2 + ... + kx_2^n \\ &\vdots \qquad\qquad \vdots \\ y_m' &= a + bx_m + cx_m^2 + ... + kx_m^n \end{aligned}$$

where $y_1', y_2', ..., y_m'$ are the expected values of y for $x = x_1, x_2, ..., x_m$ respectively. The values $y_1, y_2, ..., y_m$ are called observed values of y corresponding to $x = x_1, x_2, ..., x_m$ respectively.

The expected values are different from observed values, the difference $y_r - y_r'$ for different values of r are called *residuals*.

Introduce a new quantity U such that

$$U = \Sigma(y_r - y_r')^2 = \Sigma(y_r - a - bx_r - cx_r^2 - \ldots - kx_r^n)^2$$

The constants $a, b, c, \ldots, k$ are chosen in such a way that the sum of the squares of residuals is minimum.

Now the condition for U to be maximum or minimum is $\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b} = \frac{\partial U}{\partial c} = \ldots = \frac{\partial U}{\partial k}$.

On simplifying these relations, we get

$$\Sigma y = ma + b\Sigma x + \ldots + k\Sigma x^n$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + \ldots + k\,\Sigma x^{n+1}$$
$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + \ldots + k\,\Sigma x^{n+2}$$
$$\vdots \qquad\qquad \vdots$$
$$\Sigma x^n y = a\Sigma x^n + b\Sigma x^{n+1} + \ldots + k\,\Sigma x^{2n}$$

These are known as *Normal equations* and can be solved as simultaneous equations to give the values of the constants $a, b, c, \ldots, k$. These equations are $(n + 1)$ in number.

If we calculate the second order partial derivatives and these values are put, they give a positive value of the function, so U is minimum.

This method does not help us to choose the degree of the curve to be fitted but helps us in finding the values of the constants when the form of the curve has already been chosen.

3.18 FITTING A STRAIGHT LINE

Let (x_i, y_i), $i = 1, 2, \ldots, n$ be n sets of observations of related data and

$$y = a + bx \qquad \ldots(1)$$

be the straight line to be fitted. The residual at $x = x_i$ is

$$E_i = y_i - f(x_i) = y_i - a - bx_i$$

Introduce a new quantity U such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

By the principle of Least squares, U is minimum

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

$$\therefore \qquad 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0 \qquad \text{or} \quad \boxed{\Sigma y = na + b\Sigma x} \qquad \ldots(2)$$

and

$$2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0 \qquad \text{or} \quad \boxed{\Sigma xy = a\Sigma x + b\Sigma x^2} \qquad \ldots(3)$$

Since x_i, y_i are known, equations (2) and (3) result two equations in a and b. Solving these, the best values for a and b can be known and hence equation (1).

Note. In case of change of origin,

if n is odd then, $\quad u = \dfrac{x - (\text{middle term})}{\text{interval } (h)}$

but if n is even then, $\quad u = \dfrac{x - (\text{mean of two middle terms})}{\frac{1}{2}\,(\text{interval})}$.

EXAMPLES

Example 1. *By the method of least squares, find the straight line that best fits the following data:*

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

(U.P.T.U. 2008)

Sol. Let the straight line of best fit be $y = a + bx$

Normal equations are

$$\Sigma y = ma + b\Sigma x \quad \ldots(1)$$

and

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

$$\ldots(3)$$

Here $m = 5$

Table is as below:

x	y	xy	x^2
1	14	14	1
2	27	54	4
3	40	120	9
4	55	220	16
5	68	340	25
$\Sigma x = 15$	$\Sigma y = 204$	$\Sigma xy = 748$	$\Sigma x^2 = 55$

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

$$204 = 5a + 15b$$
$$748 = 15a + 55b$$

Solving, we get $a = 0$, $b = 13.6$

Hence required straight line is $y = 13.6x$

Example 2. *Fit a straight line to the following data by least square method:*

x:	*0*	*1*	*2*	*3*	*4*
y:	*1*	*1.8*	*3.3*	*4.5*	*6.3.*

(U.K.T.U. 2011)

Sol. Let the straight line obtained from the given data be $y = a + bx$ then the normal equations are

$$\Sigma y = ma + b\,\Sigma x \quad \ldots(1)$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \ldots(2)$$

Here,

$$m = 5$$

x	y	xy	x^2
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
$\Sigma x = 10$	$\Sigma y = 16.9$	$\Sigma xy = 47.1$	$\Sigma x^2 = 30$

From (1) and (2), $16.9 = 5a + 10b$

and $47.1 = 10a + 30b$

Solving, we get $a = 0.72, b = 1.33$

$\therefore$ Required line is $y = 0.72 + 1.33\,x$.

Example 3. *Show that the best fitting linear function for the points* $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ *may be expressed in the form*

$$\begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \end{vmatrix} = 0 \qquad (i = 1, 2, \ldots\ldots, n)$$

Show that the line passes through the mean point $(\bar{x}, \bar{y})$.

Sol. Let the best fitting linear function be $y = a + bx$ ...(1)

Then the normal equations are

$$\Sigma y_i = na + b\Sigma x_i \qquad \ldots(2)$$

and $$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 \qquad \ldots(3)$$

Equations (1), (2), (3) may be rewritten as

$$bx - y + a = 0$$

$$b\Sigma x_i - \Sigma y_i + na = 0$$

and $$b\Sigma x_i^2 - \Sigma x_i y_i + a\Sigma x_i = 0$$

Eliminating a and b between these equations

$$\begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \end{vmatrix} = 0 \qquad \ldots(4)$$

which is the required best fitting linear function for the mean point $(\bar{x}, \bar{y})$,

$$\bar{x} = \frac{1}{n}\Sigma x_i, \quad \bar{y} = \frac{1}{n}\Sigma y_i .$$

Clearly, the line (4) passes through point $(\bar{x}, \bar{y})$ as two rows of determinant being equal make it zero.

ASSIGNMENT

1. Fit a straight line to the following data regarding x as the independent variable:

(*i*)

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

(*ii*)

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

(*iii*)

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

2. (*i*) Find the best values of a and b so that $y = a + bx$ fits the given data:

x	0	1	2	3	4
y	1.0	2.9	4.8	6.7	8.6

(*ii*) Fit a straight line of the form $y = a_0 + a_1x$ to the data: *[U.P.T.U. (C.O.) 2008]*

x	1	2	3	4	6	8
y	2.4	3.1	3.5	4.2	5.0	6.0

3. Fit a straight line approximate to the data:

x	1	2	3	4
y	3	7	13	21

4. A simply supported beam carries a concentrated load P(*lb*) at its mid-point. Corresponding to various values of P, the maximum deflection Y (*in*) is measured. The data are given below. Find a law of the type $Y - a + bP$

P	100	120	140	160	180	200
Y	0.45	0.55	0.60	0.70	0.80	0.85

5. What straight line best fits the following data in the least square sense?

x	1	2	3	4
y	0	1	1	2

[G.B.T.U. (MCA) 2010]

6. The weight of a calf taken at weekly intervals are given below. Fit a straight line using method of least squares and calculate the average rate of growth per week.

Age	1	2	3	4	5	6	7	8	9	10
Weight	52.5	58.7	65	70.2	75.4	81.1	87.2	95.5	102.2	108.4

7. Find the least square line for the data points

$(-1, 10)$, $(0, 9)$, $(1, 7)$, $(2, 5)$, $(3, 4)$, $(4, 3)$, $(5, 0)$ and $(6, -1)$.

8. Find the least square line $y = a + bx$ for the data:

x_i	-2	-1	0	1	2
y_i	1	2	3	3	4

9. If P is the pull required to lift a load W by means of a pulley block, find a linear law of the form $P = mW + c$ connecting P and W, using the data:

P	12	15	21	25
W	50	70	100	120

where P and W are taken in kg-wt.

(U.P.T.U. 2007)

10. (*i*) Using the method of least squares, fit a straight line to the following data:

x	1	2	3	4	5
y	2	4	6	8	10

(*ii*) Using the method of least squares, fit a straight line from the following data:

x	0	2	4	5	6
y	5.012	10	15	21	30

(U.P.T.U. 2009)

(*iii*) Find the least square line that fits the following data, assuming that x-values are free from error: *[U.P.T.U. MCA (SUM) 2008]*

x	1	2	3	4	5	6
y	5.04	8.12	10.64	13.18	16.20	20.04

Answers

1. (*i*) $y = 1361.97 - 243.42\,x$ (*ii*) $y = 39.5454 + 0.4242\,x$ (*iii*) $y = 11.285 + 0.7\,x$

2. (*i*) $y = 1 + 1.9x$ (*ii*) $y = 2.0253 + 0.502\,x$ **3.** $y = -4 + 6x$

4. $Y = 0.004P + 0.048$ **5.** $y = -0.5 + 0.6x$ **6.** $y = 45.74 + 6.16x$, 6.16

7. $y = -1.6071429x + 8.6428571$ **8.** $y = 2.6 + (0.7)\,x$ **9.** $P = 2.2759 + 0.1879\,W$

10. (*i*) $y = 2x$ (*ii*) $y = 3.07734 + 3.86031\,x$ (*iii*) $y = 2.0253 + 2.908\,x$.

3.19 FITTING OF AN EXPONENTIAL CURVE $y = ae^{bx}$

Taking logarithm on both sides, we get

$$\log_{10} y = \log_{10} a + bx \log_{10} e$$

i.e.,
$$\boxed{Y = A + BX} \qquad \ldots(1)$$

where $Y = \log_{10} y$, $A = \log_{10} a$, $B = b \log_{10} e$ and $X = x$

The normal equations for (1) are

$$\Sigma Y = nA + B\Sigma X \quad \text{and} \quad \Sigma XY = A\Sigma X + B\Sigma X^2$$

Solving these, we get A and B.

Then a = antilog A and $b = \dfrac{B}{\log_{10} e}$.

3.20 FITTING OF THE CURVE $y = ax^b$

Taking logarithm on both sides, we get

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

i.e.,
$$\boxed{Y = A + BX} \qquad \ldots(1)$$

where $Y = \log_{10} y$, $A = \log_{10} a$, $B = b$ and $X = \log_{10} x$.