24.4 PRINCIPLE OF LEAST SQUARES

The graphical method has the obvious drawback of being unable to give a unique curve of fit. *The principle of least squares, however, provides an elegant procedure for fitting a unique curve to a given data.*

Let the curve, $y = a + bx + cx^2 + \dots + kx^{m-1}$ ...(1)

be fitted to the set of n data points $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$.

Now we have to determine the constants $a, b, c, \dots k$ such that it represents the curve of best fit. In case $n = m$, on substituting the values (x_i, y_i) in (1), we get n equations from which a unique set of n constants can be found. But when $n > m$, we obtain n equations which are more than the m constants and hence cannot be solved for these constants. So we try to determine those values of $a, b, c, \dots k$ which satisfy all the equations as nearly as possible and thus may give the best fit. In such cases, we apply the *principle of least squares*.

At $x = x_i$, the *observed* (or *experimental*) *value* of the ordinate is $y_i = P_iL_i$ and the corresponding value on the fitting curve (1) is $a + bx_i + cx_i^2 + \dots + kx_i^m = M_iL_i$ $(= \eta_i$, say) which is the *expected* (or *calculated*) *value* (Fig. 24.4). The difference of the observed and the expected values *i.e.* $y_i - \eta_i (= e_i)$ is called the *error* (or *residual*) at $x = x_i$. Clearly some of the errors $e_1, e_2, \dots e_n$ will be positive and others negative. Thus to give equal weightage to each error, we square each of these and form their sum *i.e.* $E = e_1^2 + e_2^2 + \dots e_n^2$.

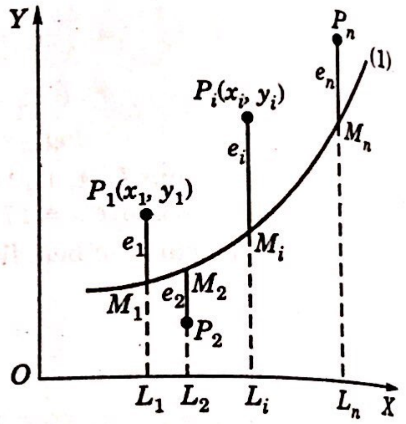

Fig. 24.4

The curve of best fit is that for which e's are as small as possible i.e., E, the sum of the squares of the errors is a minimum. This is known as the *principle of least squares* and was suggested by *Legendre** in 1806.

Obs. The principle of least squares does not help us to determine the form of the appropriate curve which can fit a given data. It only determines the best possible values of the constants in the equation when the form of the curve is known before hand. The selection of the curve is a matter of experience and practical considerations.

24.5 (1) METHOD OF LEAST SQUARES

For clarity, suppose it is required to fit the curve

$$y = a + bx + cx^2 \qquad \text{...(1)}$$

to a given set of observations $(x_1, y_1), (x_2, y_2), \dots, (x_5, y_5)$. For any x_i, the observed value is y_i and the expected value is $\eta_i = a + bx_i + cx_i^2$ so that the error $e_i = y_i - \eta_i$.

$\therefore$ the sum of the squares of these errors is

$$E = e_1^2 + e_2^2 + \dots + e_5^2$$
$$= [y_1 - (a + bx_1 + cx_1^2)]^2 + [y_2 - (a + bx_2 + cx_2^2)]^2 + \dots + [y_5 - (a + bx_5 + cx_5^2)]^2 \qquad \text{[See § 5.12 (3)]}$$

For E to be minimum, we have

$$\frac{\partial E}{\partial a} = 0 = 2[y_1 - (a + bx_1 + cx_1^2)] - 2[y_2 - (a + bx_2 + cx_2^2)] - \dots - 2[y_5 - (a + bx_5 + cx_5^2)] \qquad \text{...(2)}$$

$$\frac{\partial E}{\partial b} = 0 = -2x_1[y_1 - (a + bx_1 + cx_1^2)] - 2x_2[y_2 - (a + bx_2 + cx_2^2)] - \dots - 2x_5[y_5 - (a + bx_5 + cx_5^2)] \qquad \text{...(3)}$$

$$\frac{\partial E}{\partial c} = 0 = -2x_1^2[y_1 - (a + bx_1 + cx_1^2)] - 2x_2^2[y_2 - (a + bx_2 + cx_2^2)] - \dots - 2x_5^2[y_5 - (a + bx_5 + cx_5^2)] \qquad \text{...(4)}$$

Equation (2) simplifies to

$$y_1 + y_2 + \dots + y_5 = 5a + b(x_1 + x_2 + \dots + x_5) + c(x_1^2 + x_2^2 + \dots + x_5^2)$$

i.e.,
$$\Sigma y_i = 5a + b\Sigma x_i + c\,\Sigma x_i^2 \qquad \text{...(5)}$$

* See footnote on p. 311.

Equation (3) becomes

$$x_1y_1 + x_2y_2 + \dots + x_5y_5 = a(x_1 + x_2 + \dots + x_5) + b(x_1^2 + x_2^2 + \dots + x_5^2) + c(x_1^3 + x_2^3 + \dots + x_5^3) \quad \dots(6)$$

i.e., $$\Sigma x_iy_i = a\Sigma x_i + b\Sigma x_i^2 + c\Sigma x_i^3 \quad \dots(7)$$

Similarly (4) simplifies to $\Sigma x_i^2y_i = a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4$

The equations (5), (6) and (7) are known as *Normal equations* and can be solved as simultaneous equations in a, b, c. The values of these constants when substituted in (1) give the desired curve of best fit.

(2) Working procedure

(a) To fit the straight line $y = a + bx$

(i) Substitute the observed set of n values in this equation.

(ii) Form normal equations for each constant

i.e., $$\Sigma y = na + b\Sigma x,\ \Sigma xy = a\Sigma x + b\Sigma x^2$$

[The normal equation for the unknown a is obtained by multiplying the equations by the coefficient of a and adding. The normal equation for b is obtained by multiplying the equations by the coefficient of b ($i.e.$, x) and adding.]

(iii) Solve these normal equations as simultaneous equations for a and b.

(iv) Substitute the values of a and b in $y = a + bx$, which is the required line of best fit.

(b) To fit the parabola : $y = a + bx + cx^2$

(i) Form the normal equations $\Sigma y = na + b\Sigma x + c\Sigma x^2$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

and $$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

[The normal equation for c has been obtained by multiplying the equations by the coefficient of c ($i.e.$, x^2) and adding.]

(ii) Solve these as simultaneous equations for a, b, c.

(iii) Substitute the values of a, b, c in $y = a + bx + cx^2$, which is the required parabola of best fit.

(c) In general, the curve $y = a + bx + cx^2 + \dots + kx^{m-1}$ can be fitted to a given data by writing m normal equations.

Example 24.4. *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 following data :*

P = 12	*15*	*21*	*25*
W = 50	*70*	*100*	*120*

where P and W are taken in kg-wt. Compute P when W = 150 kg. wt. (U.P.T.U., 2007 ; V.T.U., 2002)

Solution. The corresponding normal equations are

$$\left.\begin{aligned}\Sigma P &= 4c + m\Sigma W\\ \Sigma WP &= c\Sigma W + m\Sigma W^2\end{aligned}\right\} \quad \dots(i)$$

The values of ΣW etc. are calculated by means of the following table :

W	P	W²	WP
50	12	2500	600
70	15	4900	1050
100	21	10000	2100
120	25	14400	3000
Total = 340	73	31800	6750

∴ The equations (i) becomes $73 = 4c + 340m$ and $6750 = 340c + 31800m$

i.e., $$2c + 170m = 365 \quad \dots(ii)$$

and $$34c + 3180m = 675 \quad \dots(iii)$$

Multiplying (ii) by 17 and substracting from (iii), we get

$m = 0.1879$ ∴ from (ii), $c = 2.2785$

Hence the line of best fit is

$$P = 2.2759 + 0.1879\ W$$

When $W = 150$ kg., $P = 2.2785 + 0.1879 \times 150 = 30.4635$ kg.

Obs. The calculations get simplified when the central values of x is zero. It is therefo