Least Squares Approximation of Functions

LINEAR REGRESSION:

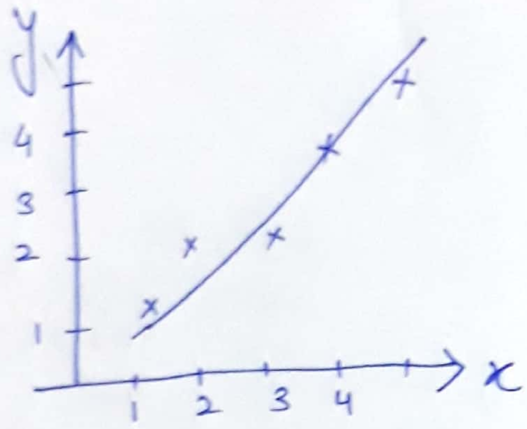

Figure represents a graph of given tabulated values. We can see that a straight line describes this situation very well.

So, we may ~~wright~~ write,

$$y_1 = a_1 x + a_0 \quad \text{——— (1)}$$

We have to find the values of a_0 and a_1 which lead to the 'best' straight line.

For this we take the deviations between the tabulated values of y and those obtained from the straight line:

$$\sum_{i=1}^{n} (y_i - \bar{y}_i)^2 = \sum_{i=1}^{n} [y_i - (a_1 x_i + a_0)]^2 \quad \text{——— (2)}$$

In obtaining eqn $y = a_1 x + a_0$, x which is the independent or controlled variable is assumed to have no errors in it.

y – dependent or measured variable is assumed to have statistical errors.

The deviation $(y_i - \bar{y}_i)$ measures the error between the measured and fitted value along the y direction. The linear-least squares fit equation obtained in this case is called as <u>linear regression of y on x</u>.

To calculate a_0 and a_1 in $y = a_1 x + a_0$, we minimise the sum of squares of error between measured values and those obtained by eqn (1).

Thus, we want to minimise S given by

$$\min S = \min \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \quad \text{——— (3)}$$

$$= \min \sum_{i=1}^{n} [y_i - (a_1 x_i + a_0)]^2 \text{— (4)}$$

S is a function of two unknown variables a_0 and a_1.

∴ To minimise S, we take the partial derivative of S wrt a_0 and a_1 and set these to zero.

$$\frac{\partial S}{\partial a_0} = \sum_{i=1}^{n} 2[y_i - a_1 x_i - a_0](-1) = 0 \text{ — (5)}$$

$$\Rightarrow \sum y_i - a_1 \sum x_i - n a_0 = 0 \text{ ——— (6)}$$

and $$\frac{\partial S}{\partial a_1} = \sum_{i=1}^{n} 2(y_i - a_1 x_i - a_0)(-x_i) = 0 \text{ —— (7)}$$

$$\Rightarrow -\sum x_i y_i + a_1 \sum x_i^2 + a_0 \sum x_i = 0 \text{ —— (8)}$$

From eqn (6) and (8),

$$n a_0 + (\Sigma x_i)\, a_1 = \Sigma y_i \qquad (9)$$

$$(\Sigma x_i) a_0 + (\Sigma x_i^2)\, a_1 = \Sigma x_i y_i \qquad (10)$$

solving eqns (9) and (10) for a_0 and a_1,

$$a_0 = \frac{\Sigma y_i \Sigma x_i^2 - \Sigma x_i \Sigma x_i y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2} \qquad (11)$$

$$a_1 = \frac{n \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2} \qquad (12)$$

These two equations are called normal equations and give the coefficients of the linear least squares fit line or regression coefficients.

Question 1 :

Fit a straight line for the given table of values :

x	1	2	4	5	6	8	9
y	2	5	7	10	12	15	19

Ans. $a_0 = 0.96$, $a_1 = 1.98$

$$y = 1.98x + 0.96$$

Algorithm for Linear Regression:

1. Read n
2. $sumx \leftarrow 0$
3. $sum\ xsq \leftarrow 0$
4. $sum\ y \leftarrow 0$
5. $sum\ xy \leftarrow 0$
6. for $i = 1$ to n do
7. Read x, y
8. $sumx \leftarrow sumx + x$
9. $sum\ xsq \leftarrow sum\ xsq + x^2$
10. $sum\ y \leftarrow sum\ y + y$
11. $sum\ xy \leftarrow sum\ xy + x \times y$
12. endfor
13. $denom \leftarrow n \times sum\ xsq - sumx \times sumx$
14. $a_0 \leftarrow (sum\ y \times sum\ xsq - sumx \times sumx) / denom$
15. $a_1 \leftarrow (n \times sumxy - sumx \times sum\ y) / denom$
16. Write a_1, a_0
17. Stop

Here n - gives the total pairs of values of x and y.

POLYNOMIAL REGRESSION

It is necessary (in general) to fit a higher degree polynomial rather than a straight line. Following is a method to fit a quadratic polynomial using n pairs of coordinates (x_i, y_i) given.

Let the quadratic be represented by,

$$y = a_2 x^2 + a_1 x + a_0 \text{ ——— } (1)$$

When $x = x_i$, $y = \bar{y_i}$ then

$$\bar{y_i} = a_2 x_i^2 + a_1 x_i + a_0 \text{ —— } (2)$$

The sum of squares of the deviations is given by,

$$S = \Sigma (y_i - \bar{y_i})^2 = \Sigma (y_i - a_2 x_i^2 - a_1 x_i - a_0)^2 \text{ —— } (3)$$

Differentiating S w.r.t a_0, a_1 and a_2 and setting each of these coefficients equal to zero, we get

$$\frac{\partial S}{\partial a_0} = \Sigma 2 (y_i - a_2 x_i^2 - a_1 x_i - a_0)(-1) = 0$$

$$\Rightarrow \Sigma (y_i - a_2 x_i^2 - a_1 x_i - a_0) = 0$$

$$\Rightarrow \Sigma y_i = a_2 \Sigma x_i^2 + a_1 \Sigma x_i + n a_0 \text{ —— } (4)$$

$$\frac{\partial S}{\partial a_1} = \Sigma 2 (y_i - a_2 x_i^2 - a_1 x_i - a_0)(-x_i) = 0$$

$$\Rightarrow \Sigma x_i y_i = a_2 \Sigma x_i^3 + a_1 \Sigma x_i^2 + a_0 \Sigma x_i \text{ —— } (5)$$

and $$\frac{\partial S}{\partial a_2} = \Sigma 2 (y_i - a_2 x_i^2 - a_1 x_i - a_0)(-x_i^2) = 0$$

$$\Rightarrow \Sigma x_i^2 y_i = a_2 \Sigma x_i^4 + a_1 \Sigma x_i^3 + a_0 \Sigma x_i^2 \text{ —— } (6)$$

These are three linear equations in three unknowns. These are called <u>normal equations</u> for quadratic regression, which may be solved by Gauss elimination procedure.

Using same idea we can fit any n^{th} degree polynomial and then there will be $(n+1)$ simultaneous equations in $(n+1)$ unknowns.

<u>Ex. 2</u> Values for y for various specified values of x are given below. Fit a quadratic curve through the points.

x_i	-4	-3	-2	-1	0	1	2	3	4	5
y_i	21	12	4	1	2	7	15	30	45	67

Ans. $a_0 = 2.07$

$a_1 = 3.0$

$a_2 = 1.98$

The eqn is: $y = 1.98x^2 + 3x + 2.07$

Fitting Exponential and Trigonometric Functions

By inspecting the given data points graphically, we assume an expression for the curve with some unknown coefficients and then find them.

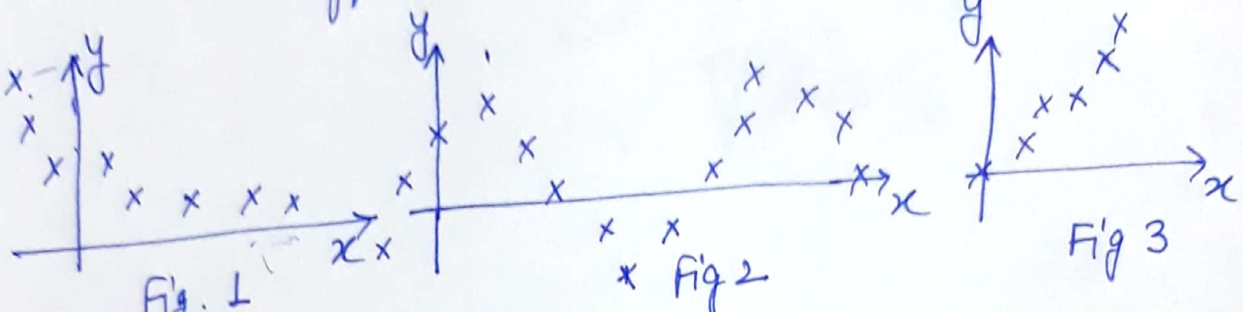

Fig. 1 Fig 2 Fig 3

Figs. 1, 2, 3 resembles a decaying exponential curve, a sinusoid and a geometric curve, respectively.

By a mathematical transformation these nonlinear functions $y(x)$ are transformed to a linear function $Z(x)$. Once these functions are transformed to linear form the coefficients of this linear function are determined by a least squares fit.

(i) Fitting an Exponential Curve :

Let $y = a e^{-bx}$ ——— (1) to be fitted.

The transformation is $Z = \log y$ ——— (2)

using this on eq^n (1),

$$Z = \log (a e^{-bx})$$

$$Z = \log a + (-bx) \quad \text{——— (3)}$$

Let $a_0 = \log a$ and $a_1 = -b$

$\therefore \quad z = a_0 + a_1 x$ —— (4)

eqn (4) is linear and we can use the normal eqns for linear regression,

$$n a_0 + (\Sigma x_i) a_1 = \Sigma z_i = \Sigma \log y_i \quad \text{—— (5)}$$

$$(\Sigma x_i) a_0 + (\Sigma x_i^2) a_1 = \Sigma x_i z_i = \Sigma x_i \log y_i \quad \text{—— (6)}$$

On solving these two equations,

$$a_0 = \frac{\Sigma \log y_i \, \Sigma x_i^2 - \Sigma x_i \, \Sigma x_i \log y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2} \quad \text{—— (7)}$$

and $$a_1 = \frac{n \Sigma x_i \, \Sigma \log y_i - \Sigma x_i \, \Sigma \log y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2} \quad \text{—— (8)}$$

From a_0 and a_1 we obtain a and b, as

$$a = e^{a_0}, \quad b = -a_1$$

Fitting a Hyperbola :

Let $y = \dfrac{1}{a + bx}$

If $z = \dfrac{1}{y}$ then $z = a + bx$

which can be solved by linear regression method.
