Let us consider the following three points:

$$x_0 = 2, \quad x_1 = 3, \quad \text{and} \quad x_2 = 4$$

Then

$$f_0 = 1.4142, \quad f_1 = 1.7321, \quad \text{and} \quad f_2 = 2$$

For $x = 2.5$, we have

$$l_0(2.5) = \frac{(2.5 - 3.0)(2.5 - 4.0)}{(2.0 - 3.0)(2.0 - 4.0)} = 0.3750$$

$$l_1(2.5) = \frac{(2.5 - 2.0)(2.5 - 4.0)}{(3.0 - 4.0)(3.0 - 2.0)} = 0.7500$$

$$l_2(2.5) = \frac{(2.5 - 2.0)(2.5 - 3.0)}{(4.0 - 2.0)(4.0 - 3.0)} = -0.125$$

$$p_2(2.5) = (1.4142)(0.3750) + (1.7321)(0.7500) + (2.0)(-0.125)$$
$$= 0.5303 + 1.2991 - 0.250 = 1.5794$$

The error is 0.0017 which is much less than the error obtained in Example 9.3

Example 9.5

Find the Lagrange interpolation polynomial to fit the following data.

i	0	1	2	3
x_i	0	1	2	3
$e^{x_i} - 1$	0	1.7183	6.3891	19.0855

Use the polynomial to estimate the value of $e^{1.5}$.

Lagrange basis polynomials are

$$l_0(x) = \frac{(x - 1)(x - 2)(x - 3)}{(0 - 1)(0 - 1)(0 - 3)}$$

$$= \frac{x^3 - 6x^2 + 11x - 6}{-3}$$

$$l_1(x) = \frac{(x - 0)(x - 2)(x - 3)}{(1 - 0)(1 - 2)(1 - 3)}$$

$$= \frac{x^3 - 5x^2 + 6x}{2}$$

```
            IF (I.NE.J) THEN
              LF = LF * (XP - X(J)) / (X(I) - X(J))
            ENDIF
20        CONTINUE
          SUM = SUM + LF * F(I)
30      CONTINUE
        FP = SUM

        WRITE(*,*)
        WRITE(*,*) 'LAGRANGIAN INTERPOLATION'
        WRITE(*,*)
        WRITE(*,*) 'Interpolated Function Value'
        WRITE(*,*) 'at X = ', XP, ' is', FP
        WRITE(*,*)

        STOP
        END
* ---------------  End of main LAGRAN ----------------- *
```

Test Run Results The program was used to compute the function value at $x = 2.5$ for the following table of data points:

x	2	3	4
f	1.4142	1.7321	2.0

The results are shown below:

```
Input number of data points(N)
3
Input data points X(I) and Function values F(I)
one set in each line
2  1.4142
3  1.7321
4  2.0
Input X value at which
interpolation is required
2.5

LAGRANGIAN INTERPOLATION

Interpolated Function Value
at X = 2.5000000 is 1.5794000

Stop - Program terminated.
```

9.5 NEWTON INTERPOLATION POLYNOMIAL

We have seen that, in Lagrange interpolation, we cannot use the work that has already been done if we want to incorporate another data point

Example 9.9

Repeat the estimation of sin 25 in Example 9.8 using Newton's backward difference formula

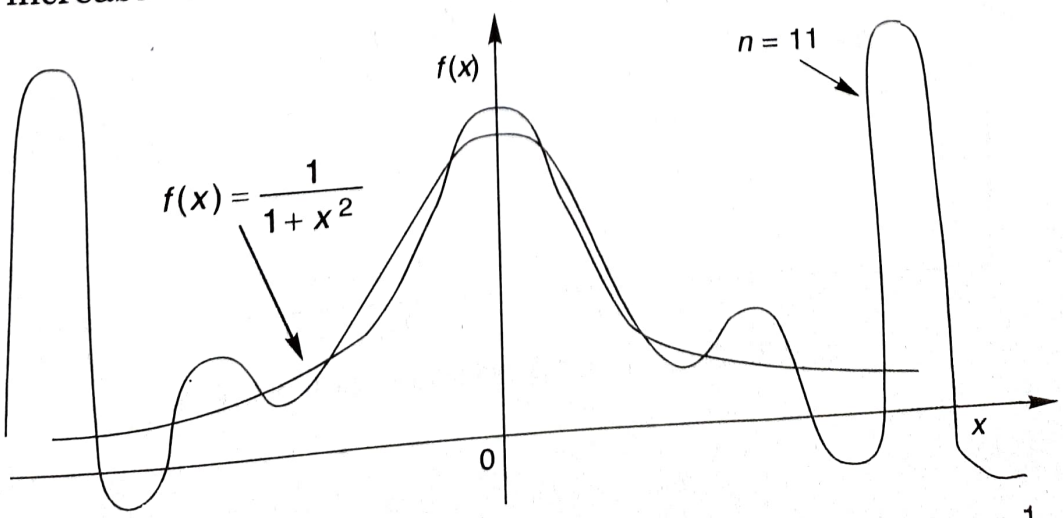
$$s = \frac{(x - x_n)}{h} = \frac{25 - 50}{10} = -2.5$$

Using Eq. (9.21), we get

$$\begin{aligned} p_4(2.5) &= 0.7660 + (-2.5)(0.1232) \\ &\quad + \frac{(-2.5)(-1.5)(-0.0196)}{2} \\ &\quad + \frac{(-2.5)(-1.5)(-0.5)(0.0044)}{6} \\ &\quad + \frac{(-2.5)(-1.5)(-0.5)(0.5)(-0.0004)}{24} \\ &= 0.4200 \end{aligned}$$

9.8 SPLINE INTERPOLATION

So far we have discussed how an interpolation polynomial of degree n can be constructed and used given a set of values of functions. There are situations in which this approach is likely to face problems and produce incorrect estimates. This is because the interpolation takes a global rather than a local view of data. It has been proved that when n is large compared to the order of the "true" function, the interpolation polynomial of degree n does not provide accurate results at the ends of the range. This is illustrated in Fig. 9.6. Note that the interpolation polynomial contains undesirable maxima and minima between the data points. This only shows that increasing the order of polynomials does not necessarily increase the accuracy.



CHAPTER 10

Curve Fitting: Regression

10.1 INTRODUCTION

In the previous chapter we discussed various methods of curve fitting for data points of well-defined functions. In this chapter, we will discuss methods of curve fitting for experimental data.

In many applications, it often becomes necessary to establish a mathematical relationship between experimental values. This relationship may be used for either testing existing mathematical models or establishing new ones. The mathematical equation can also be used to predict or forecast values of the dependent variable. For example, we would like to know the maintenance cost of an equipment (or a vehicle) as a function of age (or mileage) or the relationship between the literacy level and population growth. The process of establishing such relationships in the form of a mathematical equation is known as *regression analysis* or *curve fitting*.

Suppose the values of y for the different values of x are given. If we want to know the effect of x on y, then we may write a functional relationship

$$y = f(x)$$

The variable y is called the *dependent* variable and x the *independent variable*. The relationship may be either linear or nonlinear as shown in Fig. 10.1. The type of relationship to be used should be decided by the experiment based on the nature of scatteredness of data.

It is a standard practice to prepare a *scatter diagram* as shown in Fig. 10.2 and try to determine the functional relationship needed to fit the points. The line should best fit the plotted points. This means that the

average error introduced by the assumed line should be minimum. The parameters a and b of the various equations shown in Fig. 10.1 should be evaluated such that the equations best represent the data.

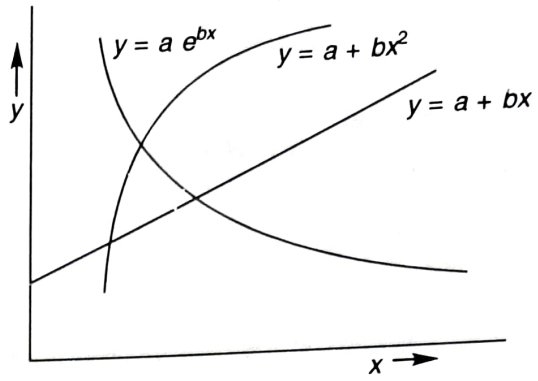


Fig. 10.1 Various relationships between x and y

We shall discuss in this chapter a technique known as *least-squares regression* to fit the data under the following situations:

1. Relationship is linear
2. Relationship is transcendental
3. Relationship is polynomial
4. Relationship involves two or more independent variables

10.2 FITTING LINEAR EQUATIONS

Fitting a straight line is the simplest approach of regression analysis. Let us consider the mathematical equation for a straight line

$$y = a + bx = f(x)$$

to describe the data. We know that a is the intercept of the line and b its slope. Consider a point (x_i, y_i) as shown in Fig. 10.2. The vertical distance of this point from the line $f(x) = a + bx$ is the error q_i. Then,

$$\begin{aligned} q_i &= y_i - f(x_i) \\ &= y_i - a - bx_i \end{aligned} \tag{10.1}$$

There are various approaches that could be tried for fitting a "best" line through the data. They include:

1. Minimise the sum of errors, i.e., minimise

$$\sum q_i = \sum (y_i - a - bx_i) \tag{10.2}$$

2. Minimise the sum of absolute values of errors

$$\sum |q_i| = \sum |(y_i - a - bx_i)| \tag{10.3}$$

3. Minimise the sum of squares of errors

$$\sum q_i^2 = \sum (y_i - a - bx_i)^2 \tag{10.4}$$

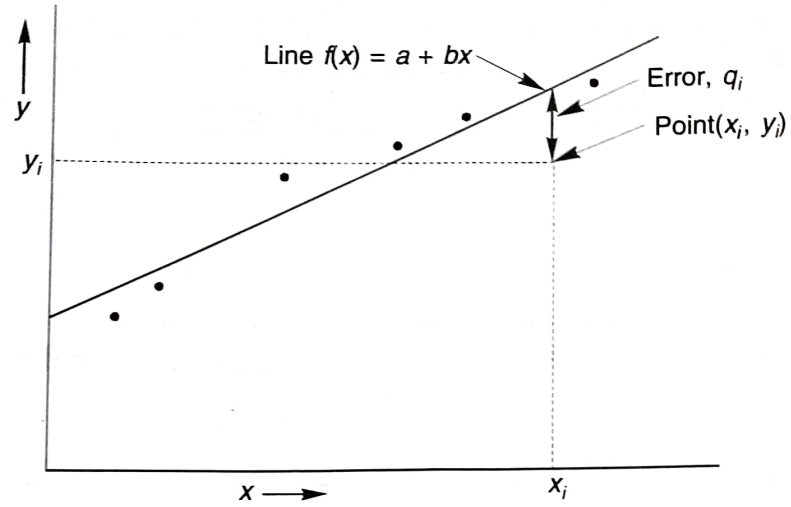


Fig. 10.2 Scatter diagram

It can be easily verified that the first two strategies do not yield a unique line for a given set of data. The third strategy overcomes this problem and guarantees a unique line. The technique of minimising the sum of squares of errors is known as *least squares regression*. In this section we consider the least-squares fit of a straight line.

Least Squares Regression

Let the sum of squares of individual errors be expressed as

$$Q = \sum_{i=1}^{n} q_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2$$

$$= \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{10.5}$$

In the method of least squares, we choose a and b such that Q is minimum. Since Q depends on a and b, a necessary condition for Q to be minimum is

$$\frac{\partial Q}{\partial a} = 0 \quad \text{and} \quad \frac{\partial Q}{\partial b} = 0$$

Then

$$\frac{\partial Q}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\frac{\partial Q}{\partial b} = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0 \tag{10.6}$$

Thus

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

$$\sum x_i y_i = a \sum x_i + b \sum x_i^2 \tag{10.7}$$

These are called *normal equations*. Solving for a and b, we get

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} = \bar{y} - b\bar{x} \tag{10.8}$$

where $\bar{x}$ and $\bar{y}$ are the averages of x values and y values, respectively.

Example 10.1

Fit a straight line to the following set of data

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

The various summations are given as follows:

	x_i	y_i	x_i^2	$x_i y_i$
	1	3	1	3
	2	4	4	8
	3	5	9	15
	4	6	16	24
	5	8	25	40
Σ	15	26	55	90

Using Eq. (10.8),

$$b = \frac{5\times 90 - 15\times 26}{5\times 55 - 15^2} = 1.20$$

$$a = \frac{26}{5} - 1.20\times\frac{15}{5} = 1.60$$

Therefore, the linear equation is

$$y = 1.6 + 1.2x$$

The *regression line* along with the data is shown in Fig. 10.3.

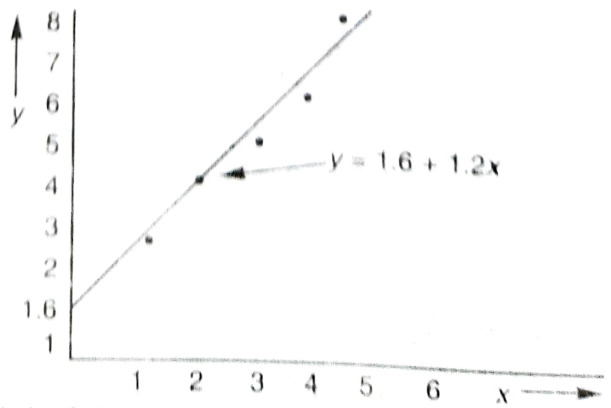


Fig. 10.3 Plot of the data and regression line of example 10.1

Algorithm

It is relatively simple to implement the linear regression on a computer. The coefficients a and b can be evaluated using Algorithm 10.1

Linear Regression

1. Read data values
2. Compute sum of powers and products

$$\Sigma x_i, \Sigma y_i, \Sigma x_i^2, \Sigma x_i y_i$$

3. Check whether the denominator of the equation for b is zero.
4. Compute b and a.
5. Print out the equation.
6. Interpolate data, if required.

Algorithm 10.1

Program LINREG

Program LINREG implements Algorithm 10.1. The program reads a table of data points and decides a straight line equation to fit the data using the method of least squares regression.

```
* ------------------------------------------------ *
    PROGRAM LINREG
* ------------------------------------------------ *
* Main program                                     *
*   This program fits a line Y = A + BX to a given *
*   set of data points by the method of least squares *
* ------------------------------------------------ *
* Functions invoked                                *
*   ABS                                            *
* ------------------------------------------------ *
* Subroutines used                                 *
*   NIL                                            *
* ------------------------------------------------ *
* Variables used                                   *
*   X, Y - Data arrays                             *
*   N - Number of data sets                        *
*   SUMX - Sum of x values                         *
*   SUMY - Sum of y values                         *
*   SUMXX - Sum of squares of x values             *
*   SUMXY - Sum of products of x and y             *
*   XMEAN - Mean of x values                       *
*   YMEAN - Mean of y values                       *
*   A - y intercept of the line                    *
*   B - Slope of the line                          *
* ------------------------------------------------ *
```

```
      WRITE(*,*)  'LINEAR REGRESSION LINE Y = A + BX'
      WRITE(*,*)
      WRITE(*,*)  'THE COEFFICIENTS ARE:'
      WRITE(*,*)  '     A = ', A
      WRITE(*,*)  '     B = ', B
      WRITE(*,*)
      STOP
      END
* ---------------- End of main LINREG ---------------- *
```

Test Run Results Shown below is the interactive data input and the linear regression line parameters computed by the program LINREG.

```
            LINEAR REGRESSION
Input number of data points N
5
Input X and Y values, one set on each line
1 3
2 5
3 7
4 9
5 11

LINEAR REGRESSION LINE Y = A + BX

THE COEFFICIENTS ARE:
  A = 1.0000000
  B = 2.0000000
Stop - Program terminated.
```

10.3 FITTING TRANSCENDENTAL EQUATIONS

The relationship between the dependent and independent variables is not always linear. Look at Fig. 10.4. The nonlinear relationship between

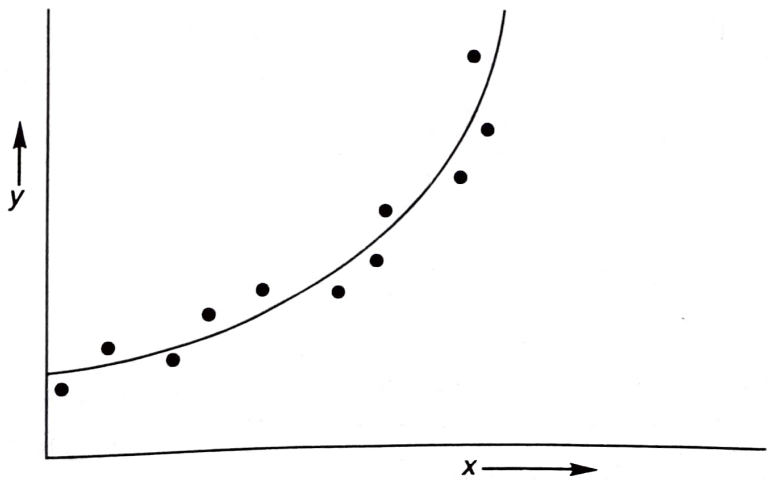


Fig.10.4 Data would fit a nonlinear curve better than a linear one

them may exist in the form of transcendental equations (or higher order polynomials). For example, the familiar equation for population growth is given by

$$P = p_0 e^{kt} \tag{10.9}$$

where p_0 is the initial population, k is the rate of growth and t is time. Another example of nonlinear model is the gas low relating to the pressure and volume, as given by

$$p = a v^b \tag{10.10}$$

Let us consider Eq. (10.10) first. If we observe values of p for various values of v, we can then determine the parameters a and b. Using the method of least squares, the sum of the squares of all errors can be written as

$$Q = \sum_{i=1}^{n} [p_i - a v_i^b]^2$$

To minimise Q, we have

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

We can prove that

$$\sum p_i v_i^b = a \sum (v_i^b)^2$$

$$\sum p_i v_i^b \ln v_i = a \sum (v_i^b)^2 \ln v_i$$

These equations can be solved for a and b. But since b appears under the summation sign, an iterative technique must be employed to solve for a and b.

However, this problem can be solved by using the algorithm given in the previous section in the following way: let us rewrite the equation using the conventional variables x and y as

$$y = a x^b$$

If we take logarithm on both the sides, we get

$$\ln y = \ln a + b \ln x \tag{10.11}$$

This equation is similar in form to the linear equation and, therefore, using the same procedure we can evaluate the parameters a and b.

$$b = \frac{n \sum \ln x_i \ln y_i - \sum \ln x_i \sum \ln y_i}{n \sum (\ln x_i)^2 - (\sum \ln x_i)^2} \tag{10.12}$$

$$\ln a = R = \frac{1}{n} (\sum \ln y_i - b \sum \ln x_i)$$

$$a = e^R \tag{10.13}$$

Similarly, we can linearise the exponential model shown in Eq. (10.9) by taking logarithm on both the sides. This would yield

$$\ln P = \ln P_0 + kt \ln e$$

Since, $\ln e = 1$,

we have $$\ln P = \ln P_o + kt \tag{10.14}$$

This is similar to the linear equation

$$y = a + bx$$

where $y = \ln P$, $a = \ln P_0$, $b = k$, and $x = t$. We can now easily determine a and b and then P_0 and k.

There is a third form of nonlinear model known as *saturation-growth-rate* equation, as shown below:

$$p = \frac{k_1 t}{k_2 + t} \tag{10.15}$$

This can be linearised by taking inversion of the terms. That is

$$\frac{1}{p} = \left(\frac{k_2}{k_1}\right)\frac{1}{t} + \frac{1}{k_1} \tag{10.16}$$

This is again similar to the linear equation

$$y = a + bx$$

where

$$y = \frac{1}{p}, \qquad x = \frac{1}{t}$$

$$a = \frac{1}{k_1}, \qquad b = \frac{k_2}{k_1}$$

Once we obtain a and b, they could be transformed back into the original form for the purpose of analysis.

Example 10.2

Given the data table

x	1	2	3	4	5
y	0.5	2	4.5	8	12.5

fit a power-function model of the form

$$y = ax^b$$

Various quantities required in equation (10.12) are tabulated below:

x_i	y_i	$\ln x_i$	$\ln y_i$	$(\ln x_i)^2$	$(\ln x_i)(\ln y_i)$
1	0.5	0	−0.6931	0	0
2	2	0.6931	0.6931	0.4805	0.4804
3	4.5	1.0986	1.5041	1.2069	1.6524
4	8	1.3863	2.0794	1.9218	2.8827
5	12.5	1.6094	2.5257	2.5903	4.0649
Sum		4.7874	6.1092	6.1995	9.0804

Using Eq. (10.12),

$$b = \frac{(5)(9.0804) - (4.7874)(6.1092)}{(5)(6.1995) - (4.7874)^2}$$

$$= \frac{45.402 - 29.2472}{30.9975 - 22.9192}$$

$$= 1.9998$$

$$\ln a = \frac{6.1092 - (1.9998)(4.7847)}{5}$$

$$= -0.6929$$

$$a = 0.5001$$

Thus, we obtain the power-function equation as

$$y = 0.5001\, x^{1.9998}$$

Note that the data have been derived from the equation

$$y = \frac{x^2}{2}$$

The discrepancy in the computed coefficients is due to roundoff errors.

Example 10.3

The temperature of a metal strip was measured at various time intervals during heating and the values are given in the table below:

time, t (min)	1	2	3	4
temp, T (°C)	70	83	100	124

If the relationship between the temperature T and time t is of the form

$$T = be^{t/4} + a$$

estimate the temperature at $t = 6$ min.

We can write the temperature equation in the form

$$y = b\, f(x) + a$$

This is similar to the linear equation except that the variable x is replaced by the function $f(x)$. Therefore, we can solve for the parameters a and b using Eq. (10.8) by replacing

$$x_i \quad \text{by} \quad f(x_i)$$

$$\sum x_i \quad \text{by} \quad \sum f(x_i)$$

$$\sum x_i^2 \quad \text{by} \quad \sum f(x_i)^2$$

Thus,

$$b = \frac{n(\sum f(x_i)y_i) - \sum f(x_i)\sum y_i}{n\sum[f(x_i)]^2 - [\sum f(x_i)]^2}$$

$$a = \frac{\sum y_i - b\sum f(x_i)}{n}$$

We can set up the following table to obtain the various terms. Note that $f(x) = e^{0.25x}$.

	x	y	$f(x)$	$y.\ f(x)$	$[f(x)]^2$
	1	70	1.28	89.89	1.65
	2	83	1.65	136.84	2.72
	3	100	2.12	211.70	4.48
	4	124	2.72	337.07	7.39
Sum		377	7.77	775.5	16.24

Now,

$$b = \frac{(4)(775.5) - (7.77)(377)}{(4)(16.24) - (7.77)^2}$$

$$= 37.62$$

$$a = \frac{377 - (37.62)(7.77)}{4}$$

$$= 21.16$$

The equation is

$$T = 37.62\, e^{0.25t} + 21.16$$

The temperature, when $t = 6$, is

$$T = 37.62\, e^{0.25(6)} + 21.16$$

$$= 189.76°C$$

10.4 FITTING A POLYNOMIAL FUNCTION

When a given set of data does not appear to satisfy a linear equation, we can try a suitable polynomial as a regression curve to fit the data. The least squares technique can be readily used to fit the data to a polynomial.

Consider a polynomial of degree $m - 1$

$$y = a_1 + a_2 x + a_3 x^2 + \ldots + a_m x^{m-1} \qquad (10.17)$$

$$= f(x)$$