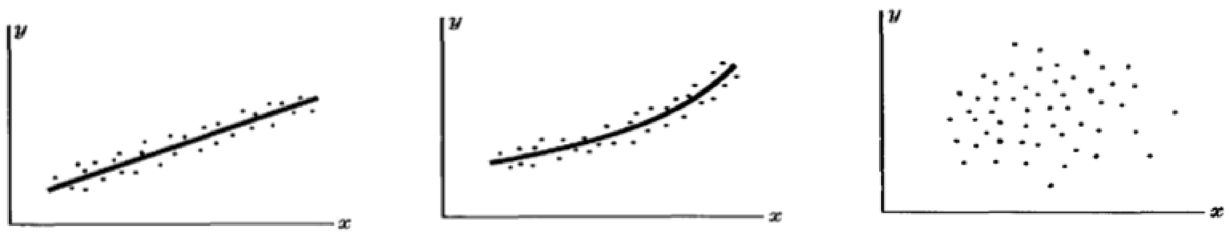
6 Principle of Least Squares

Course	**B.Sc. (H) Physics**
Semester	**VI**
Paper Name	**Advanced Mathematical Physics - II**
Unique Paper Code	**32227625**
Teacher's Name	**Ms Sonia Yogi**
Department	**Physics and Electronics, Hansraj College DU**

6.1 Introduction

Suppose x and y denote, respectively the height and weight of an adult male. Then a sample of n individuals would reveal the heights $x_1, x_2, \ldots, x_n$ and the corresponding weights $y_1, y_2, \ldots, y_n$. Our next step is to plot the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ on a rectangular coordinate system. The resulting set of points is sometimes called a *scatter diagram*.

From the scatter diagram it is often possible to visualize a smooth curve approximating the data. Such a curve is called an *approximating curve*.



(a) Linear relationship (b) Non-linear relationship (c) No relationship

Fig. 6.1: Approximating Curves

In Fig. 6.1(a), for example, the data appear to be approximated well by a straight line, and we say that a *linear relationship* exists between the variables. In Fig. 6.2(b), however, although a relationship exists between the variables, it is not a linear relationship and so we call it a *nonlinear relationship* . In Fig. 6.3(c) there appears to be no relationship between the variables.

The general problem of finding equations of approximating curves that fit given sets of data is called *curve fitting*. In practice the type of equation is often suggested from the scatter diagram. For Fig. 6.1(a) we could use a *straight line*

$$y = a + b\,x \tag{6.1}$$

while for Fig. (6.2) we could try a *parabola* or *quadratic curve*:

$$y = a + b\,x + c\,x^2 \tag{6.2}$$

Sometimes it helps to plot scatter diagrams in terms of *transformed variables*. For example, if $\log y$ vs x leads to a straight line, we would try $\log y = a + b\,x$ as an equation for the approximating curve.

6.2 The Method of Least Squares

Consider Fig. 6.2 in which the data points are $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. For given value of x, say, x_1, there will be a difference between the value y_1 and the corresponding value as determined from the curve C. We denote this difference by d_1, which is sometimes referred to as a *deviation*, *error*, or *residual* and may be positive, negative, or zero. Similarly, corresponding to the values $x_2, \ldots, x_n$, we obtain the deviations $d_2, \ldots, d_n$.

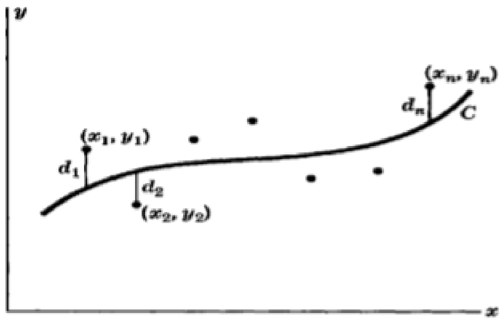


Fig. 6.2: Showing the deviations

A measure of the goodness of fit of the curve C to the set of data is provided by the quantity,

$$S = d_1^2 + d_2^2 + \cdots + d_n^2 \tag{6.3}$$

If S is small, the fit is good, if it is large, the fit is bad. We therefore make the following definition:

Definition Of all curves in a given family of curves approximating a set of n data points, a curve having the property that

$$S = d_1^2 + d_2^2 + \cdots + d_n^2 = \text{a minimum} \tag{6.4}$$

is called a *best-fitting curve* in the family.

A curve having this property is said to fit the data in the *least-squares* sense and is called a *least-squares curve*. A line having this property is called a *least-squares line*; a parabola with this property is called a *least-squares parabola*, etc.

It is customary to employ the above definition when x is the independent variable and y is the dependent variable. Unless otherwise specified, we shall consider y as the dependent and x as the independent variable.

6.3 The Least-Squares Line

By using the above definition, we will now show that the least-squares line approximating the set of points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ has the equation

$$y = a + b\,x \tag{6.5}$$

where the constants a and b are determined by solving simultaneously the equations

$$\sum y = n\,a + b \sum x$$
$$\sum x\,y = a \sum x + b \sum x^2$$

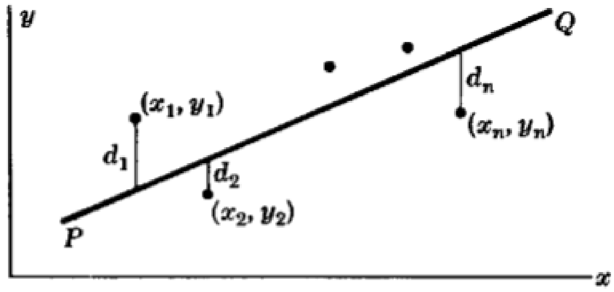


Fig. 6.3: Showing the deviations

Refer to Fig. 6.3, the values of y on the least-squares line corresponding to $x_1, x_2, \ldots, x_n$ are

$$a + b\,x_1,\; a + b\,x_2, \ldots,\; a + b\,x_n$$

The corresponding vertical deviations are

$$d_1 = a + b\,x_1 - y_1,\; d_2 = a + b\,x_2 - y_2, \ldots,\; d_n = a + b\,x_n - y_n$$

and the sum of the squares of the deviations is

$$\begin{aligned} S &= d_1^2 + d_2^2 + \cdots + d_n^2 \\ &= (a + b\,x_1 - y_1)^2 + (a + b\,x_2 - y_2)^2 + \cdots + (a + b\,x_n - y_n)^2 \\ &= \sum (a + b\,x - y)^2 \end{aligned}$$

Necessary conditions for S to be a minimum are

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

$$\text{i.e.,} \qquad \sum (a + b\,x - y) = 0 \qquad \text{and} \qquad \sum (a\,x + b\,x^2 - x\,y) = 0$$

$$\text{or,} \qquad \sum 2(a + b\,x - y) = 0 \qquad \text{and} \qquad \sum 2\,x(a + b\,x - y) = 0$$

which gives

$$\sum y = n\,a + b \sum x \tag{6.6}$$

and

$$\sum x\,y = a\sum x + b\sum x^2 \tag{6.7}$$

Solving the eqs. (6.6) and (6.7), we get

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum x\,y)}{n\,(\sum x^2) - (\sum x)^2} \tag{6.8}$$

and

$$b = \frac{n\,(\sum x\,y) - (\sum x)(\sum y)}{n\,(\sum x^2) - (\sum x)^2} \tag{6.9}$$

Note: From eq. (5.6), we have

$$\begin{aligned}
& \sum y = n\,a + b\sum x \\
\implies\quad & \frac{1}{n}\sum y = a + b\,\frac{1}{n}\sum x \\
\implies\quad & \bar{y} = a + b\,\bar{x} \qquad \left[\because\ \bar{z} = \frac{1}{n}\sum z\right] \\
\implies\quad & a = \bar{y} - b\,\bar{x}
\end{aligned}$$

where,

$$\begin{aligned}
b &= \frac{n\,(\sum x\,y) - (\sum x)(\sum y)}{n\,(\sum x^2) - (\sum x)^2} \\
&= \frac{n\,(\sum x\,y) - (n\,\bar{x})(n\,\bar{y})}{n\,(\sum x^2) - (n\,\bar{x})^2} \qquad \left[\because\ \bar{z} = \frac{1}{n}\sum z \implies \sum z = n\,\bar{z}\right] \\
&= \frac{\sum x\,y - n\,\bar{x}\,\bar{y}}{\sum x^2 - n\,\bar{x}^2} \\
&= \frac{\sum x\,y - n\,\bar{x}\,\bar{y} + n\,\bar{x}\,\bar{y} - n\,\bar{x}\,\bar{y}}{\sum x^2 - n\,\bar{x}^2 + n\,\bar{x}^2 - n\,\bar{x}^2} \\
&= \frac{\sum x\,y - \bar{x}\sum y + \sum \bar{x}\,\bar{y} - \bar{y}\sum x}{\sum x^2 - \bar{x}\sum x + \sum \bar{x}^2 - \bar{x}\sum x} \qquad \left[\because\ \bar{z} = \frac{1}{n}\sum z \to \sum z = n\,\bar{z} = \sum \bar{z}\right] \\
&= \frac{\sum [x\,y - \bar{x}\,y + \bar{x}\,\bar{y} - \bar{y}\,x]}{\sum [x^2 - 2\,x\,\bar{x} + \bar{x}^2]} \\
&= \frac{\sum [x(y-\bar{y}) - \bar{x}(y-\bar{y})]}{\sum (x-\bar{x})^2} \\
&= \frac{\sum (x-\bar{x})\,(y-\bar{y})}{\sum (x-\bar{x})^2}
\end{aligned}$$

6.4 Examples

Example 1. Find the best values of a and b so that $y = a + b\,x$ fits the data given in the table:

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

Solution Let the least-squares line to the given data be

$$y = a + b\,x \tag{6.10}$$

then normal equations are ($n = 5$)

$$\sum y = n\,a + b\sum x$$
$$\sum x\,y = a\sum x + b\sum x^2 \tag{6.11}$$

Consider the following table:

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

Eqs. (6.11) becomes

$$204 = 5\,a + 15\,b$$
$$748 = 15\,a + 55\,b \tag{6.12}$$

Solving eqs. (6.11), we get

$$a = 0 \qquad \text{and} \qquad b = 68/5 \tag{6.13}$$

Thus, the required line is

$$y = \frac{68}{5}\,x$$

Example 2. Find the best values of a and b so that $y = a\,e^{b\,x}$ fits the data given in the table:

x	2	4	6	8	10
y	4.077	11.084	30.128	81.897	222.62

Solution Given $y = a\,e^{b\,x}$; Taking logarithm both sides,

$$\therefore \qquad \ln y = \ln a + b\,x$$

Let the least-squares line to the given data be

$$Y = A + b\,x \tag{6.14}$$

Where, $Y = \ln y$ and $A = \ln a$.
then normal equations are ($n = 5$)

$$\sum Y = n\,A + b\sum x$$
$$\sum x\,Y = A\sum x + b\sum x^2 \tag{6.15}$$

Consider the following table:

x	y	$Y = \ln y$	$x\,Y$	x^2
2	4.077	1.4054	2.8108	4
4	11.084	2.4055	9.6220	16
6	30.128	3.4055	20.4330	36
8	81.897	4.4055	35.2440	64
10	222.62	5.4055	54.0550	100
$\sum x = 30$		$\sum Y = 17.0274$	$\sum x\,Y = 122.1648$	$\sum x^2 = 220$

Eqs. (6.15) becomes

$$\begin{aligned} 17.0274 &= 5\,A + 30\,b \\ 122.1648 &= 30\,A + 220\,b \end{aligned} \tag{6.16}$$

Solving eqs. (6.16), we get

$$A \approx 0.40542 \qquad \text{and} \qquad b \approx 0.50001$$

Which gives

$$\ln a \approx 0.40542 \qquad \text{and} \qquad b \approx 0.50001$$

or,

$$a \approx e^{0.40542} \qquad \text{and} \qquad b \approx 0.50001$$

Thus, the required values are

$$a \approx 1.450 \qquad \text{and} \qquad b \approx 0.50001$$

Example 3. Find the best values of a and b so that $y = a\,b^x$ fits the data given in the table:

x	02	07	13	22	28
y	09	14	26	70	130

Solution Given $y = a\,b^x$; Taking logarithm both sides,

$$\therefore \qquad \ln y = \ln a + x\,\ln b$$

Let the least-squares line to the given data be

$$Y = A + B\,x \tag{6.17}$$

Where, $Y = \ln y$, $A = \ln a$ and $B = \ln b$.
then normal equations are ($n = 5$)

$$\begin{aligned} \sum Y &= n\,A + B\sum x \\ \sum x\,Y &= A\sum x + B\sum x^2 \end{aligned} \tag{6.18}$$

Consider the following table:

x	y	$Y = \ln y$	$x\,Y$	x^2
02	09	2.1972	4.3944	04
07	14	2.6390	18.4730	49
13	26	3.2580	42.3540	169
22	70	4.2485	93.4670	484
28	130	4.8675	136.2900	784
$\sum x$ = 72		$\sum Y$ = 17.2102	$\sum x\,Y$ = 294.9784	$\sum x^2$ = 1490

Eqs. (6.18) becomes

$$\begin{aligned} 17.2102 &= 5\,A + 72\,B \\ 294.9784 &= 72\,A + 1490\,B \end{aligned} \qquad (6.19)$$

Solving eqs. (6.19), we get

$$A \approx 1.9438 \qquad \text{and} \qquad B \approx 0.10404$$

Which gives

$$\ln a \approx 1.9438 \qquad \text{and} \qquad \ln b \approx 0.10404$$

or,

$$a \approx e^{1.9438} \qquad \text{and} \qquad b \approx e^{0.10404}$$

Thus, the required values are

$$a \approx 6.9852 \qquad \text{and} \qquad b \approx 1.1096$$

Example 4. Find the best values of a and b so that $y = a\,x^b$ fits the data given in the table:

x	80	40	20	10	5
y	333	375	422	475	533

Solution Given $y = a\,x^b$; Taking logarithm both sides,

$$\therefore \qquad \ln y = \ln a + b\,\ln x$$

Let the least-squares line to the given data be

$$Y = A + b\,X \qquad (6.20)$$

Where, $Y = \ln y$, $A = \ln a$ and $X = \ln x$.
then normal equations are ($n = 5$)

$$\begin{aligned} \sum Y &= n\,A + b\sum X \\ \sum X\,Y &= A\sum X + b\sum X^2 \end{aligned} \qquad (6.21)$$

Consider the following table:

x	y	$X = \ln x$	$Y = \ln y$	$X\,Y$	X^2
80	333	4.3820	5.808	25.4507	19.2019
40	375	3.6889	5.9269	21.8637	13.6080
20	422	2.9957	6.045	18.109	8.9742
10	475	2.303	6.1633	14.1941	5.3038
5	533	1.609	6.2785	10.1021	2.5889
		$\sum X$ = 14.9786	$\sum Y$ = 30.2217	$\sum X\,Y$ = 89.7196	$\sum X^2$ = 49.6768

Eqs. (6.21) becomes

$$\begin{aligned} 30.2217 &= 5\,A + 14.9786\,b \\ 89.7196 &= 14.9786\,A + 49.6768\,b \end{aligned} \tag{6.22}$$

Solving eqs. (6.22), we get

$$A \approx 6.5532 \qquad \text{and} \qquad b \approx -0.1699$$

Which gives

$$\ln a \approx 6.5532 \qquad \text{and} \qquad b \approx -0.1699$$

or,

$$a \approx e^{6.5532} \qquad \text{and} \qquad b \approx -0.1699$$

Thus, the required values are

$$a \approx 701.4853 \qquad \text{and} \qquad b \approx -0.1699$$

6.5 Problems

1. Find a least-squares line to the data given in the table below:

x	3	5	6	8	9	11
y	2	3	4	6	5	8

2. Find the best values of a and b so that $y = a\,e^{b\,x}$ fits the data given in the table below:

x	5	10	15	20	25
y	1.09	2.09	04	7.67	14.70

3. Find the best values of a and b so that $y = a\,b^x$ fits the data given in the table below:

x	05	07	10	15	20
y	630.06	455.24	279.56	124.04	55.04

4. Find the best values of a and b so that $y = a\,x^b$ fits the data given in the table below:

x	10	20	30	40	50
y	0.91	1.51	2.04	2.51	2.96

5. Find the best values of γ and C so that $P\,V^{\gamma} = C$ fits the data given in the table below:

V	54.3	61.8	72.4	88.7	118.6	194
P	61.2	49.5	37.6	28.4	19.2	10.1

6.6 References

(a). Introduction to Probability and Statistics, 2016 by Seymour Lipschutz and John J. Schiller.