

# Chapter 11

## The Method of Least Squares

### 11.1 Introduction

The *method of least squares* has many applications. For the purposes of this book, we will examine mainly its application in the fitting of the best straight line or curve to a series of experimental results, with the aim of determining the relationship existing between two variables. The method was originally developed by Legendre in 1805, while Gauss mentions that he had already used the method in 1794, at the age of 17, to determine the orbit of the asteroid Ceres. The problem which Legendre solved is the following:

Let us assume that we have a certain number ( $n > 2$ ) of linear equations  $A_r x + B_r y = K_r$ , in which  $A_r, B_r$  and  $K_r$  are constant. We may find pairs of values  $(x, y)$  which satisfy any one of the  $n$  equations, but these pairs do not satisfy all the equations simultaneously. In other words, the equations are not consistent with each other. They form an overdetermined system of equations. The problem that arises is to find the values of  $x$  and  $y$  which satisfy all the equations in the best possible way.

The answer depends, of course, on what we mean by the phrase ‘*the best possible way*’. Legendre stated the following principle: *The most probable value of a magnitude being measured is that for which the sum of the squares of the deviations of the measurements from this value is a minimum*. As we have seen in Chap. 9 (Sect. 9.3), the normal law of errors may be used in order to prove the principle of the most probable value. Inversely, Gauss derived the normal law of errors assuming that the mean value of a series is the most probable value of the magnitude being measured. It turns out that the sum of the squares of the deviations of the measurements from their mean value is the least possible (compared to the sum of the squares of the deviations from any other value).

Similar arguments may also be used to solve the problem of fitting the best straight line or curve to a series of experimental results, as we will show below. Strictly speaking, the method of least squares is valid only in those cases where the

results of the measurements are normally distributed relative to the real values of the quantities being measured. It is, however, also applied to cases in which the distribution (if and when this is known) is only approximately normal, but, also, in general, when the relative errors are small.

## 11.2 The Theoretical Foundation of the Method of Least Squares

Let the random variable  $\mathbf{y}$  be a function of only one independent random variable  $\mathbf{x}$ . We will assume that the true values of the two variables are related through the mathematical expression

$$y_0 = y_0(x, \alpha_0, \beta_0, \dots), \quad (11.1)$$

which gives the real value of  $y$  for a given  $x$ . The parameters  $\alpha_0, \beta_0, \dots$  are unknown to us. Our aim is to determine best estimates for these parameters, using the results of  $N$  measurements by which we have found for every value  $x_i$  of  $\mathbf{x}$  the corresponding value  $y_i$  of  $\mathbf{y}$ . We will assume that errors occur only in the values  $y_i$  and that the values  $x_i$  are known with absolute certainty. The problem becomes much more complicated if we assume that both  $x_i$  and  $y_i$  have errors.

We define

$$\delta y_{0i} = y_i - y_0(x_i) \quad (11.2)$$

to be the deviation of the measured value  $y_i$  from the true value  $y_0(x_i)$  which is predicted by the function  $y_0 = y_0(x, \alpha_0, \beta_0, \dots)$  for the value  $x_i$ . The probability densities of the deviations are normal, with corresponding standard deviations  $\sigma_{0i}$  (also unknown). Thus, the probability that the value of  $\mathbf{y}$  corresponding to  $x_i$  lies between  $y_i$  and  $y_i + \delta y_{0i}$  is

$$\delta P_0 \{y_i < \mathbf{y} \leq y_i + \delta y_{0i}\} = \frac{\delta y_{0i}}{\sqrt{2\pi} \sigma_{0i}} \exp \left\{ -\frac{[y_i - y_0(x_i)]^2}{2\sigma_{0i}^2} \right\}. \quad (11.3)$$

These are illustrated for the case of a linear relationship between  $y$  and  $x$  in Fig. 11.1a.

If the deviations are mutually independent, the composite probability that the result of the first measurement lies between  $y_1$  and  $y_1 + \delta y_{01}$ , the result of the second measurement lies between  $y_2$  and  $y_2 + \delta y_{02}$  etc. for all the  $N$  measurements is

$$d^N P_0 = \frac{1}{(\sqrt{2\pi})^N \sigma_{01} \sigma_{02} \dots \sigma_{0N}} \exp \left\{ -\left[ \frac{(\delta y_{01})^2}{2\sigma_{01}^2} + \frac{(\delta y_{02})^2}{2\sigma_{02}^2} + \dots + \frac{(\delta y_{0N})^2}{2\sigma_{0N}^2} \right] \right\} \delta y_{01} \delta y_{02} \dots \delta y_{0N}. \quad (11.4)$$

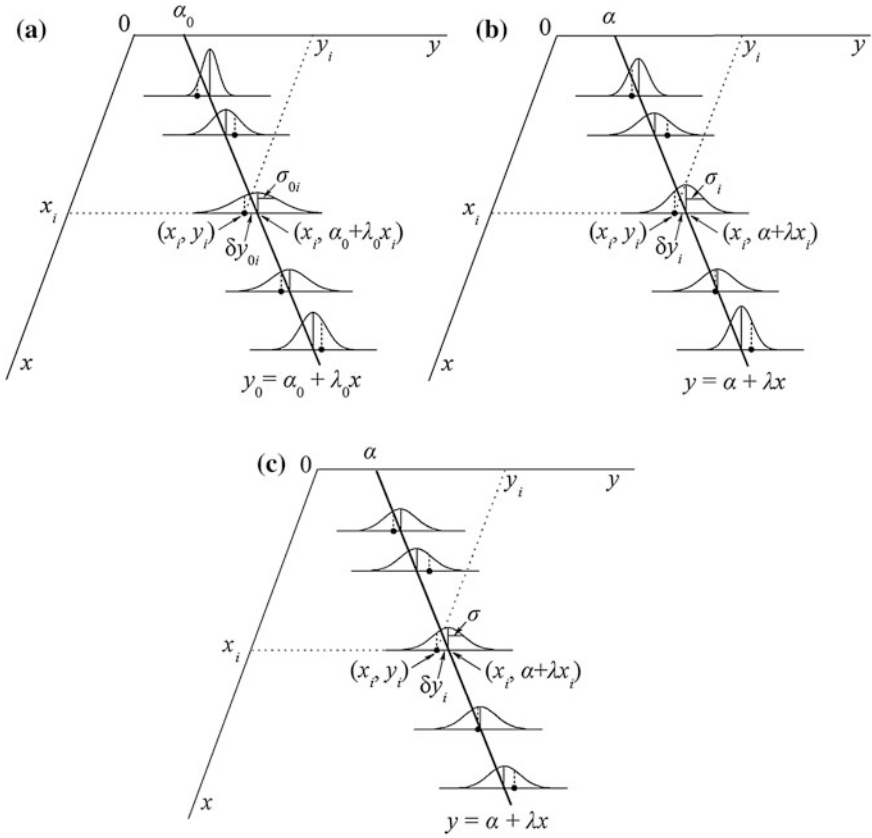


Fig. 11.1 Illustrating the method of least squares

In an  $N$ -dimensional space of errors, this probability is written as

$$\delta^N P_0 = \frac{1}{(\sqrt{2\pi})^N \sigma_0^N} e^{-\chi_0^2/2} \delta^N v_0, \quad (11.5)$$

where

$$\chi_0^2 \equiv \frac{(\delta y_{01})^2}{\sigma_{01}^2} + \frac{(\delta y_{02})^2}{\sigma_{02}^2} + \dots + \frac{(\delta y_{0N})^2}{\sigma_{0N}^2} = \sum_{i=1}^N \frac{(\delta y_{0i})^2}{\sigma_{0i}^2}, \quad (11.6)$$

$$\sigma_0^N \equiv \sigma_{01} \sigma_{02} \dots \sigma_{0N} \quad (11.7)$$

and the magnitude

$$\delta^N v_0 \equiv \delta y_{01} \delta y_{02} \dots \delta y_{0N} \quad (11.8)$$

may be considered to be the element of  $N$ -dimensional volume around the point  $(y_{01}, y_{02}, \dots, y_{0N})$ .

The values of the parameters  $\alpha_0, \beta_0, \dots$  which maximize the probability  $\delta^N P_0$  are those minimizing the quantity  $\chi_0^2$ . Thus, a number of equations equal to the number of the parameters  $\alpha_0, \beta_0, \dots$  are derived,

$$\frac{\partial \chi_0^2}{\partial \alpha_0} = 0, \quad \frac{\partial \chi_0^2}{\partial \beta_0} = 0, \quad \dots, \quad (11.9)$$

from which we would determine the parameters.

However, the true values of  $\delta y_{0i} = y_i - y_0(x_i)$  are not known. Neither do we know the true standard deviations  $\sigma_{0i}$ . We will assume a relationship between  $x$  and  $y$ ,

$$y = y(x, \alpha, \beta, \dots), \quad (11.10)$$

where  $\alpha, \beta, \dots$  are parameters which we wish to determine. These values will be the best estimates for  $\alpha_0, \beta_0, \dots$ . Instead of the deviations from the true values of  $y$ ,  $\delta y_{0i} = y_i - y_0(x_i)$ , we will use the deviations from the values given by the relation  $y = y(x, \alpha, \beta, \dots)$ ,

$$\delta y_i = y_i - y(x_i), \quad (11.11)$$

Furthermore, the standard deviations  $\sigma_{0i}$  will be replaced by the standard deviations  $\sigma_i$  estimated for the various values of  $y_i$  determined for a given value  $x_i$ . We then have

$$\chi^2 \equiv \frac{(\delta y_1)^2}{\sigma_1^2} + \frac{(\delta y_2)^2}{\sigma_2^2} + \dots + \frac{(\delta y_N)^2}{\sigma_N^2} = \sum_{i=1}^N \frac{(\delta y_i)^2}{\sigma_i^2}, \quad (11.12)$$

instead of  $\chi_0^2$ . These are illustrated in Fig. 11.1b.

The values of the parameters  $\alpha, \beta, \dots$  which maximize the probability  $\delta^N P$ , which is an estimate of  $\delta^N P_0$ , are those minimizing the quantity  $\chi^2$ . Thus, a number of equations equal to the number of the parameters  $\alpha, \beta, \dots$  are derived,

$$\frac{\partial \chi^2}{\partial \alpha} = 0, \quad \frac{\partial \chi^2}{\partial \beta} = 0, \quad \dots, \quad (11.13)$$

from which we determine the parameters.

It is noted that in the quantity  $\chi^2$  the deviations are weighted, with weights equal to  $1/\sigma_i^2$ . In all but in very rare occasions, however, the values of  $1/\sigma_i^2$  are not known. We then consider that a good approximation is that all the  $\sigma_i$  may be

substituted by one common one (also usually unknown)  $\sigma$  [see Fig. 11.1c]. The quantity to be minimized is then

$$S \equiv (\sigma \chi)^2 = \sum_{i=1}^N (\delta y_i)^2 = \sum_{i=1}^N [y_i - y(x_i)]^2. \quad (11.14)$$

### 11.3 The Fitting of Curves to Experimental Points

We will now use the theoretical result of Eq. (11.14) in specific applications.

#### 11.3.1 Straight Line

Assume that from measurements we have acquired the values of the magnitude  $y_i$  corresponding to  $N$  values of  $x_i$  ( $i = 1, 2, \dots, N$ ). We assume that the relation between  $x$  and  $y$  is of the form

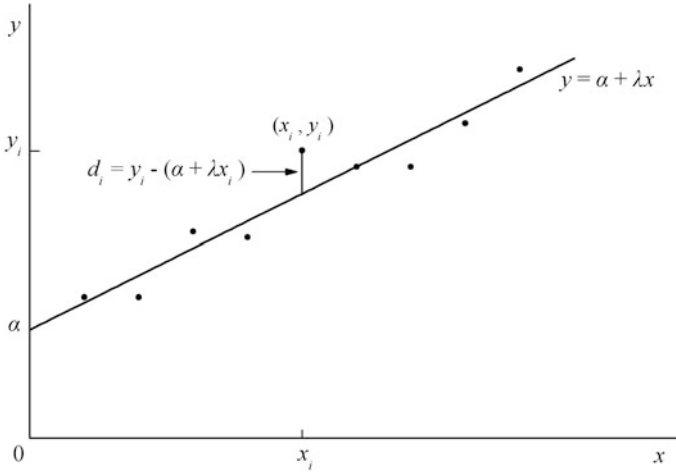
$$y = \alpha + \lambda x \quad (11.15)$$

and wish to determine the optimum values of the parameters  $\alpha$  and  $\lambda$ .

We assume that the values of the independent variable  $x$  are known with absolute accuracy. During the experimental procedure it is usually true that the variable  $x$  may be adjusted with adequate accuracy, and therefore this assumption is practically justified. The deviation of  $y_i$  from the real value  $y_{0,i}$  corresponding to the particular value  $x_i$  is governed by a Gaussian distribution with standard deviation  $\sigma$ , common to all measurements. In Fig. 11.1 (a) the true line connecting  $y$  to  $x$  is drawn, as well as the  $N$  experimental points  $(x_i, y_i)$ . For each one of them, the Gaussian distribution for the corresponding value of  $y_i$  is also drawn. The best estimates for the parameters  $\alpha$  and  $\lambda$ , according to the theory presented, are such that they maximize the probability of occurrence of the results obtained with the measurements. In Fig. 11.1 (c) the best straight line through the points is the one that will maximize the total length of the dashed lines.

Figure 11.2 shows the  $N$  points and the straight line  $y = \alpha + \lambda x$ . For the general point  $(x_i, y_i)$ , also drawn is the difference  $d_i = y_i - (\alpha + \lambda x_i)$  between the measured value  $y_i$  and the value predicted by the relation  $y = \alpha + \lambda x$  for  $x = x_i$ . The method of least squares requires the minimization of the sum

$$S \equiv \sum_{i=1}^N (y_i - y(x_i))^2 = \sum_{i=1}^N (y_i - \alpha - \lambda x_i)^2. \quad (11.16)$$



**Fig. 11.2** The fitting of a straight line to experimental results with the method of least squares. The  $N$  experimental points and the straight line  $y = \alpha + \lambda x$  are drawn. For the general point  $(x_i, y_i)$ , also shown is the difference  $d_i = y_i - (\alpha + \lambda x_i)$  between the measured value  $y_i$  and the value predicted by the relation  $y = \alpha + \lambda x$  for  $x = x_i$

This condition, which is the condition of Eq. (11.14), assumes that the weights in Eq. (11.12) are all taken to be the same. The more general case of measurements with different weights will be examined in Sect. 11.3.1.1.

Equating to zero the two partial derivatives of  $S$  with respect to  $\alpha$  and  $\lambda$ , we have the two equations:

$$\frac{\partial S}{\partial \alpha} = \frac{\partial}{\partial \alpha} \sum_{i=1}^N (y_i - \alpha - \lambda x_i)^2 = -2 \sum_{i=1}^N (y_i - \alpha - \lambda x_i) = 0 \quad (11.17)$$

$$\frac{\partial S}{\partial \lambda} = \frac{\partial}{\partial \lambda} \sum_{i=1}^N (y_i - \alpha - \lambda x_i)^2 = -2 \sum_{i=1}^N x_i (y_i - \alpha - \lambda x_i) = 0. \quad (11.18)$$

These are rewritten as

$$\alpha N + \lambda \sum_{i=1}^N x_i = \sum_{i=1}^N y_i \quad (11.19)$$

and

$$\alpha \sum_{i=1}^N x_i + \lambda \sum_{i=1}^N x_i^2 = \sum_{i=1}^N x_i y_i. \quad (11.20)$$

They are known as the *normal equations*. For convenience, we adopt the notation

$$\sum_{i=1}^N x_i \equiv [x] \quad \sum_{i=1}^N y_i \equiv [y] \quad \sum_{i=1}^N x_i^2 \equiv [x^2] \quad \sum_{i=1}^N x_i y_i \equiv [xy]. \quad (11.21)$$

Equations (11.19) and (11.20) now have the form

$$\alpha N + \lambda [x] = [y] \quad \text{and} \quad \alpha [x] + \lambda [x^2] = [xy]. \quad (11.22)$$

They are solved to give:

$$\alpha = \frac{[y][x^2] - [x][xy]}{N[x^2] - [x]^2} \quad (11.23)$$

$$\lambda = \frac{N[xy] - [x][y]}{N[x^2] - [x]^2}. \quad (11.24)$$

From Eq. (11.19), we notice that it is

$$\alpha + \lambda \frac{[x]}{N} = \frac{[y]}{N}, \quad (11.25)$$

which states that the straight line of least squares passes through the point  $(x = \frac{[x]}{N} = \bar{x}, \quad y = \frac{[y]}{N} = \bar{y})$ . The point K:  $(\bar{x}, \bar{y})$ , where  $\bar{x}$  and  $\bar{y}$  are the means of  $x$  and  $y$  respectively, may be considered to be the *center of the measurements*.

The accuracy with which we know  $\alpha$  and  $\lambda$  is a useful magnitude. We will give here the results without proof. A complete analysis is given in Appendix 1. In order to find the errors  $\delta \alpha$  and  $\delta \lambda$  in  $\alpha$  and  $\lambda$ , respectively, the standard deviation of the values  $y_i$  from the straight line must be evaluated. The best estimate for this quantity is:

$$\sigma_y = \sqrt{\frac{1}{N-2} \sum_{i=1}^N (y_i - \alpha - \lambda x_i)^2} \quad \text{or} \quad \sigma_y = \sqrt{\frac{[d^2]}{N-2}}, \quad (11.26)$$

where

$$d_i \equiv y_i - \alpha - \lambda x_i. \quad (11.27)$$

In terms of  $\sigma_y$ , the standard deviations of or the errors in  $\alpha$  and  $\lambda$  are, respectively,

$$\delta \alpha = \sigma_\alpha = \sigma_y \sqrt{\frac{[x^2]}{N[x^2] - [x]^2}} \quad (11.28)$$

and

$$\delta\lambda = \sigma_\lambda = \sigma_y \sqrt{\frac{N}{N[x^2] - [x]^2}}. \quad (11.29)$$

To make calculations easier, we note that it is

$$\delta\lambda = \delta\alpha \sqrt{\frac{N}{[x^2]}}. \quad (11.30)$$

Thus, the final value for the intercept of the  $y$ -axis by the straight line and the line's slope are:

$$\alpha \pm \delta\alpha \quad \text{and} \quad \lambda \pm \delta\lambda.$$

Having determined  $\alpha$  and  $\lambda$ , we may calculate the value of  $y$  for every value of  $x$  within the region of the validity of the law  $y = \alpha + \lambda x$ . We also need to know the error  $\delta y$  in this value. As explained in Appendix 1, the magnitudes  $\alpha$  and  $\lambda$  are not independent from each other. It would, therefore, be wrong to write the equation of the straight line as

$$y = (\alpha \pm \delta\alpha) + (\lambda \pm \delta\lambda)x \quad (11.31)$$

and the error in  $y$  as

$$\delta y = \sqrt{(\delta\alpha)^2 + (x\delta\lambda)^2}, \quad (11.32)$$

combining  $\delta\alpha$  and  $\delta\lambda$  as if they were independent of each other. In fact, the magnitudes that are mutually independent are the position  $(\bar{x}, \bar{y})$  of the center  $K$  of the least-squares straight line and its slope  $\lambda$ . The straight line is defined by its center  $(\bar{x}, \bar{y})$  and the independent from it orientation of the line, which is thought to rotate about its center.

Taking these into account, the error  $\delta y$  in  $y$ , for some value of  $x$ , is given by

$$\delta y = \frac{\sigma_y}{\sqrt{N}} \sqrt{1 + \frac{N^2}{N[x^2] - [x]^2} (x - \bar{x})^2}. \quad (11.33)$$

Finally, if it is assumed that the straight line passes through the origin, i.e. it is

$$y = \lambda x, \quad (11.34)$$



then  $\lambda$  is given by the relations

$$\lambda = \frac{[xy]}{[x^2]} = \frac{[y]}{[x]}, \quad (11.35)$$

where the second one is found by adding the terms of Eq. (11.34) over all the values of  $i$ . Thus, in this case, the least-squares straight line passes through the origin  $(0, 0)$  and the center of the measurements  $(\bar{x}, \bar{y})$ . The error in  $\lambda$  is evaluated using the relations:

$$\sigma_y = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (y_i - \lambda x_i)^2} \quad (11.36)$$

$$\delta \lambda = \frac{\sigma_y}{\sqrt{[x^2]}}. \quad (11.37)$$

Also, if the straight line is parallel to the  $x$ -axis, i.e. it is

$$y = a, \quad (11.38)$$

then

$$a = \frac{[y]}{N}, \quad (11.39)$$

the mean value of  $y$ . The error in  $a$  will, therefore, be equal to the standard deviation of the mean of the  $y$  values,

$$\delta a = \sigma_{\bar{y}}. \quad (11.40)$$

### Example 11.1

Apply the method of least squares to the measurements given in the first three columns of the table below, in order to fit to them a straight line  $y = \alpha + \lambda x$ . Find the value of  $y$  for  $x = 1.5$ .

It is  $N = 11$ .

The central point of the curve is  $K: (\bar{x}, \bar{y})$ , where

$$\bar{x} = [x]/N = 11/11 = 1.00 \quad \text{and} \quad \bar{y} = [y]/N = 35.44/11 = 3.22.$$

$i$	$x_i$	$y_i$	$x_i y_i$	$x_i^2$	$d_i$	$d_i^2$
1	0.0	0.92	0.000	0.00	-0.05090	0.00259
2	0.2	1.48	0.296	0.04	0.05892	0.00347
3	0.4	1.96	0.784	0.16	0.08874	0.00787
4	0.6	2.27	1.362	0.36	-0.05144	0.00265
5	0.8	2.61	2.088	0.64	-0.16162	0.02612
6	1.0	3.18	3.180	1.00	-0.04180	0.00175
7	1.2	3.80	4.560	1.44	0.12802	0.01639
8	1.4	4.01	5.614	1.96	-0.11216	0.01258
9	1.6	4.85	7.760	2.56	0.27766	0.07710
10	1.8	5.10	9.180	3.24	0.07748	0.00600
11	2.0	5.26	1.520	4.00	-0.21270	0.04524
Sums	11.0 = $[x]$	35.44 = $[y]$	45.344 = $[xy]$	15.40 = $[x^2]$		0.20176 = $[d^2]$

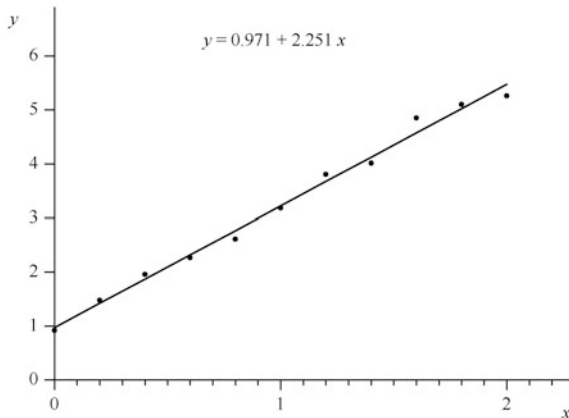
Thus,

$$\alpha = \frac{[y][x^2] - [x][xy]}{N[x^2] - [x]^2} = \frac{35.44 \times 15.40 - 11 \times 45.344}{11 \times 15.40 - 11^2} = 0.9709$$

$$\lambda = \frac{N[xy] - [x][y]}{N[x^2] - [x]^2} = \frac{11 \times 45.344 - 11 \times 35.44}{11 \times 15.40 - 11^2} = 2.2509$$

and the required straight line is  $y = 0.971 + 2.251x$ .

The experimental points and the straight line found have been drawn in the figure below.



To find the errors in  $\alpha$  and  $\lambda$  we first evaluate  $\sigma_y$ . In the table, we have calculated the deviations  $d_i \equiv y_i - \alpha - \lambda x_i$  and their squares. Thus, we find that

$$\sigma_y = \sqrt{\frac{1}{N-2} \sum_{i=1}^N (y_i - \alpha - \lambda x_i)^2} = \sqrt{\frac{0.20176}{9}} = 0.150.$$

The standard deviations of, or the errors in,  $\alpha$  and  $\lambda$  are, respectively,

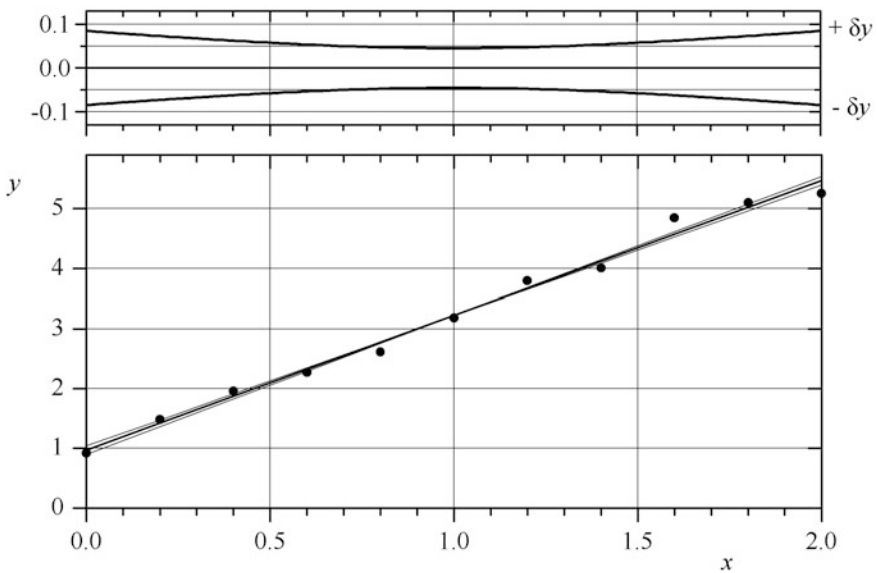
$$\delta \alpha = \sigma_\alpha = \sigma_y \sqrt{\frac{[x^2]}{N[x^2] - [x]^2}} = 0.150 \sqrt{\frac{15.40}{11 \times 15.40 - 11^2}} = 0.084$$

and

$$\delta \lambda = \sigma_\lambda = \sigma_y \sqrt{\frac{N}{N[x^2] - [x]^2}} = 0.150 \sqrt{\frac{11}{11 \times 15.40 - 11^2}} = 0.071.$$

Therefore, we have found that  $\alpha = 0.971 \pm 0.084$  and  $\lambda = 2.251 \pm 0.071$ .

In the figure below, apart from the least-squares straight line  $y = 0.971 + 2.251x$ , also given are the straight lines passing through the central point of the measurements ( $\bar{x} = 1.00$ ,  $\bar{y} = 3.22$ ) and having slopes  $\lambda = 2.251 \pm 0.071$ .



The equation  $y = 0.9709 + 2.2509x$  found, gives the value of  $y$  for every value of  $x$ .

The error in  $y$  is given by Eq. (11.29) as

$$\delta y = \frac{\sigma_y}{\sqrt{N}} \sqrt{1 + \frac{N^2}{N[x^2] - [x]^2} (x - \bar{x})^2} = 0.0452 \sqrt{1 + 2.50 \times (x - 1)^2}.$$

The variation of  $\delta y$  with  $x$  is shown in the figure. The error in  $y$  is minimum and equal to  $\delta y = 0.045$  for  $x = 1$ . For  $x = 0$  and for  $x = 2$ , the error is  $\delta y = 0.085$ .

From the relation  $y = 0.97 + 2.25x$  we find that for  $x = 1.5$  it is  $y = 4.35$ . The error in  $y$  is given by the equation  $\delta y = 0.0452 \sqrt{1 + 2.50 \times (x - 1)^2}$  as  $\delta y = 0.06$ . Therefore, for  $x = 1.5$  it is  $y = 4.35 \pm 0.06$ .

### Example 11.2 [E]

Solve the problem of Example 11.1 using Excel<sup>®</sup>.

We place the data of columns  $x_i$  and  $y_i$  in cells A1-A11 and B1-B11, respectively. We highlight these cells by left-clicking on cell A1 and then, holding the **SHIFT** key down, we draw the cursor down to cell B11. In **Insert, Charts** we select **Scatter**. A scatter chart is created, with the points  $(x_i, y_i)$ .

We press the  $\boxplus$  key at the top right corner of the plot's frame, opening **Chart Elements**. We select

**Trendline, More Options** and tick **Linear** and **Display equation on chart**. The result is  $y = 2.2509x + 0.9709$ . No errors are available for the coefficients.

Substituting  $x = 1.5$  in the equation of the line, we obtain the result 4.34727. This is  $y(1.5)$ .

### Example 11.3 [O]

Solve the problem of Example 11.1 using Origin<sup>®</sup>.

We place the data of columns  $x_i$  and  $y_i$  in columns A and B, respectively. We highlight both columns by left-clicking on the label of column A and then, holding the **Shift** key down, left-clicking on the label of column B. Then

**Analysis > Fitting > Linear Fit > Open Dialog . . .**

In the window that opens, we press **OK**. The program returns the results

**Intercept**(= $a$ ) =  $0.97091 \pm 0.08446$  and **Slope**(= $\lambda$ ) =  $2.25091 \pm 0.07138$ .

These are the results found in Example 11.1.

A graph such as the one shown in Example 11.1 is also given.

If we also want to find  $y$  for a given  $x$ , in the dialog box that opens we have to select **Find X/Y** and then tick **Find Y from X**. With the results, there is a page titled **FitLinearFindYfromX1**. We go to this page and, in a cell of the column labeled **Enter X values**: we enter the value of  $x = 1.5$ . In the adjacent column, labeled **Y value**, the result 4.34727 appears. This is  $y(1.5)$ .

The errors in the values of  $y$  may also be taken into account in the fitting. The errors should be entered in a third column which is also selected in the analysis.

### Example 11.4 [P]

Solve the problem of Example 11.1 using Python.

```

from __future__ import division
import math
import numpy as np
import matplotlib.pyplot as plt

# Enter the values of x, y:

x = np.array([0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2])
y = np.array
([0.92, 1.48, 1.96, 2.27, 2.61, 3.18, 3.8, 4.01, 4.85, 5.10, 5.26])

# Plot, size of dots, labels, ranges of values, initial values

plt.scatter(x, y)
plt.xlabel("x")

# set the x-axis label

plt.ylabel("y")

# set the y axis label

plt.grid(True)

# Evaluation

N = len(x)
X = sum(x)
XX = sum(x**2)
Y = sum(y)
XY = sum(x*y)

```

```

DENOM = N*XX-X**2
DETA = Y*XX-X*XY
DETL = N*XY-X*Y

a = DETA/DENOM
lambda = DETL/DENOM

d = y - a - lambda*x

DD = sum(d**2)

Da = math.sqrt((DD*XX)/((N-2)*DENOM))
Dlambda = math.sqrt((N*DD)/((N-2)*DENOM))

# Results

print("Value of a:", a)
print("Value of lambda:", lambda)
print("Standard error in a:", Da)
print("Standard error in λ:", Dlambda)

# Plot least-squares line

xx = np.linspace(min(x), max(x), 200)
yy = a + b * xx
plt.plot(xx, yy, '-')

plt.show()

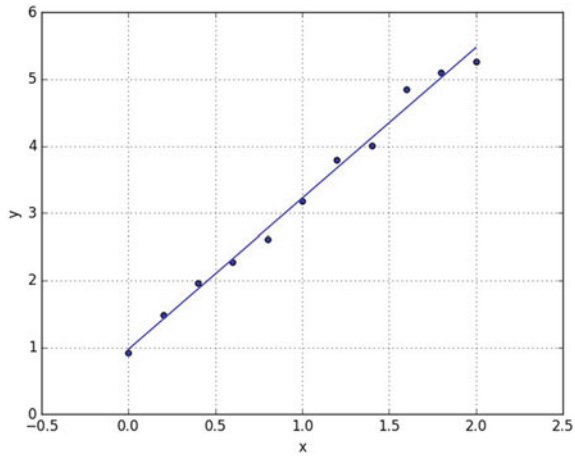
```

The plot shown below is produced.  
The values of the parameters are:

```

Value of a: 0.970909090909
Value of λ: 2.25090909091
Standard error in a: 0.08445667109784327
Standard error in λ: 0.07137891491854917

```



From the relation  $y = 0.97 + 2.25x$  we find that for  $x = 1.5$  it is  $y = 4.35$ .

**Example 11.5 [R]**

Solve the problem of Example 11.1 using R.

$i$	$x_i$	$y_i$
1	0.0	0.92
2	0.2	1.48
3	0.4	1.96
4	0.6	2.27
5	0.8	2.61
6	1.0	3.18
7	1.2	3.80
8	1.4	4.01
9	1.6	4.85
10	1.8	5.10
11	2.0	5.26

Define vectors  $x$  and  $y$ :

```
> x <- c(0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2)
> y <- c(0.92, 1.48, 1.96, 2.27, 2.61, 3.18, 3.80, 4.01, 4.85, 5.10, 5.26)
```

Plot  $y(x)$ :

```
> plot(x, y, pch=20, cex=0.5, xlab="x", ylab="y", xlim=c(0, 2), ylim=c(0, 6))
```

Find least-squares best-fit straight line:

```
> fit <- lm(y~x)
> fit
```

Read intercept and slope of line:

Call:

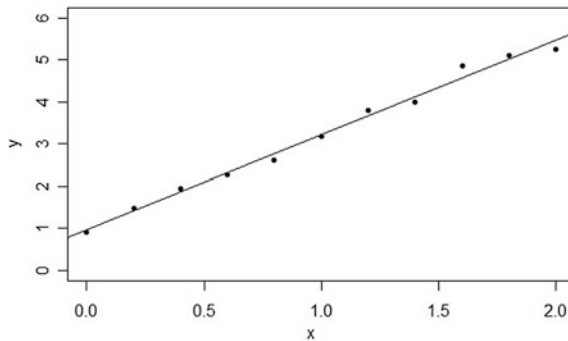
```
> lm(formula = y ~ x)
```

Coefficients:

```
(Intercept)      x
0.9709         2.2509
```

Plot least-squares best-fit straight line:

```
> abline(fit)
```



The equation of the line is  $y = 0.9709 + 2.2509x$ . For  $x = 1.5$ , it is  $y = 0.9709 + 2.2509 \times 1.5 = 4.347$ .

### Example 11.6

In this example we will demonstrate the role of the errors in the magnitudes evaluated by the method of least squares. For this reason, the experimental points were chosen to have a high dispersion, corresponding to large measurement errors.

Using the method of least squares, fit a straight line to the points:

$i$	1	2	3	4	5	6	7
$x_i$	1	2	4	6	8	9	10
$y_i$	0.2	0.8	0.4	1	0.7	1.2	0.8

Find the value of  $y$  for  $x = 5$ .

It is:  $n = 7$ ,  $\bar{x} = 5.714$  and  $\bar{y} = 0.7286$ .



$$[x] = 40 \quad [y] = 5.1 \quad [xy] = 33.8 \quad [x^2] = 302 \quad [d^2] = 0.399$$

$$a = 0.3661 \quad \lambda = 0.0634$$

and the required straight line is  $y = 0.366 + 0.0634x$ .

Also,  $\sigma_y = \sqrt{\frac{0.399}{5}} = 0.282 \quad \delta a = 0.166 \quad \delta \lambda = 0.025$ .

The errors in the  $y$  values are given by the relation

$$\delta y = \frac{\sigma_y}{\sqrt{N}} \sqrt{1 + \frac{N^2}{N[x^2] - [x]^2} (x - \bar{x})^2} = 0.1068 \sqrt{1 + 0.0953(x - 5.714)^2}$$

or  $\delta y = \sqrt{0.0469 - 0.01242x + 0.001087x^2}$

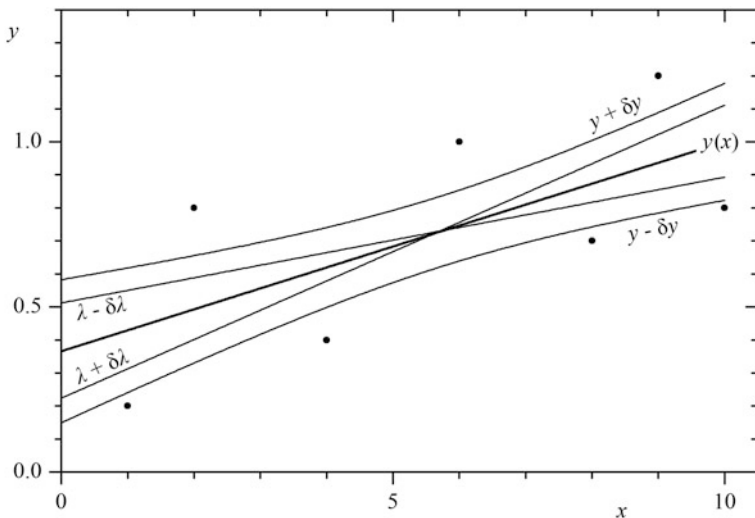
From the equation  $y = 0.366 + 0.0634x$  we find that for  $x = 5$  it is  $y = 0.68$ . The error in  $y$  is  $\delta y = 0.11$ . Therefore, for  $x = 5$  it is  $y = 0.68 \pm 0.11$ .

In the figure below were drawn:

1. The experimental points and the least-squares straight line  $y(x)$ .
2. The straight lines passing through the center  $K: (5.71, 0.73)$  of the line  $y(x)$  and having slopes  $\lambda \pm \delta \lambda$ , i.e.  $0.0634 \pm 0.0253$ . The equations of these straight lines are  $y_1 = 0.512 + 0.0381x$  and  $y_2 = 0.224 + 0.0887x$ .
3. The curves  $y(x) \pm \delta y$  or

$$y = 0.366 + 0.0634x \pm \sqrt{0.0469 - 0.01242x + 0.001087x^2}$$

which mark off the region of  $y$  values lying between  $y \pm \delta y$ .



**Example 11.7 [E]**

Using Excel<sup>®</sup> and the method of least squares, fit a straight line to the points:

$i$	1	2	3	4	5	6	7
$x_i$	1	2	4	6	8	9	10
$y_i$	0.2	0.8	0.4	1	0.7	1.2	0.8

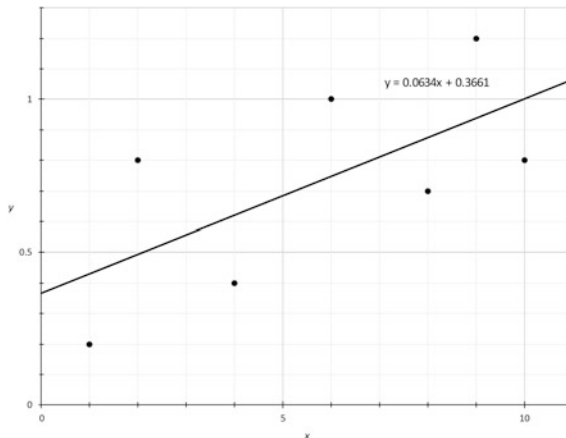
Find the value of  $y$  for  $x = 5$ .

We place the data  $x_i$  and  $y_i$  in columns A and B, respectively. We highlight columns A and B and open the **Insert** window. We open the **Recommended Charts** window and choose **Scatter**.

A scatter chart is produced. We double-click on a point and in the **Format Data Series** window that opens we select **Series Options > Marker Options** for **Fill** we select **Solid Fill** and color **Black**. Also, in **Border** we select **Solid Line**, color **Black**, **Width 0.75 pt**, **Dash Type** continuous line.

We double-click on the straight line and select **Line, Solid Line**, color **Black**, **Width 1.5 pt**, **Dash Type** continuous line. In **Trendline Options** we choose **Linear, Forecast, Forward 1.0 period, Backward 1.0 period**. We also tick the box **Display Equation on Chart**.

The graph produced is shown below.



The coefficients of the equation of line are:

$$a = 0.36615 \pm 0.21651 \quad \text{and} \quad \lambda = 0.06342 \pm 0.03296$$

The equation of the line is  $y = 0.36615 + 0.06342x$ .

For  $x = 5$  it is  $y(5) = 0.6833$ .

**Example 11.8 [O]**

Using Origin<sup>®</sup> and the method of least squares, fit a straight line to the points:

<i>i</i>	1	2	3	4	5	6	7
<i>x<sub>i</sub></i>	1	2	4	6	8	9	10
<i>y<sub>i</sub></i>	0.2	0.8	0.4	1	0.7	1.2	0.8

Find the value of *y* for *x* = 5.

We place the data *x<sub>i</sub>* and *y<sub>i</sub>* in columns A and B, respectively. We highlight columns A and B and open the **Plot** window. We select **Symbol** and then **Scatter**. A scatter plot is produced.

In the Graph1 window, and then

**Analysis > Fitting > Linear Fit > Open Dialog . . .**

In the **Linear Fit** window we select **Fitted Curves Plot** and set **X Data Type, Margin [%]** to 10. This will ensure that the straight line fitted, when plotted in the graph, will extend by 10% to the left and the right of the experimental points at the two ends. We also open the **Find X/Y** window and tick the **Find Y from X** box. Pressing **OK** produces the best fit straight line in the graph.

The program also returns the results:

**Intercept** (= *a*) = 0.36615 ± 0.21651 and **Slope** (= *λ*) = 0.06342 ± 0.03296.

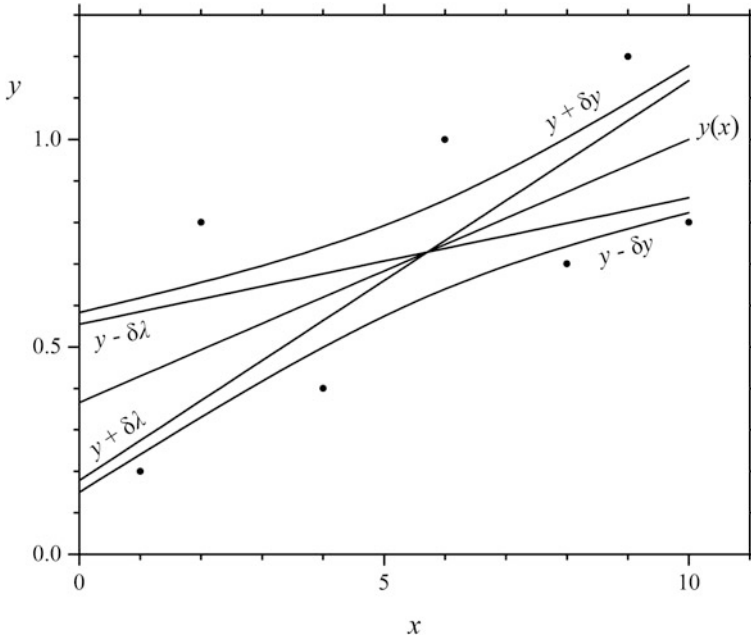
In **Book1** page **FitLinearFindYFromX1**, by typing *x* = 5 in the *x* column, we get *y*(5) = 0.6833.

The equations of the two straight lines passing through the center of the points K:(5.714, 0.7286) and having slopes *λ* ± δ*λ*, i.e. 0.06342 ± 0.03296, are

$$y_1 = 0.5546 + 0.03046x \text{ and } y_2 = 0.1779 + 0.09638x.$$

In column D we enter the values of *x* from 0 to 10 in steps of 1/3. We highlight column D and open the **Column** window, where we **Set As X** column D. For these values of *x*, we evaluate *y* in column E, *y<sub>1</sub>* in column F and *y<sub>2</sub>* in column G.

Using the expression for the error δ*y* in *y* found in Example 11.5, we evaluate δ*y* in column I and the values of *y* - δ*y* and *y* + δ*y* in columns J and K respectively. We highlight columns D, E, F, G, J and K. We open the **Plot** window and select **Line > Line**. The plot shown below is produced. The experimental points were added to this plot by right clicking on the number (1) appearing in the top left corner, selecting **Layer Contents** and including column B in the contents of the graph shown on the right. This is done by selecting **B[Y1]** from the table in the left and using the arrow to include it in the table on the right. The final result is shown in the figure below.



**Example 11.9 [P]**

Using Python and the method of least squares, fit a straight line to the points:

$i$	1	2	3	4	5	6	7
$x_i$	1	2	4	6	8	9	10
$y_i$	0.2	0.8	0.4	1	0.7	1.2	0.8

Find the value of  $y$  for  $x = 5$ .

# Enter the values of  $x$ ,  $y$ :

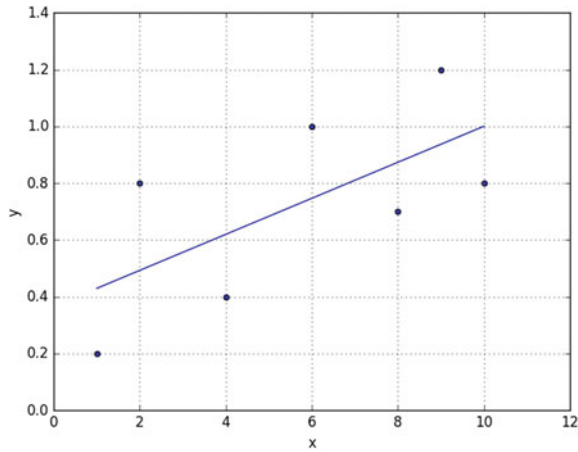
```
x = np.array([1, 2, 4, 6, 8, 9, 10])
y = np.array([0.2, 0.8, 0.4, 1, 0.7, 1.2, 0.8])
```

The rest of the program is identical to that of Example 11.4

The plot shown below is produced.

The values of the parameters are:

```
Value of a: 0.366147859922
Value of lambda: 0.0634241245136
Standard error in a: 0.21650834831061155
Standard error in lambda: 0.03296250399459645
```



From the relation  $y = 0.366 + 0.0634 x$  we find that for  $x = 5$  it is  $y = 0.683$ .

**Example 11.10 [R]**

Using R and the method of least squares, fit a straight line to the points:

$i$	1	2	3	4	5	6	7
$x_i$	1	2	4	6	8	9	10
$y_i$	0.2	0.8	0.4	1	0.7	1.2	0.8

Find the value of  $y$  for  $x = 5$ .

We form the vectors  $x$  and  $y$ . We plot the scatter plot  $y(x)$ .

```
> x <- c(1, 2, 4, 6, 8, 9, 10)
> y <- c(0.2, 0.8, 0.4, 1, 0.7, 1.2, 0.8)
> plot(x, y, pch=20, xlab="x", ylab="y", xlim=c(0, 10), ylim=c(0, 1))
```

We fit a least-squares straight line to the data:

```
> fit <- lm(y~x)
> fit
```

Call:

```
lm(formula = y ~ x)
```

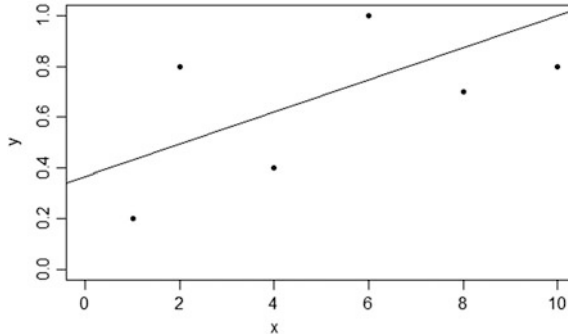
Coefficients:

```
(Intercept)      x
  0.36615      0.06342
```

We plot the straight line:

```
> abline(fit)
```

The equation of the line is  $y = 0.36615 + 0.06342x$ . For  $x = 5$  it is  $y(5) = 0.6833$ .



### 11.3.1.1 Least Squares Using Weighted Measurements

If the results of the measurements are weighted, with weight  $w_i$  for the measurement  $(x_i, y_i)$ , then the results are modified as follows: The normal equations are

$$\alpha[w] + \lambda[w x] = [w y] \quad \text{and} \quad \alpha[w x] + \lambda[w x^2] = [w x y] \quad (11.41)$$

from which it follows that

$$\alpha = \frac{[w y][w x^2] - [w x][w x y]}{[w][w x^2] - [w x]^2} \quad (11.42)$$

$$\lambda = \frac{[w][w x y] - [w x][w y]}{[w][w x^2] - [w x]^2} \quad (11.43)$$

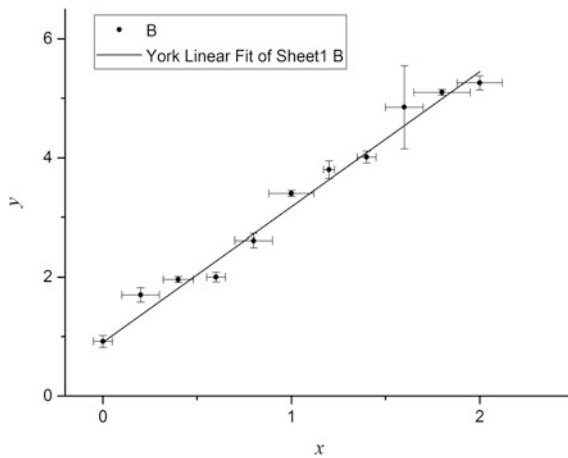
$$\delta \alpha = \sigma_\alpha = \sqrt{\frac{[w d^2]}{(N-2)} \frac{[w x^2]}{[w][w x^2] - [w x]^2}} \quad (11.44)$$

$$\delta \lambda = \sigma_\lambda = \sqrt{\frac{[w d^2]}{(N-2)} \frac{[w]}{[w][w x^2] - [w x]^2}} \quad (11.45)$$

**Example 11.11 [O]**

Fit a straight line to the set of values  $(x_i, y_i)$  with their errors given in the table below, applying the method of least squares and taking into account the errors.

	col(A)	col(B)	col(C)	col(D)
$i$	$x_i$	$y_i$	$\delta x_i$	$\delta y_i$
1	0	0.92	0.05	0.1
2	0.2	1.7	0.1	0.12
3	0.4	1.96	0.08	0.05
4	0.6	2	0.05	0.08
5	0.8	2.61	0.1	0.12
6	1	3.4	0.12	0.05
7	1.2	3.8	0.03	0.15
8	1.4	4.01	0.05	0.1
9	1.6	4.85	0.1	0.7
10	1.8	5.1	0.15	0.05
11	2	5.26	0.12	0.12



We place the data of columns  $x_i$ ,  $y_i$  and their errors  $\delta x_i$  and  $\delta y_i$ , in columns A, B, C and D, respectively. We highlight column A. Then

**Analysis > Fitting > Fit Linear with Errors > Open Dialog...**

In the window that opens, we open **Input, Input Data**. In **Range 1** we enter for **X** column **A(X)**, for **Y** column **B(Y)**, for **Y Error** column **D(Y)** and for **X Error** column **C(Y)**. Then press **OK**. The program returns the results

**Intercept**(= $a$ ) =  $0.90289 \pm 0.10616$  and **Slope**(= $\lambda$ ) =  $2.27214 \pm 0.10816$ .

The graph shown on the previous page is also produced.

The least squares method used is that of York, which uses weights for each point, based on the errors  $\delta x_i$  and  $\delta y_i$  of the measurements.

### Example 11.12 [P]

Using Python, fit a least-squares straight line to the set of values  $(x_i, y_i)$  of Example 11.11 [O], taking as weights of the points the inverses of the squares of the errors  $\delta y_i$ .

The weights will be taken to be  $w_i = 1/(\delta y_i)^2$ . The weight vector will therefore be:  
 $w = \text{np.array}([100, 69.4, 400, 156.3, 69.4, 400, 44.4, 100, 2, 400, 69.4])$

# Program:

```
from __future__ import division
import math
import numpy as np
import matplotlib.pyplot as plt

# Enter the values of x, y and their corresponding weights w:
x = np.array([0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2])
y = np.array
([0.92, 1.48, 1.96, 2.27, 2.61, 3.18, 3.8, 4.01, 4.85, 5.1, 5.26])
w = np.array([100, 69.4, 400, 156.3, 69.4, 400, 44.4, 100, 2, 400, 69.4])

# Plot, size of dots, labels, ranges of values, initial values

plt.scatter(x, y)
plt.xlabel("x") # set the x-axis label
plt.ylabel("y") # set the y-axis label
plt.grid(True)

# Evaluation
N = len(x)

W = sum(w)
WX = sum(w*x)
WXX = sum(w*x**2)
```



```

WY = sum(w*y)
WXY = sum(w*x*y)

DENOM = W*WXX - (WX)**2
DETA = WY*WXX - WX*WXY
DETL = W*WXY - WX*WY

a = DETA/DENOM
lambda = DETL/DENOM

d = y - a - lambda*x

WDD = sum(w*d**2)

Da = math.sqrt((WDD*WXX) / ((N-2)*DENOM))
Dlambda = math.sqrt((WDD*W) / ((N-2)*DENOM))

# Results

print("Value of a:", a)
print("Value of b:", lambda)
print("Standard error in a:", Da)
print("Standard error in b:", Dlambda)

# Plot least-squares line

xx = np.linspace(min(x), max(x), 200)
yy = a + lambda * xx
plt.plot(xx, yy, '-')

plt.show()

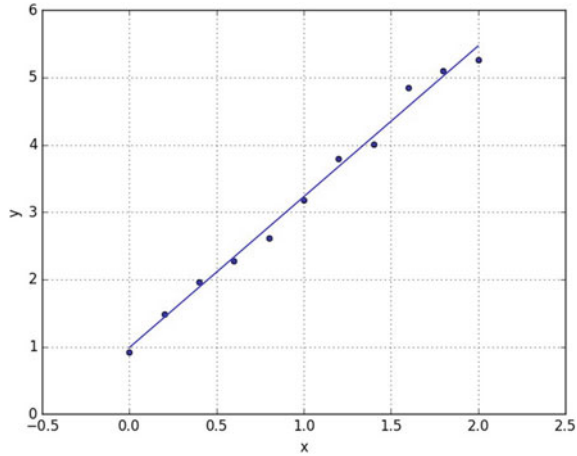
```

The plot shown in the next page is produced.  
The numerical values of the parameters are:

```

Value of a: 0.986465071722
Value of λ: 2.24054597889
Standard error in a: 0.05624562965433959
Standard error in λ: 0.04879527127536099

```



**Example 11.13 [R]**

Fit a straight line to a given set of values  $(x_i, y_i)$ , applying the method of least squares and taking into account the errors in  $y$ .

$i$	$x_i$	$y_i$	$\delta y_i$
1	0	0.92	0.1
2	0.2	1.7	0.12
3	0.4	1.96	0.05
4	0.6	2	0.08
5	0.8	2.61	0.12
6	1	3.4	0.05
7	1.2	3.8	0.15
8	1.4	4.01	0.1
9	1.6	4.85	0.7
10	1.8	5.1	0.05
11	2	5.26	0.12

This is the same problem as in Example 11.9 but taking into account only the errors in  $y$ .

```
# We form the vectors for x, y and the errors in y:
x = c(0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2)
y = c(0.92, 1.7, 1.96, 2, 2.61, 3.4, 3.8, 4.01, 4.85, 5.1, 5.26)
erry = c(0.1, 0.12, 0.05, 0.08, 0.12, 0.05, 0.15, 0.1, 0.7, 0.05, 0.12)
```

```
# The weights are taken to be  $w_i = 1/(\delta_{y_i})^2$ :
weights = (1/erry^2)
weights
[1] 100.000000 69.444444 400.000000 156.250000 69.444444 400.000000
44.444444 100.000000 2.040816 400.000000
[11] 69.444444

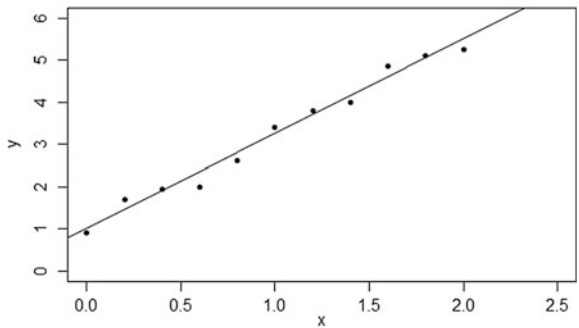
# The scatter plot of the points is created:
plot(x, y, pch=20, xlab="x", ylab="y", xlim=c(0, 2.5), ylim=c(0, 6))
fit <- lm(y~x, weights = weights)
fit

Call:
lm(formula = y ~ x, weights = weights)

Coefficients:
(Intercept)      x
      1.012      2.249

# The best straight line is drawn:
abline(fit)
```

The equation of the straight line found is:  $y = 1.012 + 2.249x$ .



### 11.3.2 Polynomial

In many cases, the relationship between  $x$  and  $y$  is not linear. In general, the relationship may be thought of as being expressed by a polynomial of the form [1]:

$$y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n. \tag{11.46}$$

The determination of the  $n + 1$  unknown coefficients is achieved by the minimization, with respect to these coefficients, of the quantity

$$S \equiv \sum_{i=1}^N (y_i - a_0 - a_1x_i - a_2x_i^2 - \dots - a_nx_i^n)^2, \tag{11.47}$$

where  $(x_i, y_i) (i = 1, 2, \dots, N)$  are the results of the  $N$  measurements we have performed. In general, it must be  $n < N - 1$ .

Differentiating Eq. (11.47) with respect to the coefficients  $a_k$  and equating to zero, we have the normal equations:

$$\begin{aligned} a_0N + a_1[x] + a_2[x^2] + \dots + a_n[x^n] &= [y] \\ a_0[x] + a_1[x^2] + a_2[x^3] + \dots + a_n[x^{n+1}] &= [xy] \\ a_0[x^2] + a_1[x^3] + a_2[x^4] + \dots + a_n[x^{n+2}] &= [x^2y] \\ \dots\dots\dots & \\ a_0[x^n] + a_1[x^{n+1}] + a_2[x^{n+2}] + \dots + a_n[x^{2n}] &= [x^ny] \end{aligned} \tag{11.48}$$

From these equations, the coefficients  $a_k$  may be found.

**11.3.2.1 Parabola**

For the case

$$y = a_0 + a_1x + a_2x^2 \tag{11.49}$$

we have the normal equations

$$\begin{aligned} a_0N + a_1[x] + a_2[x^2] &= [y] \\ a_0[x] + a_1[x^2] + a_2[x^3] &= [xy] \\ a_0[x^2] + a_1[x^3] + a_2[x^4] &= [x^2y] \end{aligned} \tag{11.50}$$

Applying Cramer’s rule, we have for  $a_0, a_1$  and  $a_2$ :

$$\begin{aligned} a_0 &= \frac{\begin{vmatrix} [y] & [x] & [x^2] \\ [xy] & [x^2] & [x^3] \\ [x^2y] & [x^3] & [x^4] \end{vmatrix}}{\begin{vmatrix} N & [x] & [x^2] \\ [x] & [x^2] & [x^3] \\ [x^2] & [x^3] & [x^4] \end{vmatrix}} = \frac{a_1}{\begin{vmatrix} N & [x] & [y] \\ [x] & [x^2] & [xy] \\ [x^2] & [x^3] & [x^2y] \end{vmatrix}} = \frac{1}{\begin{vmatrix} N & [x] & [x^2] \\ [x] & [x^2] & [x^3] \\ [x^2] & [x^3] & [x^4] \end{vmatrix}}. \end{aligned} \tag{11.51}$$

The errors in the coefficients  $\delta a_0, \delta a_1$  and  $\delta a_2$  are given by the relations:

$$\frac{(\delta a_0)^2}{\begin{vmatrix} [x^2] & [x^3] \\ [x^3] & [x^4] \end{vmatrix}} = \frac{(\delta a_1)^2}{\begin{vmatrix} N & [x^2] \\ [x^2] & [x^4] \end{vmatrix}} = \frac{(\delta a_2)^2}{\begin{vmatrix} N & [x] \\ [x] & [x^2] \end{vmatrix}} = \frac{\sigma_y^2}{\begin{vmatrix} N & [x] & [x^2] \\ [x] & [x^2] & [x^3] \\ [x^2] & [x^3] & [x^4] \end{vmatrix}}, \quad (11.52)$$

where

$$\sigma_y^2 = \frac{[d^2]}{N - 3}, \quad [d^2] \equiv \sum_{i=1}^N d_i^2 = \sum_{i=1}^N (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2. \quad (11.53)$$

**Example 11.14**

Using the method of least squares, fit a parabolic curve to the measurements given in the first three columns of the table below.

<i>i</i>	<i>t</i> (s)	<i>y</i> (m)	<i>t</i> <sup>2</sup> (s <sup>2</sup> )	<i>t</i> <sup>3</sup> (s <sup>3</sup> )	<i>t</i> <sup>4</sup> (s <sup>4</sup> )	<i>ty</i> (s m)	<i>t</i> <sup>2</sup> <i>y</i> (s <sup>2</sup> m)	<i>y</i> <sub>th</sub> (m)	<i>d</i> (m)	<i>d</i> <sup>2</sup> (m <sup>2</sup> )
1	0	1	0	0	0	0	0	2.32	1.32	1.74
2	1	8	1	1	1	8	8	6.28	-1.72	2.96
3	2	20	4	8	16	40	80	19.77	-0.23	0.05
4	3	45	9	27	81	135	405	42.80	-2.20	4.84
5	4	70	16	64	256	280	1120	75.36	5.36	28.73
6	5	120	25	125	625	600	3000	117.46	-2.54	6.45
<i>N</i> = 6	15 = [ <i>t</i> ]	264 = [ <i>y</i> ]	55 = [ <i>t</i> <sup>2</sup> ]	225 = [ <i>t</i> <sup>3</sup> ]	979 = [ <i>t</i> <sup>4</sup> ]	1063 = [ <i>ty</i> ]	4613 = [ <i>t</i> <sup>2</sup> <i>y</i> ]			44.78 = [ <i>d</i> <sup>2</sup> ]

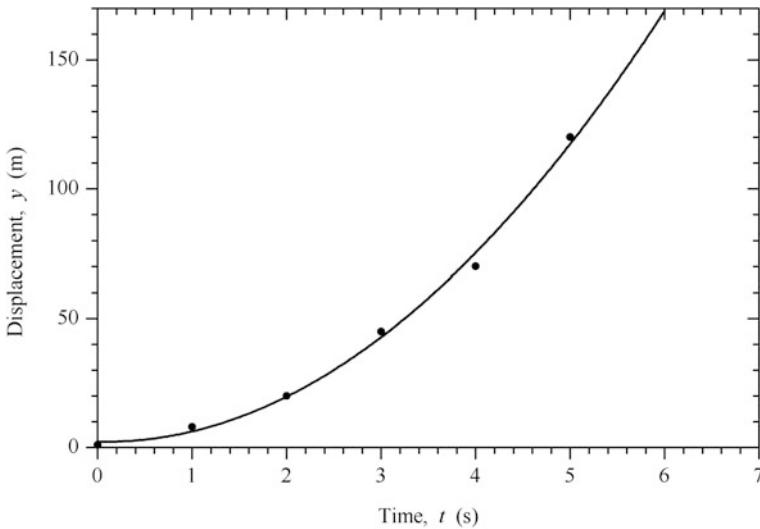
We will fit the curve  $y = a_0 + a_1 t + a_2 t^2$  to the experimental results. From Eq. (11.51), in S.I. units,

$$\begin{vmatrix} a_0 & a_1 & a_2 & 1 \\ 264 & 15 & 55 & 6 \\ 1063 & 55 & 225 & 15 \\ 4613 & 225 & 979 & 55 \end{vmatrix} = \begin{vmatrix} a_1 & a_2 & 1 \\ 6 & 264 & 55 \\ 15 & 1063 & 225 \\ 55 & 4613 & 979 \end{vmatrix} = \begin{vmatrix} a_2 & 1 \\ 6 & 15 \\ 15 & 55 \\ 55 & 225 \end{vmatrix} = \begin{vmatrix} 1 \\ 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{vmatrix}$$

we find  $\frac{a_0}{9100} = \frac{a_1}{-3178} = \frac{a_2}{18690} = \frac{1}{3920}$

and  $a_0 = 2.32 \text{ m}, a_1 = -0.811 \text{ m/s}, a_2 = 4.768 \text{ m/s}^2$ .

The curve is  $y = 2.32 - 0.811t + 4.768t^2$  (in m, when the time  $t$  is expressed in s). The experimental points and the least-squares curve are shown in the figure below.



The errors in the parameters are found using Eqs. (11.52) and (11.53)

$$\frac{(\delta a_0)^2}{\begin{vmatrix} 55 & 225 \\ 225 & 979 \end{vmatrix}} = \frac{(\delta a_1)^2}{\begin{vmatrix} 6 & 55 \\ 55 & 979 \end{vmatrix}} = \frac{(\delta a_2)^2}{\begin{vmatrix} 6 & 15 \\ 15 & 55 \end{vmatrix}} = \frac{\sigma_y^2}{\begin{vmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{vmatrix}}$$

where  $[d^2] \equiv \sum_{i=1}^N d_i^2 = \sum_{i=1}^N (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2 = 44.78 \text{ m}^2$  and

$$\sigma_y^2 = \frac{[d^2]}{N-3} = \frac{44.78}{3} = 14.9 \text{ m}^2 \text{ or } \sigma_y = 3.86 \text{ m.}$$

Therefore,  $\frac{(\delta a_0)^2}{3220} = \frac{(\delta a_1)^2}{2849} = \frac{(\delta a_2)^2}{105} = \frac{14.9}{3920} = 0.00380$  and

$$\delta a_0 = 3.5 \text{ m, } \delta a_1 = 3.3 \text{ m/s, } \delta a_2 = 0.63 \text{ m/s}^2$$

or

$$a_0 = 2.3 \pm 3.5 \text{ m, } a_1 = -0.8 \pm 3.3 \text{ m/s, } a_2 = 4.77 \pm 0.63 \text{ m/s}^2.$$

We notice that the presence of points at large values of  $t$  makes the fractional errors in  $a_0$  and  $a_1$  large, since  $a_0$  and  $a_1$  are important at low values of  $t$ . Of course, we must not forget that the values of the parameters we found depend on each other. If, in other words, we suppose a different value for one of the parameters, the optimum values of the other two will have to be modified.

**Example 11.15 [E]**

Using Excel<sup>®</sup>, fit a parabola to the data of Example 11.14.

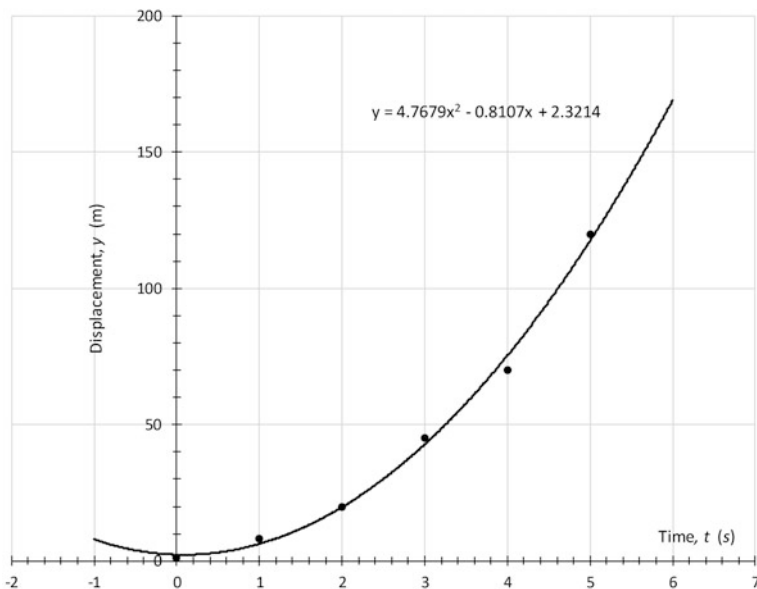
We enter the  $t$  and  $y$  values in columns A and B respectively. We highlight columns A and B. Opening the **Insert** window, we select **Scatter Plot**. We double-click on a point and change the color of the points to black and their size to 0.75 pt.

Pressing the  $\boxplus$  key at the top right hand corner of the graph, we click on **Trendline, More Options**. In **Format Trendline, Trendline Options** we select **Polynomial, Order 2**. We also select **Forecast, Forward 1 period Backward 1 period** and **Display Equation on chart**. We delete the straight line present in the graph.

The graph of the best fit parabola is produced, which, suitably formatted, looks like the figure shown here.

The equation of the parabola is found to be:

$$y = 2.3214 - 0.8107t + 4.7679t^2.$$



**Example 11.16 [O]**

Using Origin<sup>®</sup>, fit a parabola to the data of Example 11.14.

We place the data of columns  $t$  and  $y$  in columns A and B, respectively. We highlight both columns by left-clicking on the label of column A and then, holding the **Shift** key down, left-clicking on the label of column B. Then

**Analysis > Fitting > Polynomial Fit > Open Dialog...**

In the window that opens, we select: **Input Data, Range 1, X** select column A, **Y** select column B, **Polynomial Order, 2**. Press **Fit**. The program fits the parabola  $y = A + Bx + Cx^2$  to the experimental results, where  $x = t$ . It is given that

$$A(= a_0) = 2.32143 \pm 3.50266, B(= a_1) = -0.81071 \pm 3.2947 \text{ and} \\ C(= a_2) = 4.76786 \pm 0.63251.$$

The equation of the parabola is:

$$y = 2.3214 - 0.8107t + 4.7679t^2$$

These results agree with those of Example 11.3.

**Example 11.17 [P]**

Using Python, fit a parabola to the data of Example 11.14.

```
import math
from __future__ import division
import numpy as np
import matplotlib.pyplot as plt

# Enter the values of x and the corresponding y:

x = np.array([0, 1, 2, 3, 4, 5])
y = np.array([1, 8, 20, 45, 70, 120])

# Plot, size of dots, labels, ranges of values, initial values

plt.scatter(x, y)
plt.xlabel("x, (m) ") # set the x-axis label
plt.ylabel("Displacement, y (m) ") # set the y-axis label
plt.grid(True)
```



```

# Evaluation
N = len(x)

X = sum(x)
XX = sum(x**2)
XXX = sum(x**3)
XXXX = sum(x**4)
Y = sum(y)
XY = sum(x*y)
XXY = sum(x**2*y)

DENOM = N*(XX*XXXX-XXX*XXX) - X*(X*XXXX-XX*XXX) + XX*(X*XXX-XX*XX)
DET0 = Y*(XX*XXXX-XXX*XXX) - X*(XY*XXXX-XXX*XXY) + XX*(XY*XXX-XX*XXY)
DET1 = N*(XY*XXXX-XXX*XXY) - Y*(X*XXXX-XX*XXX) + XX*(X*XXY-XX*XY)
DET2 = N*(XX*XXY-XXX*XY) - X*(X*XXY-XX*XY) + Y*(X*XXX-XX*XX)

a0 = DET0/DENOM
a1 = DET1/DENOM
a2 = DET2/DENOM

d = y - a0 - a1*x - a2*x**2
S = math.sqrt(sum(d**2)/(N-3))

Da0 = S*math.sqrt(abs((XX*XXXX-XXX*XXX)/DENOM))
Da1 = S*math.sqrt(abs((N*XXXX-XX*XX)/DENOM))
Da2 = S*math.sqrt(abs((N*XX-X*X)/DENOM))

# Results

print("Value of a0:", a0)
print("Value of a1:", a1)
print("Value of a2:", a2)
print("Standard error in a0:", Da0)
print("Standard error in a1:", Da1)
print("Standard error in a2:", Da2)

# Plot least-squares line

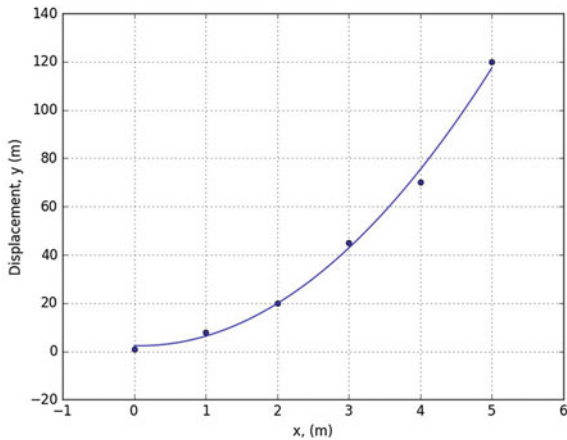
xx = np.linspace(min(x), max(x), 200)
yy = a0 + a1*xx + a2*xx**2
plt.plot(xx, yy, '-')

plt.show()

```

The plot shown is produced.  
The values of the parameters are:

Value of  $a_0$ : 2.32142857143  
 Value of  $a_1$ : -0.810714285714  
 Value of  $a_2$ : 4.76785714286  
 Standard error in  $a_0$ : 3.5026593395561  
 Standard error in  $a_1$ : 3.2947023804146847  
 Standard error in  $a_2$ : 0.632505948991947



### Example 11.18 [R]

Using R, fit a parabola to the data of Example 11.14.

```
# The data vectors
tdata = c(0, 1, 2, 3, 4, 5)
ydata = c(1, 8, 20, 45, 70, 120)

# Plot, size of dots, labels, ranges of values, initial values
plot(tdata, ydata, pch=20, xlab="Time, t (s)", ylab="Displacement, y (m)",
      xlim=c(0, 6), ylim=c(0, 150))

# Fit least-squares line

A=2
B=-10
C=5
fit = nls(ydata~A+B*tdata+C*tdata^2, start=list(A=A, B=B, C=C))
```

```
summary(fit)
Formula: ydata ~ A + B * tdata + C * tdata^2
Parameters:
  Estimate Std. Error t value Pr(>|t|)
A  2.3214   3.5027   0.663  0.55486
B -0.8107   3.2947  -0.246  0.82151
C  4.7679   0.6325   7.538  0.00484 **
Residual standard error: 3.865 on 3 degrees of freedom
Number of iterations to convergence: 1
Achieved convergence tolerance: 1.47e-07

# Plot least-squares line

new = data.frame(tdata = seq(min(tdata),max(tdata), len=200))
lines(new$tdata, predict(fit, newdata=new))

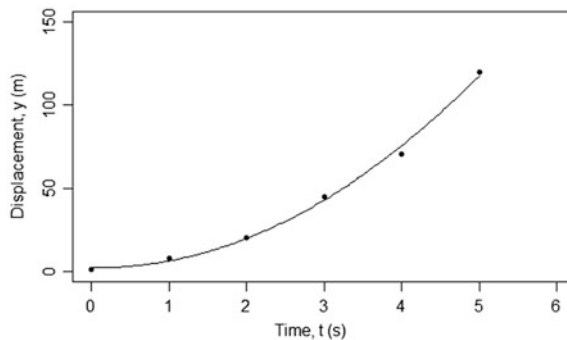
# Sum of squared residuals

sum(resid(fit)^2)
[1] 44.80714

# Parameter confidence intervals

confint(fit)

      2.5%   97.5%
A -8.825597 13.468454
B -11.295928  9.674499
C  2.754941  6.780773
```



## 11.3.2.1.1 Errors in the Values Read from a Least-Squares Parabola

It is not clear how the errors in  $a_0$ ,  $a_1$  and  $a_2$  combine to give the error  $\delta y$  in  $y$  for a given  $x$ . To get a *reasonable estimate* of  $\delta y$ , we work as follows.

We assume that  $a_0$  is known and define the new variable

$$Y = \frac{y - a_0}{x}. \quad (11.54)$$

Then,

$$Y = a'_1 + a'_2 x. \quad (11.55)$$

By the method of least squares we find

$$a'_1 = \frac{[Y][x^2] - [x][xY]}{N[x^2] - [x]^2} \quad a'_2 = \frac{N[xY] - [x][Y]}{N[x^2] - [x]^2}. \quad (11.56)$$

If

$$\sigma_Y = \sqrt{\frac{1}{N-2} \sum_{i=1}^N (Y_i - a'_1 - a'_2 x_i)^2} \quad \text{or} \quad \sigma_Y = \sqrt{\frac{[D^2]}{N-2}}, \quad (11.57)$$

with

$$D_i \equiv Y_i - a'_1 - a'_2 x_i, \quad (11.58)$$

the errors in  $a'_1$  and  $a'_2$  are

$$\delta a'_1 = \sigma_{a'_1} = \sigma_Y \sqrt{\frac{[x^2]}{N[x^2] - [x]^2}} \quad \delta a'_2 = \sigma_{a'_2} = \sigma_Y \sqrt{\frac{N}{N[x^2] - [x]^2}}. \quad (11.59)$$

According to Eq. (11.33), the error in  $Y$  at the point  $x$  is

$$\delta Y = \frac{\sigma_Y}{\sqrt{N}} \sqrt{1 + \frac{N^2}{N[x^2] - [x]^2} (x - \bar{x})^2}, \quad \text{where} \quad \bar{x} = \frac{[x]}{N}. \quad (11.60)$$

Since it is  $y = a_0 + xY$ , the error in  $y$  is given by

$$\delta y(x) = \sigma_y(x) = \sqrt{(\delta a_0)^2 + x^2(\delta Y)^2} \text{ or}$$

$$\delta y(x) = \sigma_y(x) = \sqrt{(\delta a_0)^2 + \frac{\sigma_Y^2}{N}x^2 + \frac{N\sigma_Y^2}{N[x^2] - [x]^2}x^2(x - \bar{x})^2}. \tag{11.61}$$

It should be clear what this error expresses. Considering the arguments used above,  $\delta y(x)$  gives a measure of the dispersion of the values of  $y$  derived from measurements at the point  $x$ . The error in the actual value of  $y(x)$  determined from the least squares parabola, which is the result of many measurements, is in fact much smaller. To use an analogy,  $\delta y(x)$  corresponds to what we called *error of a single observation* in the case of a series of measurements of  $y$  at the same value of  $x$ . The error in the reading of  $y(x)$  from the least squares curve would correspond to the *error of the mean value*. It is not clear how one may determine the last magnitude from a series of measurements  $y(x)$  at various values of  $x$ . We will suggest below some ideas to address this problem.

An example will help illustrate the theory presented above.

**Example 11.19**

Using the method of least squares, fit a parabola  $y(x)$  to the points:

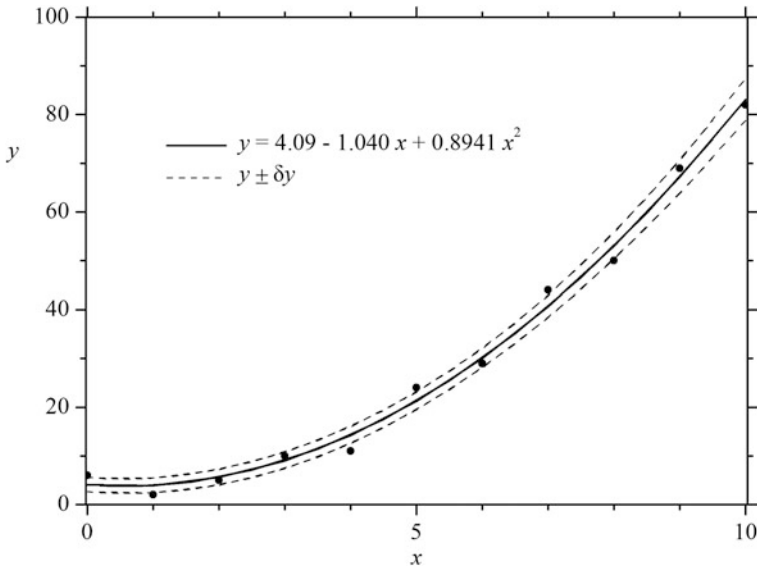
$i$	0	1	2	3	4	5	6	7	8	9	10
$x_i$	0	1	2	3	4	5	6	7	8	9	10
$y_i$	6	2	5	10	11	24	29	44	50	69	82

Find the value of  $y$  and its standard deviation from the mean for  $x = 6$ .

We evaluate the sums

$$N = 11, [x] = 55, [x^2] = 385, [x^3] = 3025, [x^4] = 25333, [y] = 332, [xy] = 2529, [x^2y] = 21,077,$$

which we use in order to fit the least-squares parabola  $y = 4.09 - 1.040x + 0.8941x^2$  to the given points. The points and the parabola are shown in the figure that follows.



We also find the deviations  $d_i$  of the points from the parabola

$i$	0	1	2	3	4	5	6	7	8	9	10
$d_i$	1.91	-1.95	-0.59	0.98	-3.24	2.77	-1.04	3.38	-2.99	1.85	-1.10

and use them to find  $[d^2] = 52.94$ ,  $\sigma_y = 2.57$  and  $\delta a_0 = 1.45$ .

We now define the variable

$$Y_i = \frac{y_i - 4.09}{x_i}$$

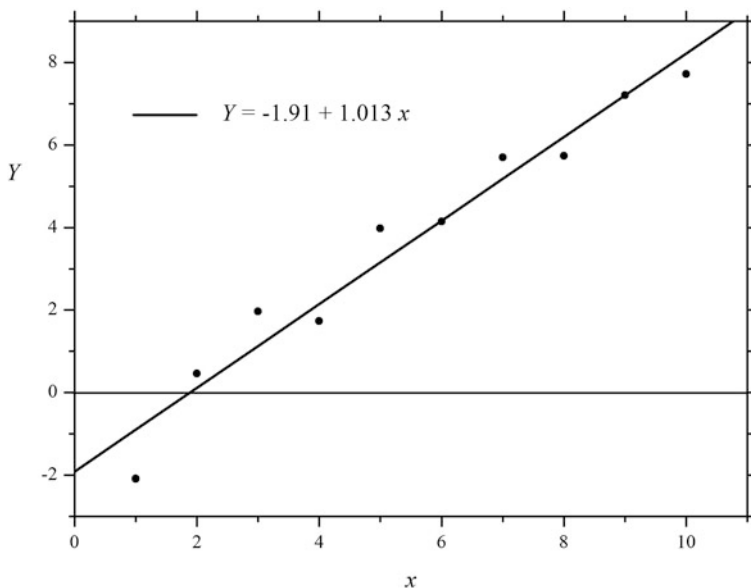
and find its values at the points  $x_i$ .

$i$	0	1	2	3	4	5	6	7	8	9	10
$x_i$	0	1	2	3	4	5	6	7	8	9	10
$Y_i$	-	-2.09	0.46	1.97	1.73	3.98	4.15	5.70	5.74	7.21	7.72
$D_i$	-	-1.20	0.34	0.84	-0.42	0.83	-0.02	0.52	-0.46	0.00	-0.43

Using the sums  $N = 10$ ,  $[x] = 55$ ,  $[x^2] = 385$ ,  $[Y] = 36.64$  and  $[xY] = 285.1$ , we find the values of  $a'_1 = -1.91$  and  $a'_2 = 1.013$ , from which we get the straight line fitted to the points  $Y_i$  by least squares

$$Y = -1.91 + 1.013x.$$

The points  $Y_i(x_i)$  and the straight line  $Y = -1.91 + 1.013x$  are shown in the figure that follows.



The errors in  $a'_1$  and  $a'_2$  can be found. Evaluating  $D_i \equiv Y_i - a'_1 - a'_2x_i$  for each value of  $x_i$ , we find

$$[D^2] = 3.77 \quad \text{and} \quad \sigma_Y = \sqrt{\frac{[D^2]}{N-2}} = \sqrt{\frac{3.77}{8}} = 0.686.$$

The errors in the coefficients are  $\delta a'_1 = \sigma_{a'_1} = 0.468$  and  $\delta a'_2 = \sigma_{a'_2} = 0.0754$ .

We may now use Eq. (11.61) to find the error in  $y$ :

$$\delta y(x) = \sigma_y(x) = \sqrt{(\delta a_0)^2 + \frac{\sigma_y^2}{N}x^2 + \frac{N\sigma_y^2}{N[x^2] - [x]^2}x^2(x - \bar{x})^2},$$

where, here,  $\delta a_0 = 1.45$  and  $\bar{x} = 5.5$ . It follows that

$$\delta y(x) = \sigma_y(x) = \sqrt{2.10 + 0.0471x^2 + 0.005704x^2(x - 5.5)^2}$$





find, using the method of least squares, the constants  $A$  and  $B$ , in terms of the co-ordinates  $(t_i, y_i)$  from  $N$  measurements.

The method of least squares requires the minimization of the quantity

$$S \equiv \sum_{i=1}^N (y_i - A \sin \omega t_i - B \cos \omega t_i)^2.$$

With partial differentiation with respect to  $A$  and  $B$ , we find

$$\begin{aligned} \frac{\partial S}{\partial A} &= -2 \sum_{i=1}^N (y_i - A \sin \omega t_i - B \cos \omega t_i) \sin \omega t_i = 0 \\ \frac{\partial S}{\partial B} &= -2 \sum_{i=1}^N (y_i - A \sin \omega t_i - B \cos \omega t_i) \cos \omega t_i = 0 \end{aligned}$$

from which the normal equations

$$\begin{aligned} [y \sin \omega t] - A[\sin^2 \omega t] - B[\sin \omega t \cos \omega t] &= 0 \\ [y \cos \omega t] - A[\sin \omega t \cos \omega t] - B[\cos^2 \omega t] &= 0 \end{aligned}$$

are obtained, where

$$\begin{aligned} [y \sin \omega t] &\equiv \sum_{i=1}^N y_i \sin \omega t_i, & [\sin \omega t \cos \omega t] &\equiv \sum_{i=1}^N \sin \omega t_i \cos \omega t_i, \\ [\sin^2 \omega t] &\equiv \sum_{i=1}^N \sin^2 \omega t_i \end{aligned}$$

etc. From these, the parameters  $A$  and  $B$  are found to be

$$\begin{aligned} A &= \frac{[\sin^2 \omega t][y \cos \omega t] - [y \sin \omega t][\sin \omega t \cos \omega t]}{[\sin^2 \omega t][\cos^2 \omega t] - [\sin \omega t \cos \omega t]^2} \\ B &= \frac{[\cos^2 \omega t][y \sin \omega t] - [y \cos \omega t][\sin \omega t \cos \omega t]}{[\sin^2 \omega t][\cos^2 \omega t] - [\sin \omega t \cos \omega t]^2}. \end{aligned}$$

These are functions of time.

### 11.3.4 The Reduction of Non-linear Relations to Linear

In certain cases, when the method of least squares is difficult or impossible to apply to a non-linear relation which is considered to apply between the variables, this relation may be transformed to a linear, relating new variables which are suitably defined.

For example, if we have measurements  $[t, R(t)]$  of the variation of the activity  $R(t)$  of the radioactive sample with time  $t$  and we wish to fit to them a relation of the form

$$R(t) = R_0 e^{-\lambda t}, \quad (11.64)$$

we may have a linear relation between the variables

$$x = t, \quad \text{and} \quad y = \ln R(t), \quad (11.65)$$

as it is obviously true that

$$\ln R(t) = \ln R_0 - \lambda t \quad (11.66)$$

and

$$y = \ln R_0 - \lambda x. \quad (11.67)$$

The method of least squares may be applied to relation (11.67) for the determination of  $R_0$  and  $\lambda$ . Of course, the method will give results which are not exactly equal to those we would have obtained by applying the method to the relation of Eq. (11.64). It is obvious that the transformation of the variables changes the relative importance of the measurements. The transformation  $y = \ln R$ , for example, increases the importance of the small values of  $R$ . This will be demonstrated in the example that follows. This somewhat arbitrary use of the method of least squares to the linearized relation is often the only solution we have. Other relations which may be linearized with the suitable change of variables will be examined in Chap. 12 (Sect. 12.4).

#### Example 11.21

$N = 6$  measurements gave the results  $(x_i, y_i)$  of the table below:

$i$	1	2	3	4	5	6
$x_i$	1	2	3	4	5	6
$y_i$	0.8	1.3	1.9	1.9	2.4	2.7

Apply the method of least squares in order to fit to these results a curve, first using the relation  $y = \alpha\sqrt{x}$  and then the linearized relation  $\ln y = \ln \alpha + \frac{1}{2} \ln x$ .

*Method 1*

Using  $y = \alpha\sqrt{x}$ , the deviations of the experimental points are  $d_i = y_i - \alpha\sqrt{x_i}$ .

The magnitude to be minimized is  $S \equiv \sum_{i=1}^N d_i^2 = \sum_{i=1}^N (y_i - \alpha\sqrt{x_i})^2$ .

From  $\frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^N (y_i - \alpha\sqrt{x_i})\sqrt{x_i}$ , we get, for  $\frac{\partial S}{\partial \alpha} = 0$ ,

$$\sum_{i=1}^N (\sqrt{x_i}y_i - \alpha x_i) = [\sqrt{xy}] - \alpha[x] = 0.$$

This gives the value

$$\alpha = \frac{[\sqrt{xy}]}{[x]}.$$

We form the table

$i$	$x_i$	$y_i$	$\sqrt{x_i}$	$\sqrt{x_i}y_i$
1	1	0.8	1	0.8
2	2	1.3	1.414	1.838
3	3	1.9	1.732	3.291
4	4	1.9	2	3.8
5	5	2.4	2.236	5.366
6	6	2.7	2.449	6.612
Sums:	$21 = [x]$		$11.831 = [\sqrt{x}]$	$21.707 = [\sqrt{xy}]$

From the sums of which we get

$$\alpha = \frac{[\sqrt{xy}]}{[x]} = \frac{21.707}{21} = 1.033 \quad \alpha = 1.033.$$

*Method 2*

Linearizing the relation  $y = \alpha\sqrt{x}$ , we get  $\ln y = \ln \alpha + \frac{1}{2} \ln x$ .

Defining  $d_i = \ln y_i - \ln \alpha - \frac{1}{2} \ln x_i$  and  $S \equiv \sum_{i=1}^N d_i^2 = \sum_{i=1}^N (\ln y_i - \ln \alpha - \frac{1}{2} \ln x_i)^2$  and

demanding that  $\frac{\partial S}{\partial \alpha} = 0$  or  $-\frac{2}{\alpha} \sum_{i=1}^N (\ln y_i - \ln \alpha - \frac{1}{2} \ln x_i) = 0$ , we obtain the equation  $[\ln y] - N \ln \alpha - \frac{1}{2} [\ln x] = 0$  from which  $\ln \alpha = \frac{1}{N} ([\ln y] - \frac{1}{2} [\ln x])$  or  $\alpha = \exp\{\frac{1}{N} ([\ln y] - \frac{1}{2} [\ln x])\}$ .

This may also be written as  $\alpha = \left\{ \frac{y_1 y_2 \dots y_N}{\sqrt{x_1 x_2 \dots x_N}} \right\}^{1/N}$  or  $\alpha = (\alpha_1 \alpha_2 \dots \alpha_N)^{1/N}$ ,

where  $\alpha_i = \frac{y_i}{\sqrt{x_i}}$ .

The values of the table give

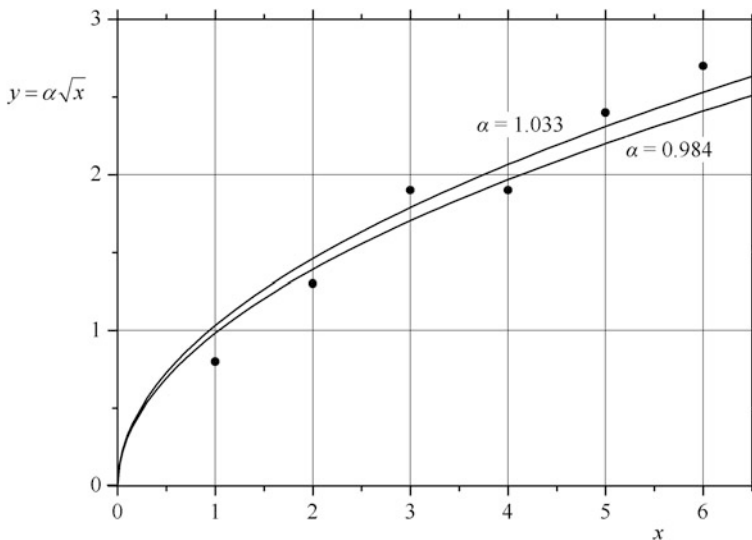
$$\prod_i x_i = 720 \quad \prod_i \sqrt{x_i} = 26.83 \quad \prod_i y_i = 24.329,$$

from which we get

$$\alpha = \left\{ \frac{y_1 y_2 \cdots y_N}{\sqrt{x_1 x_2 \cdots x_N}} \right\}^{1/N} = \left\{ \frac{24.329}{26.83} \right\}^{1/6} = 0.9068^{1/6} = 0.984$$

or, finally  $\alpha = 0.984$ .

The two methods give the slightly different values  $\alpha = 1.033$  and  $\alpha = 0.984$ , respectively. The curves for these two values of  $\alpha$  are drawn in the figure below. The difference is clearly visible.

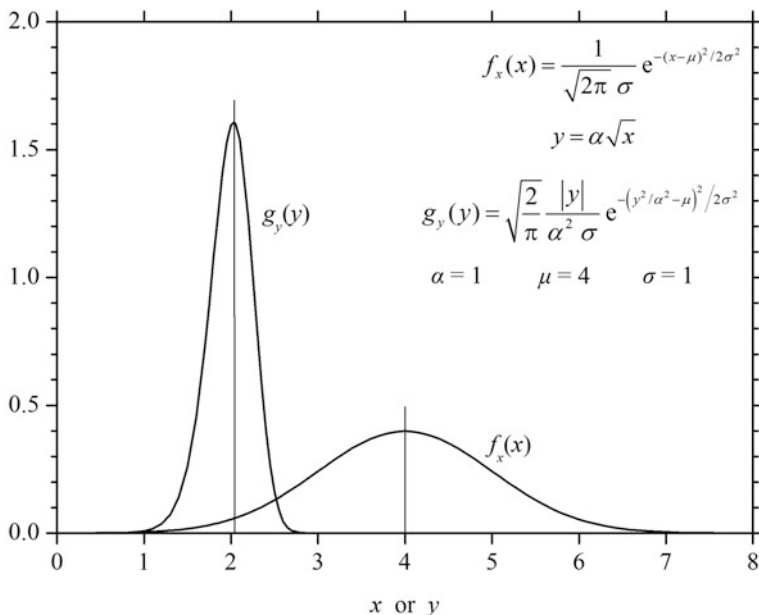


What is the effect of a transformation of variables on their probability densities? Assume that the variable  $x$  has a probability density  $f_x(x)$ . Let this variable be changed into  $y = y(x)$ . We want to determine the probability density  $g_y(y)$  of  $y$ . If to an interval  $dx$  there corresponds an interval  $dy$  and equating the probability  $g_y(y)dy$  of a result in the region between  $y$  and  $y + dy$  with that for a result in the corresponding region between  $x$  and  $x + dx$ , i.e.  $f_x(x)dx$ , we have  $g_y(y)dy = f_x(x)dx$ , from which we finally get

$$g_y(y) = \frac{f_x(x)}{|dy/dx|}, \quad (11.68)$$

where we have taken the absolute value of the derivative since the probability density must be positive. This relation is true for a relation  $y = y(x)$  describing a one-to-one correspondence between the variables  $x$  and  $y$  [2].

As an example, if it is  $f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$  and  $y = \alpha\sqrt{x}$ , the probability density of  $y$  is  $g_y(y) = \sqrt{\frac{2}{\pi}} \frac{|y|}{\alpha^2\sigma} e^{-(y^2/\alpha^2 - \mu)^2/2\sigma^2}$ . These two probability densities are drawn in the figure that follows.



The differences in the two distributions are visible. Apart from the shift on the axis, the Gaussian  $f_x(x)$  is changed into the asymmetrical function  $g_y(y)$ . The two methods, therefore, are bound to give different results.

**Example 11.22 [O]**

Using Origin<sup>®</sup> fit a parabola to the data of Example 11.21.

We place the data of columns  $x$  and  $y$  in columns A and B, respectively. We highlight both columns by left-clicking on the label of column A and then, holding the **Shift** key down, left-clicking on the label of column B. Then,

**Analysis > Fitting > Nonlinear Curve Fit > Open Dialog...**

In the window that opens, we select: **Settings: Function Selection, Category: Power, Function: Power 1**. The function Power 1 is  $y = A|x - x_c|^p$ . We wish to fit

the function  $y = a\sqrt{x}$ , so we must set  $x_c = 0$  and  $p = 1/2$ . To do this, we open **Parameters**. For  $x_c$  we tick the box **Fixed** and enter the value 0. For  $p$  we tick the box **Fixed** and enter the value 0.5 (not 1/2). Then we press **Fit**.

The result returned is:  $A(= a) = 1.03379 \pm 0.03866$ , so that it is  $y = 1.034\sqrt{x}$ . This agrees well with the results of Example 11.16.

## 11.4 The Choice of the Optimum Function Fitted to a Set of Experimental Results

The method of least squares gives us the best coefficients for the function we chose to fit to a series of experimental results. What it does not give us is the best function to be used. It can, however, tell us which of the various functions we have tried has a better fit to the experimental results.

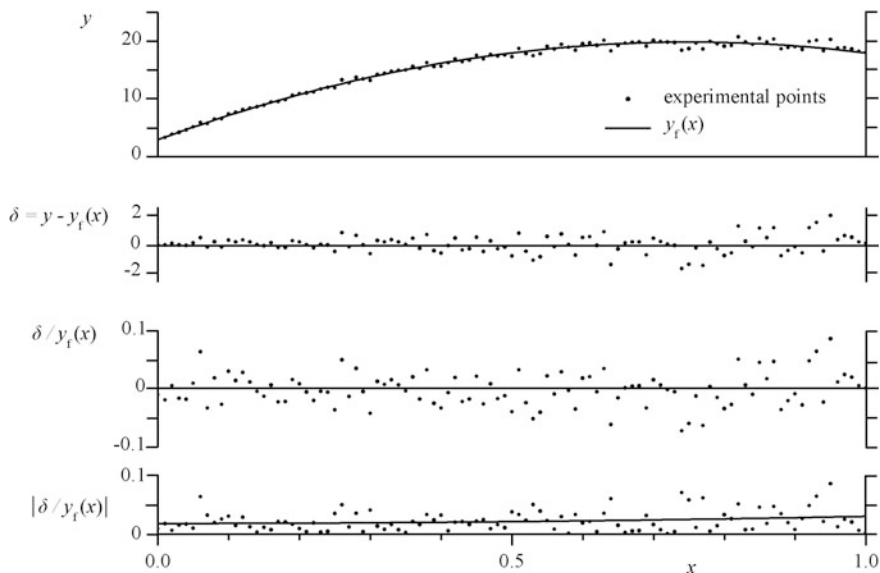
Let the two functions which were used,  $\alpha$  and  $\beta$ , have  $n_\alpha$  and  $n_\beta$  parameters respectively (2 for a straight line, 3 for a parabola etc.). For the  $N$  values  $x_i$ , which are assumed as known with absolute accuracy, the two functions give, respectively, the values  $y_{\alpha,i}$  and  $y_{\beta,i}$ . We evaluate the magnitudes

$$\Phi_\alpha \equiv \frac{1}{N - n_\alpha} \sum_{i=1}^N (y_i - y_{\alpha,i})^2 \quad \text{and} \quad \Phi_\beta \equiv \frac{1}{N - n_\beta} \sum_{i=1}^N (y_i - y_{\beta,i})^2. \quad (11.69)$$

It is proved that the function with the smaller value of  $\Phi$  gives the best fit to the experimental results.

## 11.5 The Fractional Absolute Deviation of the Experimental Values from the Values of the Curve

Assume that a curve has been fitted to the scatter data of an experiment, such as the ones shown in Fig. 11.3, passing between the experimental points, either using the least squares method or by applying smoothing to the data (see next section). If we read the value of  $y$  as given by this curve for a particular  $x$ , what is a measure of dispersion for this  $y$  value? If, as is the case for a straight line or a simple curve used in the method of least squares, the standard deviation in the  $y$  values is given by a formula, then there is no problem. In most cases, however, this is not possible. The results obtained in Chap. 4 do not apply here, as we do not have many measurements of a physical quantity under the same experimental conditions but many measurements performed at different values of the independent variable  $x$ . We may obtain a *measure* for the scatter of the points about the smoothed curve by working as described below.



**Fig. 11.3** The curve  $y_f(x)$  is fitted to the experimental results. The deviation  $\delta = y - y_f(x)$ , the fractional deviation  $\delta/y_f(x)$  and the absolute fractional deviation  $|\delta/y_f(x)|$  of the experimental points from the values given by the graph are evaluated for all the experimental points. A parabola is fitted by the method of least squares to the points  $|\delta/y_f(x)|$  (line in the lower part of the figure)

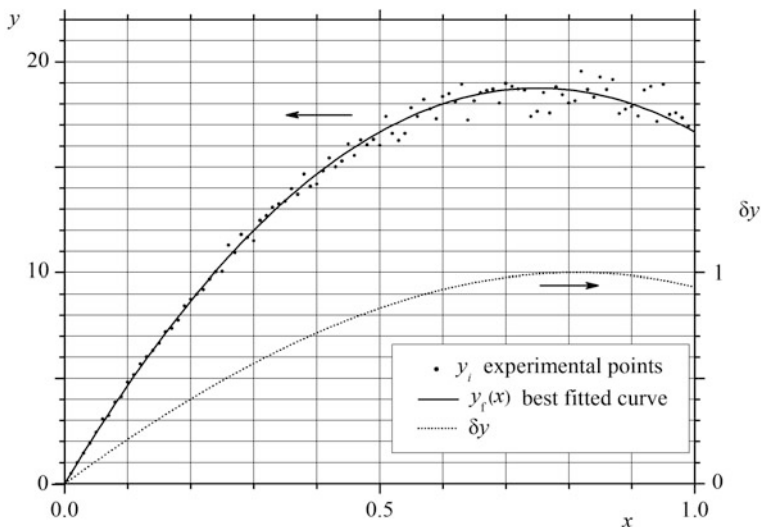
We wish to evaluate an estimate for the fractional absolute deviation of the experimental values from the curve,

$$\frac{\delta(x)}{y_f(x)} = \frac{y_{\text{exper.}}(x) - y_f(x)}{y_f(x)}, \tag{11.70}$$

as a function of  $x$ . Here,  $y_{\text{exper.}}(x)$  is the value of  $y$  at  $x$ , expected from an experimental measurement and  $y_f(x)$  is the value given by the curve at  $x$ . The steps of the procedure followed are shown in Fig. 11.3.

Using the experimental results  $y_i$  and the curve fitted to them,  $y_f(x)$ , we evaluate the deviation  $\delta = y - y_f(x)$ , the fractional deviation  $\delta/y_f(x)$  and the absolute fractional deviation  $|\delta/y_f(x)|$  for each experimental point. Figure 11.4 shows the experimental points,  $y_i$ , the best curve fitted to them,  $y_f(x)$ , and the deviations of the experimental points from the curve,  $\delta y$ , as functions of  $x$ .

The values of  $\delta y$  evaluated by the method described above, simply gives a measure of the dispersion of the experimental points about the fitted curve. It does not give the error in a value of  $y$  read off the curve. What was found above is the equivalent of the standard deviation of the measurements about their mean. We need the equivalent of the standard deviation of the mean, which may also be considered to be the error in  $y$ . A suggestion on how an estimate for such a magnitude may be obtained will be given below.



**Fig. 11.4** The experimental points,  $y_i$ , the best curve fitted to them,  $y_f(x)$ , and the deviations of the experimental points from the curve,  $\delta y$ , as functions of  $x$

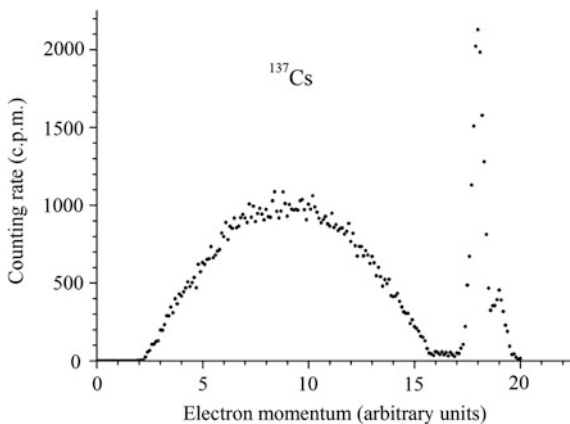
## 11.6 Smoothing

It is often impossible to fit a simple curve to the experimental values by the method of least squares or otherwise. Usually, the reason is that a curve which would agree sufficiently with the experimental points does not have the form of a polynomial or other simple functions. An example is given in Fig. 11.5. There is a very rich library of specialized functions used in particular branches of science, e.g. in optical, dielectric or gamma-ray spectroscopy. Even so, in many cases, curves appear which do not have a known or simple structure enabling the fitting to them of a curve of known mathematical form. In some cases we can settle for a curve through the points which is smooth enough so that we can read the result of a possible measurement at any value of the independent variable. This is achieved with a procedure called *smoothing*.

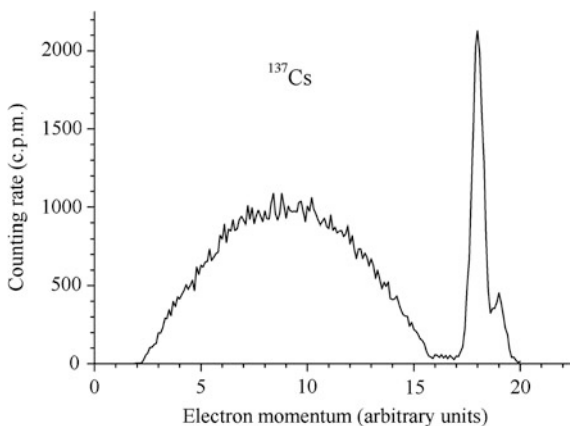
Figure 11.5 shows the momentum spectrum of electrons emitted by the radioisotope  $^{137}\text{Cs}$ . The details do not concern us here, but in essence the points represent a histogram of the momenta of the electrons emitted, each point representing the electrons counted in a narrow interval of values of the momentum. The dispersion of the points is due to the statistical fluctuations in the numbers of the electrons counted. This dispersion is made more obvious in Fig. 11.6, in which consecutive points have been joined by straight lines. It is clear that it would not be easy to apply to the whole curve the method of least squares without destroying the fine structure of the spectrum in the region of the two narrow peaks at the large values of momentum.



**Fig. 11.5** The momentum spectrum of the electrons emitted from the radioisotope  $^{137}\text{Cs}$ , as recorded by a multichannel analyzer. Apart from the continuous spectrum which is due to the  $\beta$  emission from the nucleus, two narrow peaks are also observed, of monoenergetic electrons due to internal conversion: a large one with electrons from the K shell and a smaller one with electrons from the L shell



**Fig. 11.6** The spectrum of Fig. 11.5. Consecutive points have been joined with straight lines



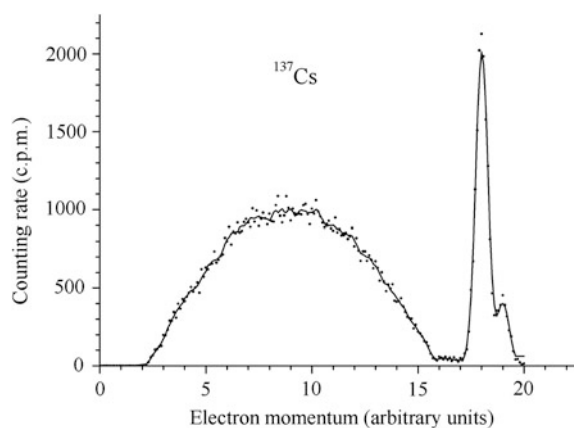
The smoothing of a curve is achieved by applying the method of least squares or some procedure of averaging on parts of the curve separately. Using this method is much simpler when the experimental points are at the same distance between them along the axis of the independent variable ( $x$ ), as is the case in Fig. 11.5. A new value of  $y$  is calculated for every value of  $x$ , by fitting a curve to only  $2N + 1$  consecutive points, with central point that at which the new value of  $y$  is being evaluated. The values of  $2N + 1$  are 3, 5, 7 etc. and the curve fitted to these points is a simple polynomial of the second or not very much higher degree. There are various equations for the application of the method, for points which are mutually equidistant or not, or others which take into account, for example, that at the edges the points available are not sufficient for the calculations. Relative weights may also be attributed to each point, depending on its distance from the central point. It is possible, of course, to apply the same procedure two or more times in succession. This must be avoided if the number of points involved is too large, as it will lead to

an over-smoothing of the curve, in which points at very large distances from the central point affect its value.

Various smoothing methods are available in data analysis computer programs. The simplest method is that of averaging the  $y$ -values of  $2N + 1$  points symmetrically situated about each experimental point in succession [e.g.  $y'_n = (y_{n-2} + y_{n-1} + y_n + y_{n+1} + y_{n+2})/5$  for the  $n$ th point of the data]. Here,  $y'_n$  replaces  $y_n$  in the smoothed curve. A better and more popular method is known as the *Savitzki-Golay method*. This uses, for each experimental point,  $2N + 1$  points symmetrically situated about the central one and fits a least-squares polynomial to them. The value of  $y$  at the central point is then evaluated using the resulting polynomial. Obviously, the value of  $N$  must be decided taking into account the total number of points available. The degree of the polynomial must not be too high, otherwise the effectiveness of the method is reduced. In the limit, if the degree of the polynomial is equal to  $N$ , there will be no change in the re-calculated  $y$ -values of the points! The method was applied to the data of Fig. 11.5, with  $2N + 1 = 7$  points and a polynomial of the second degree (parabola). The differences between the curves of Figs. 11.6 and 11.7 are obvious.

Smoothing must be applied with great caution and only when it would offer an improvement to a scatter plot or to a table of data. It is useful to remember that smoothing is equivalent to ‘filtering’ the curve by a filter that cuts off the high frequencies. In other words, the process removes the high-frequency variations from the curve. In the final analysis, this is equivalent to the diminishing of the power of discrimination of the experimental method used. In Fig. 11.7, this is demonstrated by the broadening of the two narrow peaks. Greater smoothing might possibly make invisible the small peak at the higher values of momentum. The fine structure in the data, which may be of great physical importance, could be made to disappear by an excessive use of smoothing.


**Fig. 11.7** The spectrum of Fig. 11.5, after smoothing by the method of Savitzki-Golay, using  $2N + 1 = 7$  points and a polynomial of the second degree. The scatter points are also shown for comparison



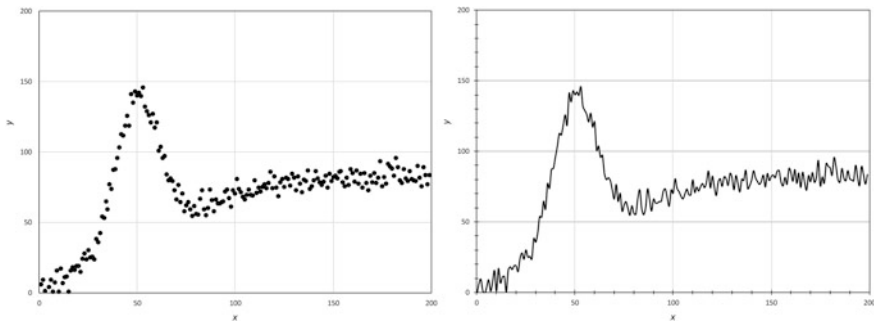
**Example 11.23 [E]**

A number of experimental results are given, which are shown in a scatter plot below. Use Excel® to transform the data to a smoothed curve.

We enter the values of  $x$  in column A and those of  $y$  in column B. We highlight the two columns and, through **Insert**, we produce a scatter plot of the experimental points. This is shown in the left hand side figure below.

While in the chart page, we right-click on one of the points and, in the window that opens, we select **Format Data Series**. The **Format Data Series** task pane opens. Click the **Fill and Line** icon . Select **Solid Line** and then check the **Smoothed Line** box. Click **OK**. A line appears in the plot, joining the points.

Right-click on the line and, in the window that opens, select **Change Series Chart Type**. Select **Line** plot and **Scatter With Smooth Lines**. The dots will disappear. After some formatting, the graph looks like the right-hand figure shown below.

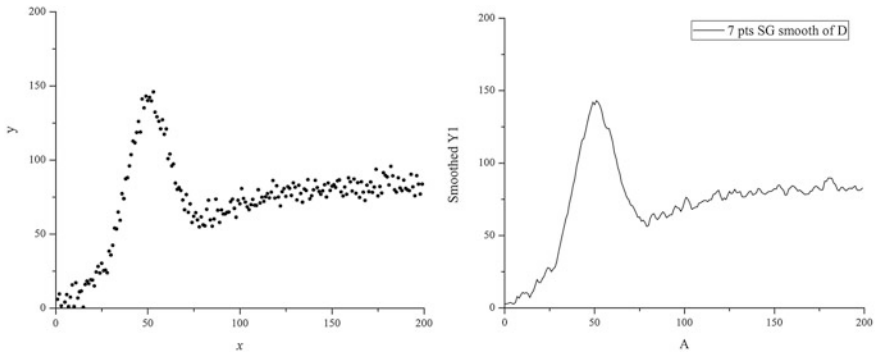


Strictly speaking, what Excel® does here is not smoothing. It just joins the dots with straight lines and rounds off the corners.

**Example 11.24 [O]**

A number of experimental results are given, which are shown in a scatter plot below. Using Origin Origin®, perform a 7-point parabola Savitzki-Golay smoothing operation on these data and show the result.

We import the data  $(x, y)$  and place them in columns A and D, respectively. The scatter plot of the data is shown in the figure on the left.



We select column D by left clicking on its label. Then,

**Analysis > Signal Processing > Smooth > Open Dialog**

In the window that opens, we select:

**Method: Savitzki–Golay, Points of Window: 7, Polynomial Order: 2.**

Press **OK**. The smoothed data appear in a new column. Give the instructions:

**Plot > Line > Line.**

The smoothed data are plotted as shown above, in the figure on the right.

### Example 11.25 [P]

A number of experimental results are given, which are shown in a scatter plot below. Use Python to transform the data to a smoothed curve.

First the data vector  $y$  is entered (we omit this operation for brevity), and we create a vector  $x$ , containing a series of integers from 0 to the length of  $y$ . We calculate a corresponding vector of smoothed data using the `savgol_filter` function from the `scipy.signal` sub-package. The function accepts three parameters: the original data, the number of window points and the polynomial order; like in the previous example, we use a 7-point window and a 2nd degree polynomial for the Savitzky-Golay smoothing operation.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.signal import savgol_filter

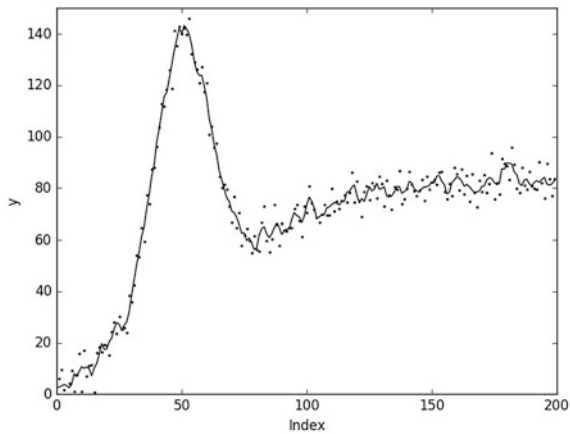
y = np.array([-1.43019, 6.04592, 9.58303, ... 83.65553])
x = np.arange(0, len(y))
plt.scatter(x, y, s=2, color="black")
```

```

plt.xlim(0, 200)
plt.ylim(0, 150)
plt.xlabel("Index")
plt.ylabel("y")
ysmooth = savgol_filter(y, 7, 2)
plt.plot(x, ysmooth, '-', color="blue")
plt.show()

```

The following figure is produced, combining the scatter plot of the original data and the smoothed curve.



### Example 11.26 [R]

A number of experimental results are given, which are shown in a scatter plot below. Use R to transform the data to a smoothed curve.

The points are the same as in the two previous examples. We will achieve smoothing by the use of a cubic spline.

```

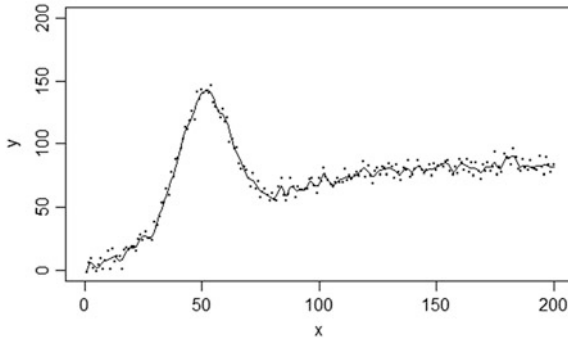
# The data vectors are entered:
> y <- c(-1.43019, 6.04592, 9.58303, 1.54254, ...
... 75.90812, 89.37784, 83.67599, 77.00079, 83.65553)
> x <- seq(1, length(y), len = 201)

# The scatter plot is drawn:
> s02 <- smooth.spline(y, spar = 0.2)
> plot(y, pch = 20, cex = 0.5, xlab = "x", ylab = "y", xlim=c
(0, 200), ylim=c(0, 200), col.main = 2)
>

```

```
# The smoothed curve is drawn:
> lines(predict(s02, x))
```

The results are shown in the figure below.



## 11.7 The Error in a Value Read off a Smoothed Curve

We need an estimate for the standard deviation or error of a point on a curve which was obtained by the method of least squares or by smoothing of data. This is possible in the case of a curve obtained by smoothing using the simple averaging procedure. For example, if the smoothing of a curve is done by finding the average

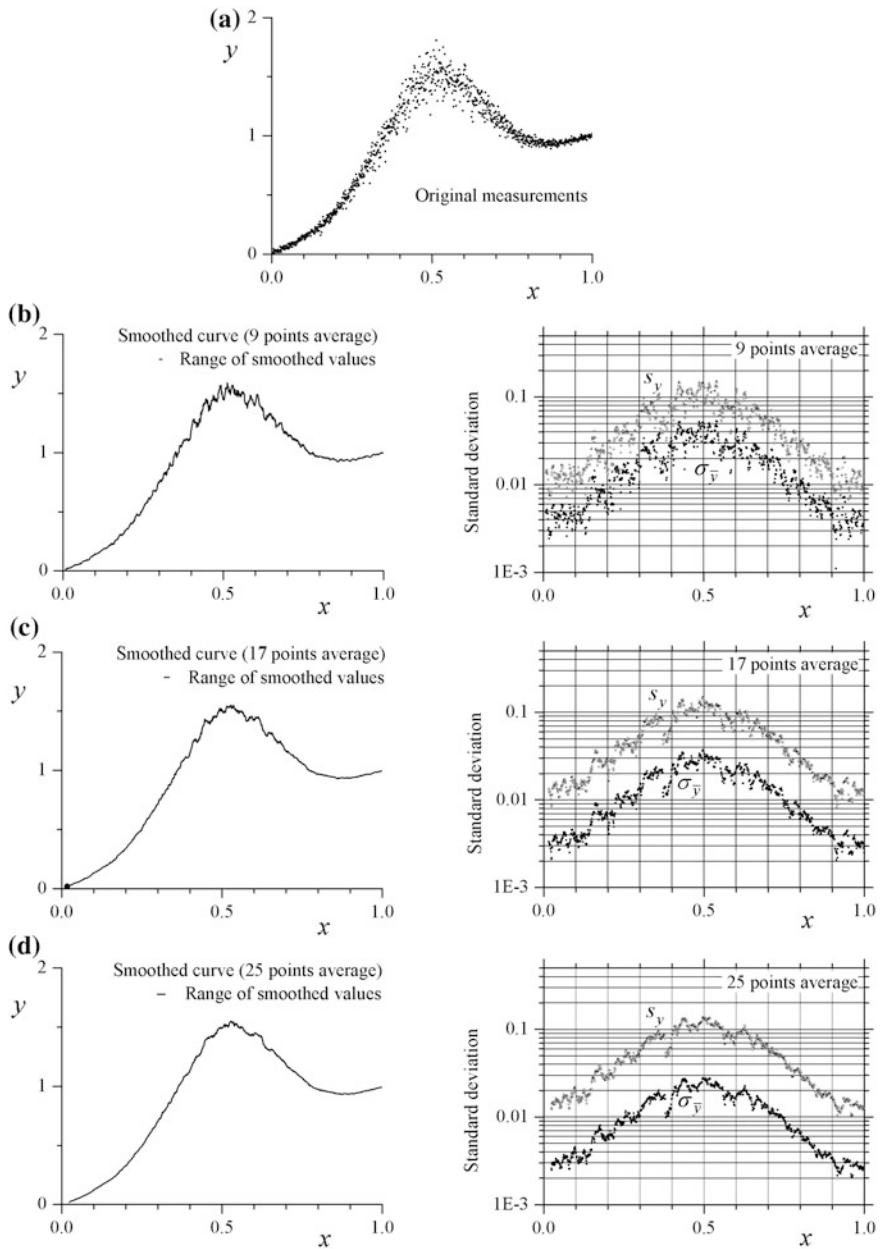
$$y'_n = (y_{n-k} + y_{n-k+1} + \dots + y_n + \dots + y_{n+k-1} + y_{n+k}) / (2k + 1), \quad (11.71)$$

then we may consider the  $2k + 1 = N$  measurements at slightly different values of  $x$ , as measurements performed under approximately the same conditions and evaluate their mean,  $\bar{y}$ , standard deviation  $s_y$  and standard deviation of their mean  $\sigma_{\bar{y}}$ .

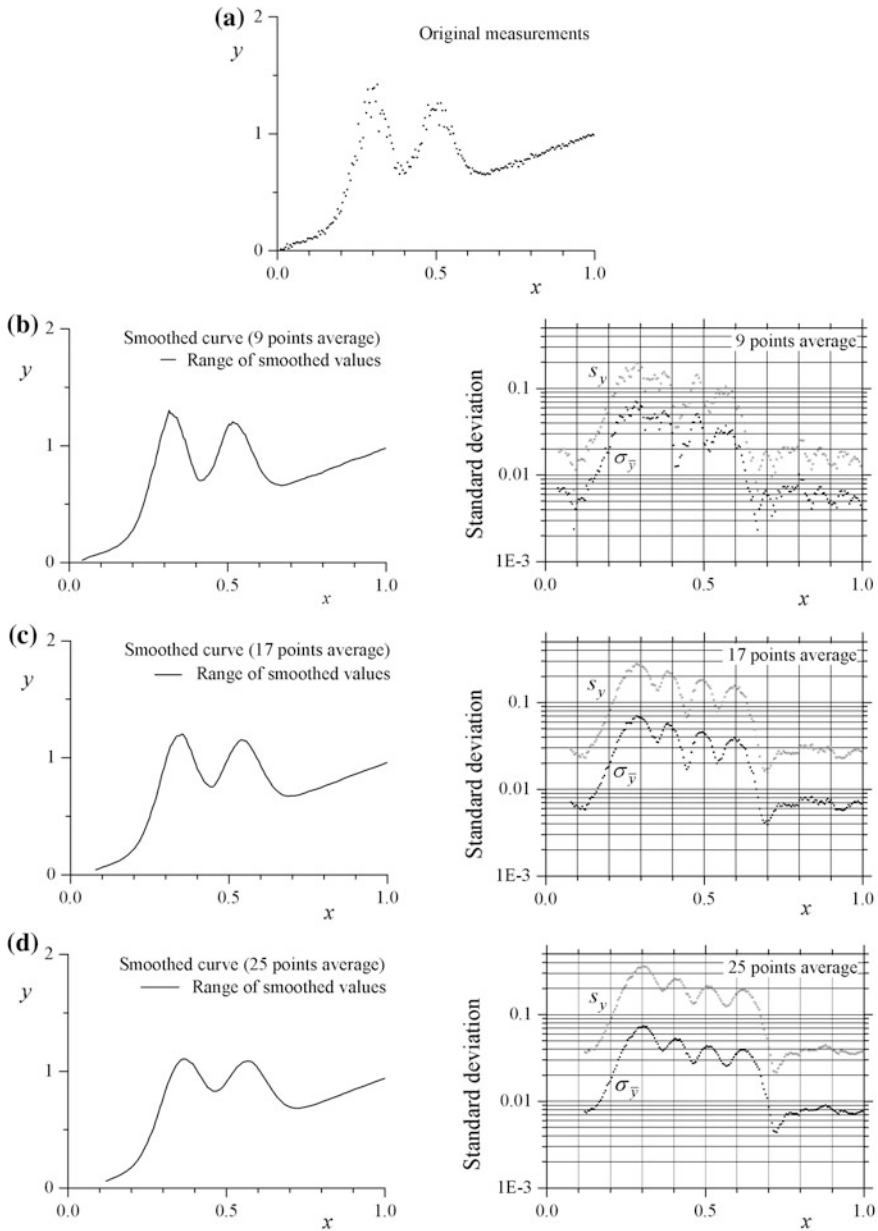
An example is shown in Fig. 11.8. The noisy original signal is shown in (a). In three different cases, smoothing is performed by averaging 9, 17 and 25 adjacent points [figures (b), (c) and (d), respectively]. In each case, estimates of the standard deviations  $s_y$  and  $\sigma_{\bar{y}}$  are evaluated. It is seen that  $s_y$ , as expected, tends to stabilize at some value, while  $\sigma_{\bar{y}}$  decreases, as the number of points averaged ( $N$ ) increases.

This is as expected, since it is  $\sigma_{\bar{y}} = s_y / \sqrt{N - 1}$ .

The question which arises concerns the optimum number of points to be used in the smoothing of the curve and, therefore, in the evaluation of  $\sigma_{\bar{y}}$ . No quantitative criterion exists, so we are obliged to make a subjective judgment, trying to minimize  $\sigma_{\bar{y}}$  as much as possible (using a large  $N$ ) while not deforming the curve too



**Fig. 11.8** The smoothing of a curve consisting of 1001 experimental points, **a**, by taking the averages of various numbers of points (9, 17 and 25 here) and the evaluation of the corresponding estimates for the standard deviation of the points,  $s_y$ , and of their mean  $\sigma_{\bar{y}}$  (**b**, **c** and **d**)



**Fig. 11.9** The smoothing of a curve consisting of 201 experimental points, **a**, by taking the averages of various numbers of points (9, 17 and 25 here) and the evaluation of the corresponding estimates for the standard deviation of the points,  $s_y$ , and of their mean  $\sigma_{\bar{y}}$  [**b**, **c** and **d**]



much (by using a small  $N$ ). In the example of Fig. 11.8, the series of measurements consists of 1001 results. Given the variation of the signal, an averaging using 25 points does not seem unreasonable. It represents 1/40th of the whole range, and it is seen that the signal does not change significantly over this range. Figure 11.9 shows another case. There are only 201 points and in the smoothing, the 17 or 25 points used cover a significant part of the whole range of values (the ranges are shown in the graphs by small horizontal lines). As a result, over the range of the smoothing, the signal varies significantly and, consequently,  $s_y$  and  $\sigma_{\bar{y}}$  increase with increasing  $N$ . The loss of detail in the smoothed curves is also obvious. It is seen, that in this case, using more than 9 points in the averaging for the smoothing does not offer any advantage.

Ideally, there should be a strict mathematical method for finding the error ( $\sigma_{\bar{y}}$ ) at any point of a series of measurements such as that of Fig. 11.8a. This estimate would depend on the values at all the measurements. Such a method, however, is not available. We are thus forced to use the somewhat arbitrary method described above, based on smoothing. In all cases, we should consider the results obtained using the method described above as giving an order of magnitude estimate for the error in  $y$  as a function of the independent variable,  $x$ .

## 11.8 The Regression Line and the Coefficient of Correlation

In Chap. 6, Sect. 6.2.3, we found the mean value and the standard deviation of a function  $Q = Q(x, y)$  of two variables  $x, y$ . Making use of those results in the case of  $N$  pairs of values  $(x_i, y_i) (i = 1, 2, \dots, N)$ , if we expand the function  $Q = Q(x, y)$  in a Taylor series in the region of the point  $(\bar{x}, \bar{y})$ , where  $\bar{x}$  and  $\bar{y}$  are the means of  $x$  and  $y$ , we have

$$Q(x, y) = Q(\bar{x}, \bar{y}) + \left( \frac{\partial Q}{\partial x} \right)_{\bar{x}, \bar{y}} (x - \bar{x}) + \left( \frac{\partial Q}{\partial y} \right)_{\bar{x}, \bar{y}} (y - \bar{y}) + \dots \quad (11.72)$$

and find for the mean of the function, approximately,

$$\bar{Q} = Q(\bar{x}, \bar{y}). \quad (11.73)$$

The standard deviation of  $Q$  is found from the relation

$$\sigma_Q^2 = \frac{1}{N} \sum_i \left[ \left( \frac{\partial Q}{\partial x} \right)_{\bar{x}, \bar{y}} (x_i - \bar{x}) + \left( \frac{\partial Q}{\partial y} \right)_{\bar{x}, \bar{y}} (y_i - \bar{y}) \right]^2, \quad (11.74)$$

$$\begin{aligned} \sigma_Q^2 = & \left( \frac{\partial Q}{\partial x} \right)_{\bar{x}, \bar{y}}^2 \frac{1}{N} \sum_i (x_i - \bar{x})^2 + \left( \frac{\partial Q}{\partial y} \right)_{\bar{x}, \bar{y}}^2 \frac{1}{N} \sum_i (y_i - \bar{y})^2 \\ & + 2 \left( \frac{\partial Q}{\partial x} \right)_{\bar{x}, \bar{y}} \left( \frac{\partial Q}{\partial y} \right)_{\bar{x}, \bar{y}} \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \end{aligned} \quad (11.75)$$

This expression may be written in the form

$$\sigma_Q^2 = \left( \frac{\partial Q}{\partial x} \right)_{\bar{x}, \bar{y}}^2 \sigma_x^2 + \left( \frac{\partial Q}{\partial y} \right)_{\bar{x}, \bar{y}}^2 \sigma_y^2 + 2 \left( \frac{\partial Q}{\partial x} \right)_{\bar{x}, \bar{y}} \left( \frac{\partial Q}{\partial y} \right)_{\bar{x}, \bar{y}} \sigma_{xy} \quad (11.76)$$

where  $\sigma_x$  and  $\sigma_y$  are the standard deviations of  $x$  and  $y$ , and

$$\sigma_{xy} \equiv \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \quad (11.77)$$

is the *covariance* of  $x$  and  $y$ . This is a property of the sample of the measurements. The best estimate for the covariance of the parent population is

$$\hat{\sigma}_{xy} = \frac{N}{N-1} \sigma_{xy} = \frac{1}{N-1} \sum_i (x_i - \bar{x})(y_i - \bar{y}). \quad (11.78)$$

Equation (11.76) gives the standard deviation of  $Q$  whether  $x$  and  $y$  are independent of each other or not. If they are independent of each other, their covariance tends to zero as the number of measurements tends to infinity.

In the case of fitting a straight line to the points  $(x_i, y_i)$  using the method of least squares, we have found that the equation of the line may be written in the form

$$\alpha + \lambda \frac{[x]}{N} = \frac{[y]}{N}, \quad (11.79)$$

i.e. that the straight line passes through the point  $K : (\bar{x} = [x]/N, \bar{y} = [y]/N)$ , which we called center of the line. If we define the variables

$$X \equiv x - \bar{x} \quad \text{and} \quad Y \equiv y - \bar{y}, \quad (11.80)$$

the equation of the straight line is

$$Y = \lambda X \quad (11.81)$$

and, according to the method of least squares, Eq. (11.35), it will be

$$\lambda = \frac{[XY]}{[XX]}. \quad (11.82)$$

Therefore, the straight line is given by the equation

$$y - \bar{y} = \frac{[XY]}{[XX]}(x - \bar{x}). \quad (11.83)$$

Using the relations  $[XX] = N\sigma_x^2$ ,  $[YY] = N\sigma_y^2$  and  $[XY] = N\sigma_{xy}$ , and defining the (*Pearson*) *coefficient of linear correlation*

$$r \equiv \frac{[XY]}{\sqrt{[XX][YY]}}, \quad (11.84)$$

we may write Eq. (11.83) as

$$\frac{y - \bar{y}}{\sigma_y} = r \frac{x - \bar{x}}{\sigma_x}. \quad (11.85)$$

This straight line is called *regression line of y on x*.

The correlation coefficient is written in the forms

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}, \quad (11.86)$$

where the sums are performed over all the values  $i = 1, 2, \dots, N$ .

The coefficient of correlation  $r$  is a measure of how well the points  $(x_i, y_i)$  are described by the regression line. It may take values

$$-1 \leq r \leq 1. \quad (11.87)$$

If the coefficient of correlation  $r$  is near the values  $\pm 1$ , the points are near a straight line. If it has values near 0, the points are not correlated and there is no line that could be fitted to them satisfactorily. Let us note that, if all the points lie on the straight line  $y = \alpha + \lambda x$ , then it is  $y_i = \alpha + \lambda x_i$  for every  $i$  and, also,  $\bar{y} = \alpha + \lambda \bar{x}$ . Subtracting, we find that  $y_i - \bar{y} = \lambda(x_i - \bar{x})$  for every point. Therefore, Eq. (11.86) gives

$$r = \frac{\lambda \sum (x_i - \bar{x})^2}{\sqrt{\sum (x_i - \bar{x})^2 \lambda^2 \sum (x_i - \bar{x})^2}} = \frac{\lambda}{|\lambda|} = \pm 1. \quad (11.88)$$

The conclusion is: if all the points lie exactly on a straight line, then  $r = \pm 1$  and the sign is that of the line's slope.

At the other end, if  $x$  and  $y$  are not correlated to each other, then the sum  $\sum (x_i - \bar{x})(y_i - \bar{y})$  tends to zero as the number of points increases, since the terms are equally probable to be positive or negative. For a finite number of uncorrelated measurements, the coefficient of correlation  $r$  has values near 0.

**Table 11.1** The probability  $P\{|r| \geq r_0\}$  for the absolute value of the coefficient of linear correlation  $|r|$  of a number  $N$  of points  $(x_i, y_i)$  to be greater than or equal to some value  $r_0$  due to a coincidence and not due to the correlation of the variables  $x$  and  $y$  with each other

$N$	$P\{ r  \geq r_0\}$										
	$r_0$										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
4	1	0.90	0.87	0.81	0.74	0.67	0.59	0.51	0.41	0.29	0
6	1	0.85	0.70	0.56	0.43	0.31	0.21	0.12	0.056	0.014	0
8	1	0.81	0.63	0.47	0.33	0.21	0.12	0.053	0.017	0.002	0
10	1	0.78	0.58	0.40	0.25	0.14	0.067	0.024	0.005		0
12	1	0.76	0.53	0.34	0.20	0.098	0.039	0.011	0.002		0
14	1	0.73	0.49	0.30	0.16	0.069	0.023	0.005	0.001		0
16	1	0.71	0.46	0.26	0.12	0.049	0.014	0.003			0
18	1	0.69	0.43	0.23	0.10	0.035	0.008	0.001			0
20	1	0.67	0.40	0.20	0.081	0.025	0.005	0.001			0
25	1	0.63	0.34	0.15	0.048	0.011	0.002				0
30	1	0.60	0.29	0.11	0.029	0.005					0
35	1	0.57	0.25	0.080	0.017	0.002					0
40	1	0.54	0.22	0.060	0.011	0.001					0
45	1	0.51	0.19	0.045	0.006						0
50	1	0.49	0.16	0.034	0.004						0

When no value is given, the probability is less than 0.0005

Having fitted a straight line to a group of experimental points, it would be very useful to know whether the two variables are not correlated with each other and that the curve fit is simply the result of a coincidence. Given in Table 11.1 is, for a given number of measurements  $N$ , the probability  $P\{|r| \geq r_0\}$  for the value of the correlation coefficient to be greater than or equal to some value  $r_0$  due to a coincidence and not because of the correlation of the variables  $x$  and  $y$  with each other.

For example, for  $N = 10$  points, a coefficient of linear correlation greater than or equal to  $r_0 = 0.5$  has a probability of 0.14 (or 14%) to be due to a coincidence and not to a correlation of  $x$  and  $y$  with each other. For the same number of points, the value  $r_0 = 0.8$  has a probability of 0.005 (or 0.5%) to be due to a coincidence.

**Example 11.27**

Find the coefficient of linear correlation for the points  $(x_i, y_i)$  of Example 11.1 and the probability for the linear relationship between  $x$  and  $y$  to be due to a coincidence.

In Example 11.1 we found  $\bar{x} = 1.00$  and  $\bar{y} = 3.22$ .

For the evaluation of  $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$  we complete the table below:

$i$	$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	0.0	0.92	-1.0	-2.3	1.00	5.2900	2.300
2	0.2	1.48	-0.8	-1.74	0.64	3.0276	1.392
3	0.4	1.96	-0.6	-1.26	0.36	1.5876	0.756
4	0.6	2.27	-0.4	-0.95	0.16	0.9025	0.380
5	0.8	2.61	-0.2	-0.61	0.04	0.3721	0.122
6	1.0	3.18	0	-0.04	0	0.0016	0
7	1.2	3.80	0.2	0.58	0.04	0.3364	0.116
8	1.4	4.01	0.4	0.79	0.16	0.6241	0.316
9	1.6	4.85	0.6	1.63	0.36	2.6569	0.978
10	1.8	5.10	0.8	1.88	0.64	3.5344	1.504
11	2.0	5.26	1.0	2.04	1.00	4.1616	2.040
Sums					4.40	22.5	9.90

Therefore,  $r = \frac{9.90}{\sqrt{4.40 \times 22.5}} = 0.995$ .

The probability that this value of the coefficient of linear correlation is due to a coincidence is extremely small, as seen in Table 11.1.

### Example 11.28 [E]

Using Excel<sup>®</sup>, evaluate the correlation coefficient for the data of Example 11.27.

Copy the  $x$  values into column A and the  $y$  values into column B. Highlight an empty cell. From **Formulas > More Functions, Statistical**, select **Correl**. This opens the correlation window. Fill **Array1** by right-clicking on cell A3 and dragging the cursor to A13. Similarly, fill **Array2** with the values in the cells B3 to B13. Pressing **OK** returns the value for the coefficient of correlation as  $r = 0.99551$ .

### Example 11.29 [O]

Using Origin<sup>®</sup>, evaluate the correlation coefficient for the data of Example 11.27.

We place the data of columns  $x$  and  $y$  of the table of Example 11.27 in columns A and B, respectively. We highlight both columns by left-clicking on the label of column A and then, holding the **Shift** key down, left-clicking on the label of column B. Then

**Statistics > Descriptive Statistics > Correlation Coefficient > Open Dialog...**

In the window that opens, we select: **Correlation Types: Pearson**. Press **OK**.

The result returned is: AB or BA Pearson Correlation Coefficient = 0.99551. This is the same result as the one found in Example 11.27.

**Example 11.30 [P]**

Evaluate the correlation coefficient for the data of Example 11.27.

The `scipy.stats` subpackage includes the function `pearsonr` to calculate the Pearson correlation coefficient. We first enter the data into two vectors, and then invoke the function: it returns the value of  $r$  and the two-tailed  $p$ -value for testing non-correlation.

```
import numpy as np
from scipy.stats.stats import pearsonr

x = np.array([0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0])
y = np.array
([0.92, 1.48, 1.96, 2.27, 2.61, 3.18, 3.80, 4.01, 4.85, 5.10, 5.26])
pearsonr(x, y)
```

The result is  $r = 0.99551$ , with a  $p$ -value of  $1.5913E-10$ .

**Example 11.31 [R]**

Evaluate the correlation coefficient for the data of Example 11.27.

Enter the data vectors:

```
> x <- c(0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0)
> y <- c(0.92, 1.48, 1.96, 2.27, 2.61, 3.18, 3.80, 4.01, 4.85, 5.10, 5.26)
```

Calculate the Pearson correlation coefficient:

```
> cor(x, y, method = "pearson")
[1] 0.9955053
```

The result,  $r = 0.99551$ , is in agreement with those of the two previous Examples.

## 11.9 The Use of the Method of Least Squares in the Solution of a System of Overdetermined Linear Equations

The method of least squares was used by Legendre in order to find the optimum solutions of systems of linear equations, in those cases when the number of equations is larger than the number of unknowns and the equations are not all satisfied by a certain set of values of the unknowns.

### 11.9.1 Equations in Two Variables

Let the linear equations involve two variables,  $x$  and  $y$ . Given are  $N > 2$  equations

$$a_i x + b_i y = h_i \quad (i = 1, 2, \dots, N) \quad (11.89)$$

where  $a_i, b_i$  and  $h_i$  are unknown constants. The problem is overdetermined, in the sense that there exist more equations than needed for the unique determination of the unknowns  $x$  and  $y$ . The equations are said to form an *overdetermined system of equations*.

To find the *most probable values* of  $x$  and  $y$ , the method of least squares is used as follows:

Defining the 'error' of the  $i$ th equation as

$$e_i \equiv a_i x + b_i y - h_i, \quad (11.90)$$

we find the values of  $x$  and  $y$  which minimize the sum

$$S \equiv \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (a_i x + b_i y - h_i)^2. \quad (11.91)$$

Differentiating  $S$  partially with respect to  $x$  and with respect to  $y$  and equating both to zero, we obtain the normal equations

$$[a^2]x + [ab]y = [ah] \quad (11.92)$$

$$[ab]x + [b^2]y = [bh] \quad (11.93)$$

the solutions of which are given by the relations

$$\frac{x}{\begin{vmatrix} [ah] & [ab] \\ [bh] & [b^2] \end{vmatrix}} = \frac{y}{\begin{vmatrix} [a^2] & [ah] \\ [ab] & [bh] \end{vmatrix}} = \frac{1}{\begin{vmatrix} [a^2] & [ab] \\ [ab] & [b^2] \end{vmatrix}}. \quad (11.94)$$

In cases where weights are attributed to the equations, with the  $i$ th equation having a weight equal to  $w_i$ , the normal equations are

$$[wa^2]x + [wab]y = [wah] \quad (11.95)$$

$$[wab]x + [wb^2]y = [wbh] \quad (11.96)$$

and the solutions are suitably readjusted.

In order to find the errors in  $x$  and  $y$ , we define the residuals

$$d_i \equiv a_i x + b_i y - h_i, \quad (11.97)$$

and their standard deviation,

$$\sigma = \sqrt{\frac{[d^2]}{N-2}}. \quad (11.98)$$

The errors  $\delta x$  and  $\delta y$  in  $x$  and  $y$  are given by the relations:

$$\frac{(\delta x)^2}{[b^2]} = \frac{(\delta y)^2}{[a^2]} = \frac{\sigma^2}{\begin{vmatrix} [a^2] & [ab] \\ [ab] & [b^2] \end{vmatrix}}. \quad (11.99)$$

### Example 11.32

Find the most probable solutions of the equations

$$x + y = 5.3 \quad 2x - y = 0.8 \quad x - y = -0.6 \quad 3x + 2y = 11.2$$

and their errors.

We construct the following table:

$i$	$a_i$	$b_i$	$h_i$	$a_i^2$	$b_i^2$	$a_i b_i$	$a_i h_i$	$b_i h_i$
1	1	1	5.3	1	1	1	5.3	5.3
2	2	-1	0.8	4	1	-2	1.6	-0.8
3	1	-1	-0.6	1	1	-1	-0.6	0
4	3	2	11.2	9	4	6	33.6	22.4
Sums	7	1	16.7	15	7	4	39.9	27.5

The normal Eq. (11.94) give  $15x + 4y = 39.9$  the solutions of which are  $4x + 7y = 27.5$   
 $x = 1.90$   $y = 2.84$ . These are the most probable solutions of the equations given.  
 The residuals of the four equations are, respectively:

$$d_1 = x + y - 5.3 = -0.56 \quad d_2 = 2x - y - 0.8 = 0.16$$

$$d_3 = x - y + 0.6 = -0.34 \quad d_4 = 3x + 2y - 11.2 = 0.18$$

Therefore,  $[d^2] = 0.487$  and  $\sigma = \sqrt{\frac{[d^2]}{N-2}} = \sqrt{\frac{0.487}{2}} = 0.494$ .

$$\frac{(\delta x)^2}{7} = \frac{(\delta y)^2}{15} = \frac{\sigma^2}{\begin{vmatrix} 15 & 4 \\ 4 & 7 \end{vmatrix}} = \frac{0.244}{89} = 0.002742$$

and, finally,  $\delta x = 0.139$   $\delta y = 0.203$ .

The most probable values of  $x$  and  $y$  are, therefore,  $x = 1.90 \pm 0.14$ ,  
 $y = 2.84 \pm 0.20$ .



**Example 11.33 [E]**

Using Excel<sup>®</sup>, find the most probable solutions of the equations

$$x + y = 5.3 \quad 2x - y = 0.8 \quad x - y = -0.6 \quad 3x + 2y = 11.2$$

as well as their errors.

We enter the 4 coefficients  $a_i$  in column A (cells A3 to A6), the 4 coefficients  $b_i$  in column B (cells B3 to B6) and the  $h_i$ 's in column C (cells C3 to C6).

We evaluate [aa] = **sumsq(A3:A6)** = 15 in cell B9.

We evaluate [bb] = **sumsq(B3:B6)** = 7 in cell D9.

We evaluate [ab] = **sumproduct(A3:A6;B3:B6)** = 4 in cell F9.

We evaluate [ah] = **sumproduct(A3:A6;C3:C6)** = 39.9 in cell B11.

We evaluate [bh] = **sumproduct(B3:B6;C3:C6)** = 27.5 in cell D11.

Using Eq. (11.94), we find:  $x = 1.9022$  and  $y = 2.8416$ .

We calculate the values of  $d_i$  in column E: In cell E3 type **1.9022\*A3 + 2.8416\*B3 - C3** and press **ENTER**. This puts  $d_1$  in cell E1. We fill down to cell E6.

In cell B13 we calculate **sumsq(E3:E6)**.

Using Eq. (11.99), we find  $\delta x = 0.1384$  and  $\delta y = 0.2026$ .

The most probable values of  $x$  and  $y$  are, therefore,  $x = 1.90 \pm 0.14$ ,  $y = 2.84 \pm 0.20$ .

**Example 11.34 [O]**

Using Origin<sup>®</sup>, find the most probable solutions of the equations

$$x + y = 5.3 \quad 2x - y = 0.8 \quad x - y = -0.6 \quad 3x + 2y = 11.2$$

as well as their errors.

We enter the 4 coefficients  $a_i$  in column A (cells A1 to A4), the 4 coefficients  $b_i$  in column B (cells B1 to B4) and the  $h_i$ 's in column C (cells C1 to C4).

Using **Column > Set Column Values...** we evaluate  $a^2$ ,  $b^2$ ,  $ab$ ,  $ah$  and  $bh$  in columns D, E, F, G and H, respectively.

In each column we highlight those cells containing data and, using the  $\Sigma$  operator, we evaluate:

$$\begin{aligned} [\text{aa}] &= 15 \text{ in cell M2} & [\text{bb}] &= 7 \text{ in cell M3} & [\text{ab}] &= 4 \text{ in cell M4} \\ [\text{ah}] &= 39.9 \text{ in cell M5} & [\text{bh}] &= 27.5 \text{ in cell M6.} \end{aligned}$$

Using Eq. (11.94), we find:  $x = 1.9022$  and  $y = 2.8416$ .

We calculate the values of  $d_i^2$  in column J: We highlight column J and, using **Column > Set Column Values...**, we evaluate **(1.9022\*col(A) + 2.8416\*col(B) - col(C))^2** in column J. Summing these values, we find [dd] = 0.48708. This value gives  $\sigma = 0.4935$ .

Using Eq. (11.99), we find  $\delta x = 0.1384$  and  $\delta y = 0.2026$ .

The most probable values of  $x$  and  $y$  are, therefore,  $x = 1.90 \pm 0.14$ ,  $y = 2.84 \pm 0.20$ .

### Example 11.35 [P]

Using Python, find the most probable solutions of the equations

$$x + y = 5.3 \quad 2x - y = 0.8 \quad x - y = -0.6 \quad 3x + 2y = 11.2$$

as well as their errors.

```

from __future__ import division
import math
import numpy as np

# Enter the values of the coefficients a, b and h:
a = np.array([1, 2, 1, 3])
b = np.array([1, -1, -1, 2])
h = np.array([5.3, 0.8, -0.6, 11.2])

# Evaluation
AA = sum(a**2)
BB = sum(b**2)
AB = sum(a*b)
AH = sum(a*h)
BH = sum(b*h)

DENOM = AA*BB-AB*AB
DETX = AH*BB-AB*BH
DETY = BH*AA-AB*AH

x = DETX/DENOM
y = DETY/DENOM

d = x*a + y*b - h
S = math.sqrt(sum(d**2)/2)

DX = S*math.sqrt(BB/DENOM)
DY = S*math.sqrt(AA/DENOM)

# Results
print("Value of x:", x)
print("Value of y:", y)
print("Standard error in x:", DX)
print("Standard error in y:", DY)

```

```
Value of x: 1.90224719101
Value of y: 2.84157303371
Standard error in x: 0.1384007891490002
Standard error in y: 0.2025980103399658
```

**Example 11.36 [R]**

Using R, find the most probable solutions of the equations

$$x + y = 5.3 \quad 2x - y = 0.8 \quad x - y = -0.6 \quad 3x + 2y = 11.2$$

as well as their errors.

```
# Enter data vectors:
> a <- c(1, 2, 1, 3)
> b <- c(1, -1, -1, 2)
> h <- c(5.3, 0.8, -0.6, 11.2)

# Calculate sums of products:
> AA = sum(a^2)
> BB = sum(b^2)
> AB = sum(a*b)
> AH = sum(a*h)
> BH = sum(b*h)

# Calculate determinants:
> DENOM = AA*BB-AB*AB
> DETX = AH*BB-AB*BH
> DETY = BH*AA-AB*AH

# Find x and y:
> x = DETX/DENOM
> x
[1] 1.902247
> y = DETY/DENOM
> y
[1] 2.841573

# Calculate  $\sigma$ :
> d = 1.902247*a+2.841573*b-h
> S = sqrt(sum(d^2)/2)
> S
[1] 0.493497
```

```
# Calculate errors in x and y:
> DX = S*sqrt(BB/DENOM)
> DX
[1] 0.1384008
> DY = S*sqrt(AA/DENOM)
> DY
[1] 0.202598
```

The final results are:  $x = 1.90 \pm 0.14$ ,  $y = 2.84 \pm 0.20$ .

### 11.9.2 Equations in Three Variables

Let the linear equations involve three variables,  $x$ ,  $y$  and  $z$ . Given are  $N > 3$  equations

$$a_i x + b_i y + c_i z = h_i \quad (i = 1, 2, \dots, N) \quad (11.100)$$

where  $a_i$ ,  $b_i$ ,  $c_i$  and  $h_i$  are unknown constants. The problem is overdetermined, in the sense that there exist more equations than needed for the unique determination of the unknowns  $x$ ,  $y$  and  $z$ . The equations are said to form an overdetermined system of equations.

To find the *most probable values* of  $x$ ,  $y$  and  $z$ , we work as in Sect. 11.9.1. The equations derived are just presented here:

$$a_i x + b_i y + c_i z = h_i \quad (i = 1, 2, \dots, N) \quad (11.101)$$

$$S \equiv \sum_{i=1}^N (a_i x + b_i y + c_i z - h_i)^2 \quad (11.102)$$

The normal equations are

$$[a^2]x + [ab]y + [ac]z = [ah] \quad (11.103)$$

$$[ab]x + [b^2]y + [bc]z = [bh] \quad (11.104)$$

$$[ac]x + [bc]y + [c^2]z = [ch] \quad (11.105)$$

and their solutions,

$$\begin{aligned} \overline{\begin{matrix} x \\ [ah] & [ab] & [ac] \\ [bh] & [b^2] & [bc] \\ [ch] & [bc] & [c^2] \end{matrix}} &= \overline{\begin{matrix} y \\ [a^2] & [ah] & [ac] \\ [ab] & [bh] & [bc] \\ [ac] & [ch] & [c^2] \end{matrix}} = \overline{\begin{matrix} z \\ [a^2] & [ab] & [ah] \\ [ab] & [b^2] & [bh] \\ [ac] & [bc] & [ch] \end{matrix}} \\ &= \overline{\begin{matrix} 1 \\ [a^2] & [ab] & [ac] \\ [ab] & [b^2] & [bc] \\ [ac] & [bc] & [c^2] \end{matrix}}. \end{aligned} \tag{11.106}$$

In cases where weights are attributed to the equations, with the  $i$ th equation having a weight equal to  $w_i$ , the normal equations are

$$[wa^2]x + [wab]y + [wac]z = [wah] \tag{11.107}$$

$$[wab]x + [wb^2]y + [wbc]z = [wbh] \tag{11.108}$$

$$[wac]x + [wbc]y + [wc^2]z = [wch] \tag{11.109}$$

and the solutions are suitably readjusted.

To find the errors  $\delta x$ ,  $\delta y$  and  $\delta z$  in the variables  $x$ ,  $y$  and  $z$ , we define

$$d_i \equiv a_i x + b_i y + c_i z - h_i \tag{11.110}$$

and

$$\sigma = \sqrt{\frac{[d^2]}{N - 3}}, \tag{11.111}$$

in which case we have the relations

$$\overline{\begin{matrix} (\delta x)^2 \\ [b^2] & [bc] \\ [bc] & [c^2] \end{matrix}} = \overline{\begin{matrix} (\delta y)^2 \\ [a^2] & [ac] \\ [ac] & [c^2] \end{matrix}} = \overline{\begin{matrix} (\delta z)^2 \\ [a^2] & [ab] \\ [ab] & [b^2] \end{matrix}} = \overline{\begin{matrix} \sigma^2 \\ [a^2] & [ab] & [ac] \\ [ab] & [b^2] & [bc] \\ [ac] & [bc] & [c^2] \end{matrix}}. \tag{11.112}$$

---

**Programs**

---

**Excel**

---

- Ch. 11. Excel—Least Squares—Overdetermined Equations—2 Variables
- Ch. 11. Excel—Least Squares—Overdetermined Equations—3 Variables
- Ch. 11. Excel—Least Squares—Smoothing—Adjacent Averaging
- Ch. 11. Excel—Least Squares Fit—Straight Line
- Ch. 11. Excel—Least Squares Fit—Straight Line—Weighted Points
- Ch. 11. Excel—Least Squares Fit—Straight Line Through Origin
- Ch. 11. Excel—Least Squares Fit—Straight Line Through Origin—Weighted Points

(continued)

(continued)

**Programs**

- 
- Ch. 11. Excel—Least Squares Fit—Parabola*
  - Ch. 11. Excel—Least Squares Fit—Cubic*
  - Ch. 11. Excel—Least Squares Fit—Curve of 4th Degree*
  - Ch. 11. Excel—Least Squares Fit—Curve of 5th Degree*
  - Ch. 11. Excel—Least Squares Fit—Curve of 6th Degree*
  - Ch. 11. Excel—Least Squares Fit—Exponential*
- 

**Origin**

- 
- Ch. 11. Origin—Least Squares—Overdetermined Equations—2 Variables*
  - Ch. 11. Origin—Least Squares—Overdetermined Equations—3 Variables*
  - Ch. 11. Origin—Least Squares—Smoothing—Adjacent Averaging and Savitzki-Golay*
  - Ch. 11. Origin—Least Squares Fit—Straight Line*
  - Ch. 11. Origin—Least Squares Fit—Straight Line—Weighted Points*
  - Ch. 11. Origin—Least Squares Fit—Straight Line Through Origin*
  - Ch. 11. Origin—Least Squares Fit—Straight Line Through Origin—Weighted Points*
  - Ch. 11. Origin—Least Squares Fit—Parabola*
  - Ch. 11. Origin—Least Squares Fit—Cubic*
  - Ch. 11. Origin—Least Squares Fit—Curve of 4th Degree*
  - Ch. 11. Origin—Least Squares Fit—Curve of 5th Degree*
  - Ch. 11. Origin—Least Squares Fit—Curve of 6th Degree*
  - Ch. 11. Origin—Least Squares Fit—Power*
  - Ch. 11. Origin—Least Squares Fit—Exponential*
  - Ch. 11. Origin—Least Squares Fit—Gaussian*
  - Ch. 11. Origin—Least Squares Fit—Poisson*
- 

**Python**

- 
- Ch. 11. Python—Least Squares—Overdetermined Equations—2 Variables*
  - Ch. 11. Python—Least Squares—Overdetermined Equations—3 Variables*
  - Ch. 11. Python—Least Squares—Smoothing—Savitzki-Golay*
  - Ch. 11. Python—Least Squares Fit—Straight Line*
  - Ch. 11. Python—Least Squares Fit—Straight Line—Weighted Points*
  - Ch. 11. Python—Least Squares Fit—Straight Line Through Origin*
  - Ch. 11. Python—Least Squares Fit—Straight Line Through Origin—Weighted Points*
  - Ch. 11. Python—Least Squares Fit—Parabola*
  - Ch. 11. Python—Least Squares Fit—Cubic*
  - Ch. 11. Python—Least Squares Fit—Curve of 4th Degree*
  - Ch. 11. Python—Least Squares Fit—Curve of 5th Degree*
  - Ch. 11. Python—Least Squares Fit—Curve of 6th Degree*
  - Ch. 11. Python—Least Squares Fit—Exponential*
- 

**R**

- 
- Ch. 11. R—Least Squares—Overdetermined Equations—2 Variables*
  - Ch. 11. R—Least Squares—Overdetermined Equations—3 Variables*
  - Ch. 11. R—Least Squares—Smoothing—Cubic Spline*
  - Ch. 11. R—Least Squares Fit—Straight Line*
  - Ch. 11. R—Least Squares Fit—Straight Line—Weighted Points*
  - Ch. 11. R—Least Squares Fit—Straight Line Through Origin*
  - Ch. 11. R—Least Squares Fit—Straight Line Through Origin—Weighted Points*
  - Ch. 11. R—Least Squares Fit—Parabola*
  - Ch. 11. R—Least Squares Fit—Cubic*
  - Ch. 11. R—Least Squares Fit—Curve of 4th Degree*
- 

(continued)

(continued)

**Programs**

Ch. 11. R—Least Squares Fit—Curve of 5th Degree

Ch. 11. R—Least Squares Fit—Curve of 6th Degree

Ch. 11. R—Least Squares Fit—Exponential

**Problems**

The reader is reminded of the fact that most scientific hand-held calculators have the possibility of evaluating the quantities mentioned in this book. The  $n$  pairs of values  $x, y$  are entered using the key  $\Sigma +$  (and  $\Sigma -$  for correcting erroneous entries). When entering the data is complete, the calculator's memories contain the quantities  $n, [x], [y], [x^2], [y^2], [xy], \bar{x}, \bar{y}, s_x, \sigma_{\bar{x}}, s_y, \sigma_{\bar{y}}$ , which may be used to evaluate magnitudes such as the ones mentioned so far in this book. Some scientific calculators also return the parameters of the regression line,  $\alpha, \lambda$  and  $r$ . Of course, statistics calculators offer even more.

11.1 [E.O.P.R.] Given the experimental results

$x_i$	0	1	2	3	4	5	6	7	8	9
$y_i$	3.8	11.3	18.5	24.5	31.1	37.7	45.8	52.7	60.5	66.2

- (a) In a figure draw the straight line that you consider to fit better to these points.
- (b) Find the least-squares straight line  $y = a + \lambda x$  fitted to the points. Draw this line in the figure drawn in (a).
- (c) Find the coefficient of correlation  $r$  of the least-squares line.
- (d) What are the errors in the values of  $a$  and  $\lambda$  ?

11.2 [E.O.P.R.] Measurements of  $y$  as a function of  $x$  gave the results

$x_i$	0.8	2.2	3.6	4.8	6.2	7.8	9.0
$y_i$	8.0	6.8	6.1	5.2	4.4	4.0	2.8

- (a) Find the parameters  $a \pm \delta a$  and  $\lambda \pm \delta \lambda$  of the straight line  $y = a + \lambda x$  fitted to these data using the method of least squares. Assuming that  $a$  and  $\lambda$  are correlated to a negligible degree, so that from the relation  $y = (a \pm \delta a) + (\lambda \pm \delta \lambda)x$  the error in  $y$  to be given by  $\delta y = \sqrt{(\delta a)^2 + x^2(\delta \lambda)^2}$ , find:
  - (b) the value of  $y$  and its error  $\delta y$  for  $x = 5$  and
  - (c) for which value of  $x$  (and its error,  $\delta x$ )  $y$  is equal to 0.

11.3 To the pairs of experimental values  $(x_i, y_i)$  ( $i = 1, 2, \dots, N$ ) we wish to fit a straight line  $y = a + \lambda x$ .

(a) Assuming that  $a$  is known with great accuracy, show that the method of least squares gives  $\lambda = \frac{[xy] - a[x]}{[x^2]}$ .

(b) Assuming that  $\lambda$  is known with great accuracy, show that the method of least squares gives  $a = \frac{1}{N}([y] - \lambda[x])$ .

11.4 **[E.O.P.R.]** Using the method of least squares, fit a parabolic curve to the experimental points

$x_i$	0	0.1	0.2	0.3	0.4	0.5	0.6
$y_i$	2.8	4.2	8.4	16.0	27.5	41.9	59.3

11.5 If in Example 11.4 the function was  $y = (A \sin \omega t + B \cos \omega t)e^{-\kappa t}$ , where  $\omega$  and  $\kappa$  are known, what does the method of least squares give for the parameters  $A$  and  $B$ ?

11.6 A sample contains two radioisotopes, whose decay constants,  $\lambda_1$  and  $\lambda_2$ , are known with great accuracy. If  $N_{01}$  and  $N_{02}$  are the initial numbers of nuclei of the two isotopes, the total activity of the sample at time  $t$  is  $R = R_1 + R_2$ ,

$$R(t) = \lambda_1 N_{01} e^{-\lambda_1 t} + \lambda_2 N_{02} e^{-\lambda_2 t}.$$

From  $N$  measurements  $(t_i, R_i)$ , find  $N_{01}$  and  $N_{02}$  by the method of least squares.

(*Suggestion:* For convenience, use the notation  $x_i \equiv \lambda_1 e^{-\lambda_1 t_i}$  and  $y_i \equiv \lambda_2 e^{-\lambda_2 t_i}$ . The values of  $x_i$  and  $y_i$  are known for every value  $t_i$ .)

11.7 **[E.O.P.R.]** The viscosity of water,  $\eta$  (in units of centipoise) varies with the temperature in the following way, as determined by measurements:

$t$ ( $^{\circ}\text{C}$ )	10	20	30	40	50	60	70
$\eta$	1.308	1.005	0.801	0.656	0.549	0.469	0.406

Assume that a relation of the form  $\eta = Ae^{\lambda/T}$  holds, where  $T(\text{K}) = t(^{\circ}\text{C}) + 273.15$  is the absolute temperature. Using  $x = 1/T$  as variable and the methods of curve fitting, determine  $A$  and  $\lambda$ . Find also the errors in these parameters.

11.8 In an experiment for the determination of the radius of the Earth,  $R$ , by measuring the acceleration of gravity as a function of height  $H$  above the surface of the Earth, the results were as follows:



$H$ (m)	0	500	1000	1500	2000	2500	3000
$g$ (m/s <sup>2</sup> )	9.8070	9.8051	9.8044	9.8020	9.8015	9.7990	9.7976

The theoretical relation for the acceleration of gravity as a function of height is  $g = \frac{g_0}{(1+H/R)^2}$ . From this we have  $\frac{1}{\sqrt{g}} = \frac{1}{\sqrt{g_0}} + \frac{1}{R\sqrt{g_0}}H$ . Putting  $x = H$ ,  $y = 1/\sqrt{g}$ ,  $a = 1/\sqrt{g_0}$  and  $\lambda = \frac{1}{R\sqrt{g_0}}$ , it follows that  $y = a + \lambda x$ .

Using the method of least squares determine  $a \pm \delta a$  and  $\lambda \pm \delta \lambda$ , and then  $g_0 \pm \delta g_0$  and  $R \pm \delta R$ .

This method, which would not give accurate results, assumes that  $g$  may be measured with sufficient accuracy and that its variation is due solely to the change in height.

- 11.9 The rolling resistance  $F$  for a vehicle was measured at various speeds  $v$  and found to be

$v_i$ (m/s)	1	2	3	4	5
$F_i$ (N)	15	20	30	40	60

Assuming a relation of the form  $F = a + \lambda v^2$  and using as variable  $x = v^2$ , find the least-squares straight line for  $F(x)$  and from it find the coefficients  $a$  and  $\lambda$ .

**[E.O.P.R.]** Using non-linear curve fitting, find the parabola of the form  $F = a + \lambda v^2$  that gives the best fit to the points.

- 11.10 The activity  $R(t)$  of a radon sample is initially equal to  $R_0$ . The variation of the ratio  $R(t)/R_0$  in measurements which were made at intervals of one day each from the other is:

$t_i$ (d)	0	1	2	3	4	5	6	7	8
$R(t)/R_0$	1	0.835	0.695	0.580	0.485	0.405	0.335	0.280	0.235

Assuming that it is  $R(t)/R_0 = e^{-\lambda t}$  and, therefore,  $\ln[R(t)/R_0] = -\lambda t$ , find the value of  $\lambda$  applying the method of least squares to the last relation.

**[E.O.P.R.]** Using non-linear curve fitting, find the curve  $R(t)/R_0 = e^{-\lambda t}$  that gives the best fit to the points.

- 11.11 **[E.O.P.R.]** Measurements of  $y$  as a function of  $x$  gave the following results:

$x_i$	2	6	8	12	16	18	22	28
$y_i$	2	4	8	8	10	14	16	18

- (a) Using the method of least squares, find the straight line  $y(x)$ , when  $x$  is considered to be the independent variable.
- (b) Using the method of least squares, find the straight line  $x(y)$ , when  $y$  is considered to be the independent variable.

- (c) Draw both lines in a graph.  
 (d) Show that both lines pass through the point  $(\bar{x}, \bar{y})$ .

- 11.12 **[E.O.P.R.]** The main measurements of the speed of light performed between 1900 and 1956 are given in the table below.

#	Researcher	$t$	$c$ (km/s)	#	Researcher	$t$	$c$ (km/s)
1	Rosa, Dorsey	1906	299 781	10	Houston	1950	299 775
2	Mercier	1923	299 782	11	Bol, Hansen	1950	299 789.3
3	Michelson	1926	299 796	12	Aslakson	1951	299 794.2
4	Karolus, Mittelstaedt	1928	299 778	13	Rank, Ruth, Ven der Sluis	1952	299 776
5	Michelson, Pease, Pearson	1932	299 774	14	Froome	1952	299 792.6
				15	Florman	1954	299 795.1
6	Huettel	1940	299 768	16	Rank, Shearer, Wiggins	1954	299 789.8
7	Anderson	1941	299 776				
8	Bergstrand	1950	299 792.7	17	Edge	1956	299 792.9
9	Essen	1950	299 792.5				

Using the method of least squares, fit a straight line of the form  $c = a + \lambda(t - 1956)$  to the measurements, where  $t$  is the year each measurement was performed. Investigate the possibility that the results support the hypothesis that the speed of light varies with time.

- 11.13 From the equations

$$3x + 2y = 5.8 \quad x - 4y = 1.8 \quad 4x - 6y = 3.8$$

find the most probable values of  $x$  and  $y$ .

- 11.14 Find the most probable values of  $x$ ,  $y$  and  $z$ , as these are determined from the equations:

$$x + 2y + 3z = 12.1 \quad 2x - 2y + 3z = 3.2 \quad x + 6y - 6z = 15.1 \\ 3x + 2y = 14.9.$$

- 11.15 Find the most probable values of  $x$  and  $y$ , and their errors, as these are determined by applying the method of least squares to the equations

$$x + 2y = 31.8 \quad x - 4y = -4.8 \quad x - 2y = 3.6 \quad 2x + 6y = 67.2.$$

## References

1. For a complete analysis of the problem of fitting polynomial curves to experimental points, see, for example, F.B. Hildebrand, *Introduction to Numerical Analysis*, 2nd edn. (McGraw-Hill Inc., New York, 1974). Chap. 7
2. For the general expression, see, for example, Athanasios Papoulis, *Probability, Random Variables and Stochastic Processes*, 3rd edn. (McGraw-Hill Co. Inc., New York, 1991). Chap. 5, Sect. 5-2