Chapter 6

Least squares I

6.1 Introduction

Establishing and understanding the relationship between quantities are principal goals in the physical sciences. As examples, we might be keen to know how the:

- size of a crystal depends on the growth time of the crystal;
- output intensity of a light emitting diode varies with the emission wavelength;
- amount of light absorbed by a chemical species depends on the species concentration;
- electrical power supplied by a solar cell varies with optical power incident on the cell;
- viscosity of an engine oil depends upon the temperature of the oil;
- rate of flow of a fluid through a hollow tube depends on the internal diameter of the tube.

Once an experiment is complete and the data presented in the form of an x–y graph, an examination of the data assists in answering important qualitative questions such as: Is there evidence of a clear trend in the data? If so, is that trend linear, and do any of the data conflict with the general trend? A qualitative analysis often suggests which quantitative methods of analysis to apply.

There are many situations in the physical sciences in which prior knowledge or experience suggests a relationship between measured quantities. Perhaps we are already aware of an equation which predicts how one quantity depends on another. Our goal in this situation might be to discover how well the equation can be made to 'fit' the data.

As an example, in an experiment to study the variation of the viscosity of engine oil between room temperature and 100 °C, we observe that the viscosity of the oil decreases with increasing temperature, but we would like to know more.

- What is the quantitative relationship between viscosity and temperature? Does viscosity decrease linearly with increasing temperature, or in some other way?
- Can we find an equation that represents the variation of viscosity with temperature? Perhaps this would allow us to predict values of viscosity at any given temperature, i.e. permit interpolation between measured temperatures.
- Is it possible to use the data to test a theory which predicts how viscosity should depend on temperature?
- Can a useful parameter be estimated from viscosity–temperature data, such as the change in viscosity for a temperature rise of 1 °C?

A powerful and widely used method for establishing a quantitative relationship between quantities is that of *least squares*, also known as *regression*, and it is this method that we will focus upon in this chapter.

The method of least squares is computationally demanding, especially if there are many data to be considered. We will use Excel to assist with least squares analysis at various points in this chapter after the basic principles have been considered.

6.2 The equation of a straight line

It is quite common for there to be a linear relationship between two quantities measured in an experiment. The data obtained through an experiment devised to study the relationship between two quantities are routinely represented as points on an x–y graph. By fitting a straight line to the data, a quantitative expression may be found that relates the two quantities.

What we would really like to do is find the equation that describes the *best* line that passes through (or at least close to) the data. We assume that the equation of the 'best line' is the closest we can come to finding the relationship between the quantities with the data available.

Figure 6.1 shows a straight line drawn on an x–y graph. The y value, y_i, of any point on the line is related to the corresponding x value, x_i, by the equation[1]

[1] It is common to find the equation of a straight line written in other ways, such as $y_i = mx_i + c$, $y_i = mx_i + b$ or $y_i = b_0 + b_1x_i$.

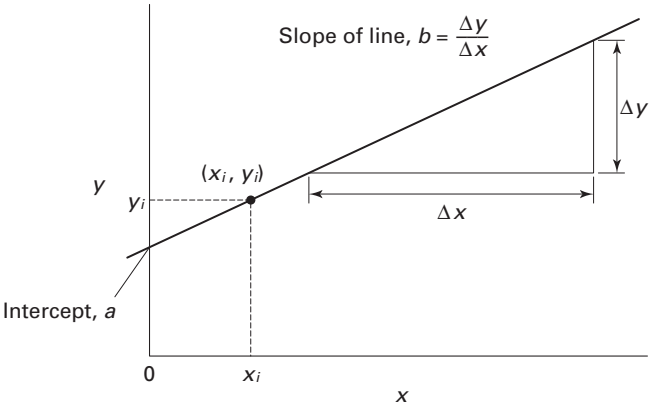

Figure 6.1. Straight line relationship between x and y.

$$y_i = a + bx_i,$$

where a is the intercept and b is the slope of the line.

If all the x–y data gathered in an experiment were to lie along a straight line, there would be no difficulty in determining a and b and our discussion would end here. We would simply use a rule to draw a line through the points. Where the line intersects the y axis at $x = 0$ gives a, and b is found by dividing Δy by Δx, as indicated in figure 6.1.

In situations in which 'real' data are considered, even if the underlying relationship between x and y is linear, it is highly unlikely that all the points will lie on a straight line, as sources of error act to scatter the data. So how do we find the best line through the points in circumstances where data are scattered?

6.2.1 The 'best' straight line through x–y data

When dealing with experimental data, we commonly plot the quantity that we are able to control on the x axis. This quantity is often referred to as the *independent* (or the *predictor*) variable. A quantity that changes in response to changes in the independent variable is referred to as the *dependent* (or the *response*) variable, and is plotted on the y axis.

As an example, consider an experiment in which the velocity of sound in air is measured at different temperatures. Here temperature is the independent variable and velocity is the dependent variable. Table 6.1 shows temperature–velocity data for sound travelling through dry air.

Table 6.1. *Temperature–velocity data for sound travelling through dry air.*

Temperature, θ (°C)	Velocity, v (m/s) (± 5 m/s)
–13	322
0	335
9	337
20	346
33	352
50	365

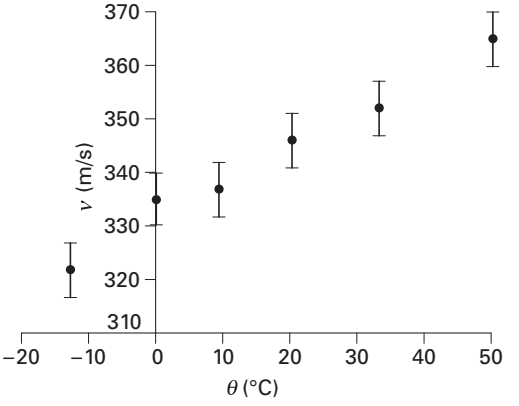

Figure 6.2. An *x*–*y* graph showing velocity of sound versus temperature.

The data in table 6.1 are plotted in figure 6.2. Error bars are attached to each point to indicate the standard uncertainty in the values of velocity.

From an inspection of figure 6.2, it appears reasonable to propose that there is a linear relationship between velocity, v, and temperature, θ. This relationship can be written as

$$v = A + B\theta, \tag{6.1}$$

where A and B are constants.

Finding the best straight line through the points *means* finding best estimates for A and B, as it is these two parameters that describe the relationship between v and θ. We can draw a line 'by eye' through the points using a transparent plastic rule and from that line estimate A and B. The difficulty with this approach is that, when data are scattered, it is difficult to find the position of the best line through the points.

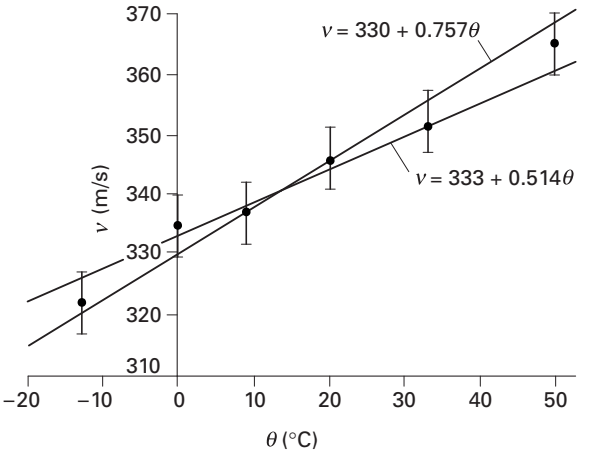

Figure 6.3. Two lines fitted to velocity of sound versus temperature data. The equation describing each line is shown.

Figure 6.3 shows two attempts at drawing a line through the velocity versus temperature data along with the equation that describes each line. Can either of the two lines be regarded as the best line through the points? If the answer is no, then how *do* we find the best line?

The guesswork associated with drawing a line by eye through data can be eliminated by applying the technique of least squares.

6.2.2 Unweighted least squares

To find the best line through x–y data, we need to decide upon a numerical measure of the 'goodness of fit' of the line to the data. One approach is to take that measure to be the 'sum of squares of residuals', which we will discuss for the case where there is a linear relationship between x and y. The least squares method discussed in this section rests on the assumptions described in table 6.2.

Figure 6.4 shows a line drawn through the x–y data. The vertical distances from each point to the line are labelled Δy_1, Δy_2, Δy_3, etc., and are referred to as the *residuals* (or *deviations*). A residual is defined as the difference between the observed y value and the y value on the line for the same x value. Referring to the ith observed y value as y_i, and the ith predicted value found using the equation of the line as $\hat{y}_i$, the residual, Δy_i, is written

Table 6.2. *Assumptions upon which the unweighted least squares method is based.*

Assumption	Comment
The y values are influenced by random errors only.	Any measurement is affected by errors introduced by such sources as noise and instrument resolution (see chapter 5). Systematic errors cannot be accounted for using least squares.
The y values measured at a particular value of x have a normal distribution.	If errors in measured values are normally distributed, then measured values will exhibit the characteristics of the normal distribution (see chapter 3).
The standard deviation of y values at all values of x are the same.	If this assumption is true then every point on an x–y graph is as reliable as any other and, in using the least squares method to fit a line to data, one point must not be favoured or 'weighted' more than any other point. This is referred to as *unweighted* least squares.
There are no errors in the x values.	This is the most troublesome assumption. In most circumstances the x quantity is controlled in an experiment. It is therefore likely to be less influenced by random errors than the y quantity. But this is not always true. We will consider one situation in which the assumption is not valid, i.e. where the errors in the x values are much larger than the errors in y values.

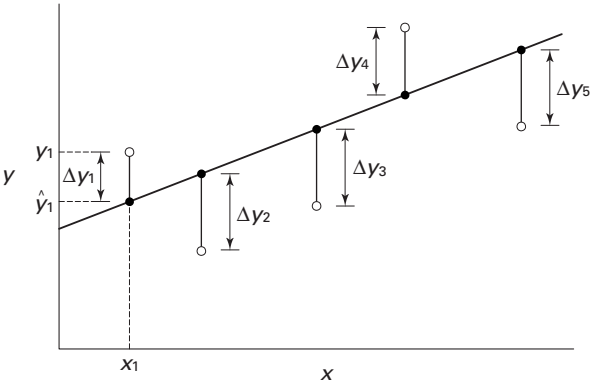

Figure 6.4. An x–y graph showing residuals.

$$\Delta y_i = y_i - \hat{y}_i, \tag{6.2}$$

where

$$\hat{y}_i = a + bx_i. \tag{6.3}$$

We propose that the underlying or 'true' relationship between x and y is given by

$$y = \alpha + \beta x, \tag{6.4}$$

where α is the true intercept and β is the true slope. We can never know α and β exactly, but best estimates of α and β, written as a and b respectively, are obtained by applying the 'Principle of Maximum Likelihood' as discussed in appendix 3. The outcome of applying this principle is that the best line is given by values of a and b which minimise the *sum of the square of the residuals*, *SSR*. *SSR* is given by

$$SSR = \sum (y_i - \hat{y}_i)^2, \tag{6.5}$$

where $\hat{y}_i$ is given by equation 6.3.

In principle, values of a and b that minimise *SSR* could be found by trial and error, or by a systematic numerical search using a computer. However, when a straight line is fitted to data, an equation for the best line can be found analytically.

If the assumptions in table 6.2 are valid, a and b are given by[2]

$$a = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} \tag{6.6}$$

and

$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}, \tag{6.7}$$

where n is the number of data points and each summation is carried out between $i = 1$ and $i = n$.

Example 1

Table 6.3 contains x–y data which are shown plotted in figure 6.5.

Using the data in table 6.3:

(i) find the value for the slope and intercept of the best line through the points;
(ii) draw the line of best fit through the points;
(iii) calculate the sum of squares of residuals, *SSR*.

[2] For derivations of equations 6.6 and 6.7 see section A3.2 in appendix 3.

ANSWER

To calculate a and b we need the sums appearing in equations 6.6 and 6.7, namely $\sum x_i$, $\sum y_i$, $\sum x_i y_i$, and $\sum x_i^2$. Many pocket calculators are able to calculate these quantities[3] (in fact, some pocket calculators are able to perform unweighted least squares fitting to give a and b directly).

A word of caution here: as there are many steps in the calculations of a and b, it is advisable *not* to round numbers in the intermediate calculations,[4] as rounding can significantly influence the values of a and b.

(i) Using the data in table 6.3 we find that $\sum x_i = 30$, $\sum y_i = 284$, $\sum x_i y_i = 1840$ and $\sum x_i^2 = 220$. Substituting these values into equations 6.6 and 6.7 (and noting that the number of points, $n = 5$) gives

$$a = \frac{220 \times 284 - 30 \times 1840}{5 \times 220 - (30)^2} = 36.4,$$
$$b = \frac{5 \times 1840 - 30 \times 284}{5 \times 220 - (30)^2} = 3.4.$$

(ii) The line of best fit through the data in table 6.3 is shown in figure 6.6.

(iii) The squares of residuals and their sum, *SSR*, are shown in table 6.4.

Exercise A

Use least squares to fit a straight line to the velocity versus temperature data in table 6.1. Calculate the intercept, a, the slope, b, of the line and the sum of squares of the residuals, *SSR*.

Table 6.3. *Linearly related x–y data.*

x	y
2	43
4	49
6	59
8	63
10	70

[3] If there are many data, a spreadsheet is preferred to a pocket calculator for calculating the sums.

[4] This problem is acute when a calculation involves subtracting two numbers that are almost equal, as can happen in the denominator of equations 6.6 and 6.7. Premature rounding can cause the denominator to tend to zero, resulting in very large (and very likely incorrect) values for a and b.

Table 6.4. *Calculation of sum of squares of residuals*

x_i	y_i	$\hat{y}_i = 36.4 + 3.4x_i$	$(y_i - \hat{y}_i)^2$
2	43	43.2	0.04
4	49	50.0	1.00
6	59	56.8	4.84
8	63	63.6	0.36
10	70	70.4	0.16
			$SSR = 6.4$

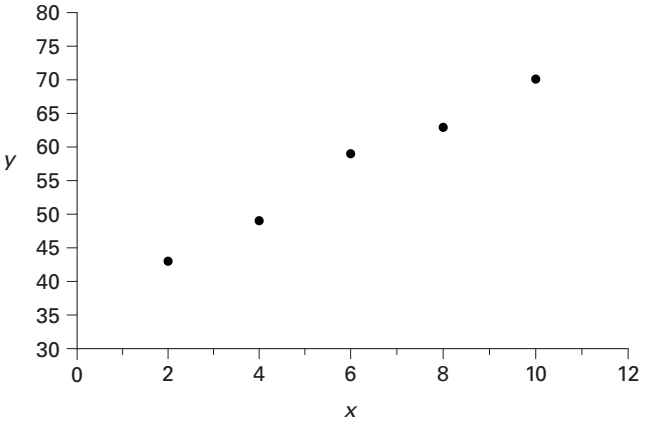

Figure 6.5. Linearly related x–y data.

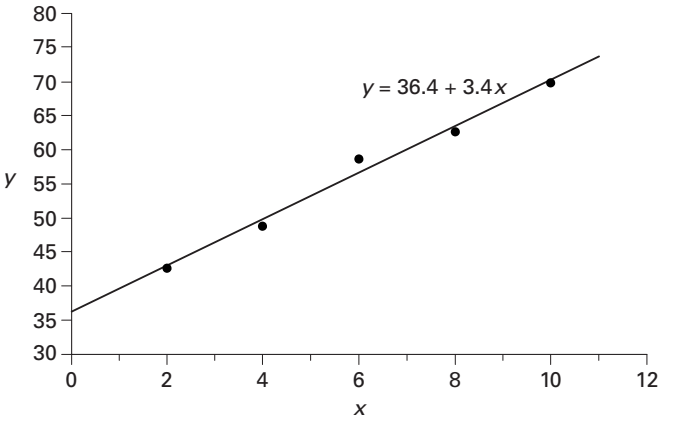

Figure 6.6. Line of best fit through data given in table 6.3.

6.2.3 Trendline in Excel

Excel may be used to add the best straight line to an x–y graph using the Add Trendline option. Excel uses equations 6.6 and 6.7 to determine the intercept and slope. Section 2.7.1 describes plotting an x–y graph using Excel and how a line of best fit can be added using Trendline.

Advantages of using Excel's Trendline include the following.

(i) It requires that data be plotted first so that we are encouraged to consider whether it is *reasonable* to draw a straight line through the points.
(ii) The best line is added automatically to the graph.
(iii) The equation of the best line can be displayed if required.
(iv) Excel is capable of drawing the best line through points for relationships between x and y other than linear, such as a logarithmic or power relationship.
(v) If data are changed then the points on the graph, the line of best and the equation of the best line are updated and displayed immediately, as the graph and the line of best fit are linked 'dynamically' to the data.

Though excellent for determining and displaying the best line through points, Trendline does not,

(i) give the standard uncertainties in a and b;
(ii) allow 'weighted fitting'. This is required when there is evidence to suggest that some x–y values are more reliable than others. In this situation the best line should be 'forced' to pass close to the more reliable points. Weighted fitting is considered in section 6.10;
(iii) plot residuals. Residuals are extremely helpful for assessing whether it is appropriate to fit a straight line to data in the first place. Residuals are considered in section 6.7.

Despite the usefulness of the Trendline option in Excel, there are often situations in which we need to extract more from the data than just a and b. In particular, the uncertainties in a and b, expressed as the standard uncertainties in a and b, are important as they are required for the calculation of the coverage intervals for the intercept and slope. We consider uncertainties in intercept and slope next.

6.2.4 Uncertainty in a and b

One of the basic assumptions made when fitting a line to data using least squares is that the dependent variable is subject to random error. It is reasonable to expect therefore that a and b are themselves influenced by the errors in the dependent variable. A preferred way of expressing the uncertainties in a and b is in terms of

their respective standard uncertainties,[5] as this permits us to calculate coverage intervals for a and b. Standard uncertainties in a and b may be found using the ideas of propagation of uncertainties discussed in chapter 5. Provided the uncertainty in each y value is the same,[6] the standard uncertainties in a and b are given by s_a and s_b, where[7]

$$s_a = \frac{s\left(\sum x_i^2\right)^{1/2}}{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]^{1/2}}, \tag{6.8}$$

$$s_b = \frac{sn^{1/2}}{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]^{1/2}}, \tag{6.9}$$

and where s is the standard deviation of the observed y values about the fitted line. The calculation of s is similar to the calculation of the estimate of the population standard deviation, s, of univariate data given by equation 1.16.

Here, s is given by

$$s = \left[\frac{1}{n-2}\sum (y_i - \hat{y}_i)^2\right]^{1/2}. \tag{6.10}$$

Example 2

Absorbance–concentration data shown in table 6.5 were obtained during an experiment in which standard silver solutions were analysed by flame atomic absorption spectrometry.

Assuming that absorbance is linearly related to concentration of the silver solution, use unweighted least squares to find the intercept and slope of the best line through the points and the standard uncertainties in these quantities.

ANSWER

Regarding the concentration as the independent variable, x, and the absorbance as the dependent variable, y, we write

$$y = a + bx.$$

Using the data in table 6.5 we find, $\sum x_i = 105$ (ng/mL), $\sum y_i = 2.667$, $\sum x_i y_i = 57.585$ (ng/mL) and $\sum x_i^2 = 2275$ (ng/mL)^2.

[5] Many data analysis and statistics texts refer to the standard error or standard deviation of parameter estimates. Consistent with the vocabulary introduced in chapter 5, we will use the term *standard uncertainty* when describing the uncertainty in parameter estimates.

[6] If repeat measurements are made of y at a particular value of x, the standard deviation in the y values remains the same regardless of the value of x chosen.

[7] See appendix 4 for derivations of equations 6.8 and 6.9.

Table 6.5. *Data obtained in a flame atomic absorption experiment.*

Concentration (ng/mL)	Absorbance (arbitrary[8] units)
0	0.002
5	0.131
10	0.255
15	0.392
20	0.500
25	0.622
30	0.765

Using equations 6.6 and 6.7,

$$a = \frac{2275(\text{ng/mL})^2 \times 2.667 - 105(\text{ng/mL}) \times 57.585(\text{ng/mL})}{7 \times 2275\,(\text{ng/mL})^2 - (105(\text{ng/mL}))^2} = 4.286 \times 10^{-3},$$

$$b = \frac{7 \times 57.585\,(\text{ng/mL}) - 105(\text{ng/mL}) \times 2.667}{7 \times 2275\,(\text{ng/mL})^2 - (105(\text{ng/mL}))^2} = 2.511 \times 10^{-2}\text{mL/ng}.$$

In this example units have been included explicitly[9] in the calculation of a and b to emphasise that, in most situations in the physical sciences, we deal with quantities that have units and these must be carried through to the 'final answers' for a and b.

In order to use equations 6.8 and 6.9, first calculate s as given by equation 6.10.

Table 6.6 has been constructed to assist in the calculation of s.

Summing the values in the last column of table 6.6 gives

$$\sum (y_i - \hat{y}_i)^2 = 3.2686 \times 10^{-4}.$$

Using equation 6.10,

$$s = \left[\frac{1}{(7-2)} \times 3.2686 \times 10^{-4}\right]^{1/2} = 0.008085.$$

Now use equations 6.8 and 6.9 to find s_a and s_b:

[8] When values indicate the relative size of a quantity (in this case absorbance), but are not expressed in a recognised unit of measure, such as that based on the SI system, we speak of that the unit of measure on that scale as being 'arbitrary'. Arbitrary units are often found on the y-axis of a graph.

[9] In other examples in this chapter, units do not appear (for the sake of brevity) in the intermediate calculations of a and b or s_a and s_b.

Table 6.6. *Calculation of squares of residuals.*

x_i (ng/mL)	y_i (arbitrary units)	$\hat{y}_i = 4.286 \times 10^{-3} + 2.511 \times 10^{-2} x_i$	$(y_i - \hat{y}_i)^2$
0	0.002	0.004286	5.2245×10^{-6}
5	0.131	0.129857	1.3061×10^{-6}
10	0.255	0.255429	1.8367×10^{-7}
15	0.392	0.381000	1.2100×10^{-4}
20	0.500	0.506571	4.3184×10^{-5}
25	0.622	0.632143	1.0288×10^{-4}
30	0.765	0.757714	5.3082×10^{-5}

$$s_a = \frac{s\left(\sum x_i^2\right)^{1/2}}{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]^{1/2}} = \frac{0.008085 \times (2275)^{1/2}}{\left[7 \times 2275 - (105)^2\right]^{1/2}} = 5.509 \times 10^{-3},$$

$$s_b = \frac{sn^{1/2}}{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]^{1/2}} = \frac{0.008085 \times (7)^{1/2}}{\left[7 \times 2275 - (105)^2\right]^{1/2}} = 3.056 \times 10^{-4}\text{mL/ng}.$$

Using the properties of the normal distribution,[10] we can say that the true value for the intercept has a probability of approximately 0.7 of lying between $(4.3–5.5) \times 10^{-3}$, and $(4.3 + 5.5) \times 10^{-3}$, i.e. the 70% coverage interval for α is between approximately -1.2×10^{-3} and 9.8×10^{-3}. Similarly, the 70% coverage interval for β is between 2.480×10^{-2} mL/ng and 2.542×10^{-2} mL/ng.

We must admit to a misdemeanour in applying the normal distribution here: In this example we are dealing with a small number of values (seven only), so we should use the t distribution rather than the normal distribution when calculating coverage intervals for α and β. We discuss this further in section 6.2.6.

Exercise B

Calculate s_a and s_b for the data in table 6.3.

6.2.5 Least squares, intermediate calculations and significant figures

In this text we adopt the convention that uncertainties are rounded to two significant figures. a and b are then presented to the number of figures consistent with the magnitude of the uncertainties. Where a pocket calculator or a

[10] See section 3.5.

spreadsheet has been used to determine a and b, all intermediate calculations are held to the full internal precision of the calculator, rounding only occurring in the presentation of the final parameter estimates.

6.2.6 Coverage intervals for α and β

We are able to determine the coverage interval for the true intercept, α, and that of the true slope, β, of a straight line fitted to data. We adopt the 'rule of thumb' that whenever there are fewer than 30 data points, it is appropriate to use the t distribution rather than the normal distribution when quoting coverage intervals. The 95% coverage interval for α is written as

$$\alpha = a \pm t_{95\%,\nu} s_a, \tag{6.11}$$

where ν is the number of degrees of freedom, and $t_{95\%,\nu}$ is the critical t value corresponding to the 95% coverage level.[11]

When fitting a straight line, the experimental data are used to calculate a and b. By using the data to estimate two parameters, the number of degrees of freedom is reduced by two, so that ν is given by[12]

$$\nu = n - 2,$$

where n is the number of data.

Similarly, the 95% coverage interval for β is written

$$\beta = b \pm t_{95\%,\nu} s_b, \tag{6.12}$$

where a and b are calculated using equations 6.6 and 6.7 respectively, and s_a and s_b are calculated using equations 6.8 and 6.9.

When the X% coverage interval is required, $t_{95\%,\nu}$ is replaced in equations 6.11 and 6.12 by $t_{X\%,\nu}$. Table 2 in appendix 1 gives the values of t for various confidence levels, X%, and degrees of freedom, ν.

Example 3

Using information given in example 2, calculate the 95% coverage interval for α and β.

ANSWER

Relevant information from example 2:

[11] It would be quite acceptable to substitute the coverage factor, k, for $t_{95\%,\nu}$.

[12] See Devore (2007) for more information on degrees of freedom.

$a = 4.286 \times 10^{-3}, s_a = 5.5 \times 10^{-3},$
$b = 2.511 \times 10^{-2}\text{mL/ng}, s_b = 3.1 \times 10^{-4}\text{mL/ng},$
$\nu = n - 2 = 7 - 2 = 5.$

Using table 2 in appendix 1, $t_{95\%,5} = 2.571$.

From equation 6.11, the 95% coverage interval for α is

$$4.286 \times 10^{-3} \pm 2.571 \times 5.5 \times 10^{-3},$$

i.e. $\alpha = (4 \pm 14) \times 10^{-3}$.

Using equation 6.12,

$$\beta = (2.511 \times 10^{-2} \pm 2.571 \times 3.1 \times 10^{-4})\,\text{mL/ng},$$

i.e.

$$\beta = (2.511 \pm 0.080) \times 10^{-2}\,\text{mL/ng}.$$

Exercise C

(1) Calculate the 99% coverage intervals for α and β in example 2.
(2) The data in table 6.7 were obtained in an experiment to study the variation of the electrical resistance, R, with temperature, θ, of a tungsten wire.

Assuming the relationship between R and θ can be written $R = A + B\theta$, calculate:

(i) the values of the intercept, a, and slope, b, of the best line through the resistance–temperature data;
(ii) the standard uncertainties in a and b;
(iii) the 95% coverage intervals for A and B.

6.3 Excel's LINEST() function

Calculating a and b and their respective standard uncertainties by using equations 6.6 to 6.10 is tedious, especially if there are many x–y values to consider. It is possible to use a spreadsheet to perform the calculations and this lessens the effort considerably. An even quicker method of calculating a and b is to use the LINEST()

Table 6.7. *Variation of resistance of a tungsten wire with temperature.*

θ (°C)	1	4	10	19	23	28	34	40	47	60	66	78	82
R (Ω)	10.2	10.3	10.7	11.0	11.2	11.4	11.8	12.2	12.5	12.8	13.2	13.5	13.6

function in Excel. This function estimates the parameters of the line of best fit and returns those estimates into an array of cells in the spreadsheet. The LINEST() function is versatile and we will consider it again in chapter 7. For the moment we use it to calculate a, b, s_a and s_b. The syntax of the function is as follows.

LINEST(y values, x values, constant, statistics)

y values:	Give range of cells containing the y values.
x values:	Give range of cells containing the x values.
constant:	Set to True to fit the equation $y = a + bx$ to the data. If we want to force the line through the origin (0, 0) (i.e. fit the equation $y = bx$ to data) we set the constant to False. See problem 11 at the end of the chapter for a brief consideration of fitting the equation $y = bx$ to data.
statistics:	Set to True to calculate the standard uncertainties[13] in a and b.

Example 4

Consider the x–y data shown in sheet 6.1. Use the LINEST() function to find the parameters a, b, s_a and s_b for the best line through these data.

ANSWER

Data are entered into columns A and B of the Excel spreadsheet as shown in sheet 6.1. We require Excel to return a, b, s_a and s_b. To do this:

(1) Move the cursor to cell D3. With the left hand mouse button held down, pull down and across to cell E4. Release the mouse button. Values returned by the LINEST() function will appear in the four cells, D3 to E4.
(2) Type **=LINEST(B2:B9,A2:A9,TRUE,TRUE)**.
(3) Hold down the Crtl and Shift keys together, then press the Enter key.

Figure 6.7 shows part of the screen as it appears after the Enter key has been pressed. Labels have been added to the figure to identify a, b, s_a and s_b.

Sheet 6.1. *The x–y data for example 4.*

	A	B
1	x	y
2	1	−2.3
3	2	−8.3
4	3	−11.8
5	4	−15.7

[13] In fact, the LINEST() function is able to return other statistics, but we will focus on the standard uncertainties for the moment.

Sheet 6.1. (*cont.*)

	A	B
6	5	−20.8
7	6	−25.3
8	7	−34.2
9	8	−37.2

Exercise D

Consider the x–y data in table 6.8.

(i) Use the LINEST() function to determine the intercept and slope of the best line through the x–y data, and the standard uncertainty in intercept and slope.

(ii) Plot the data on an x–y graph and show the line of best fit.

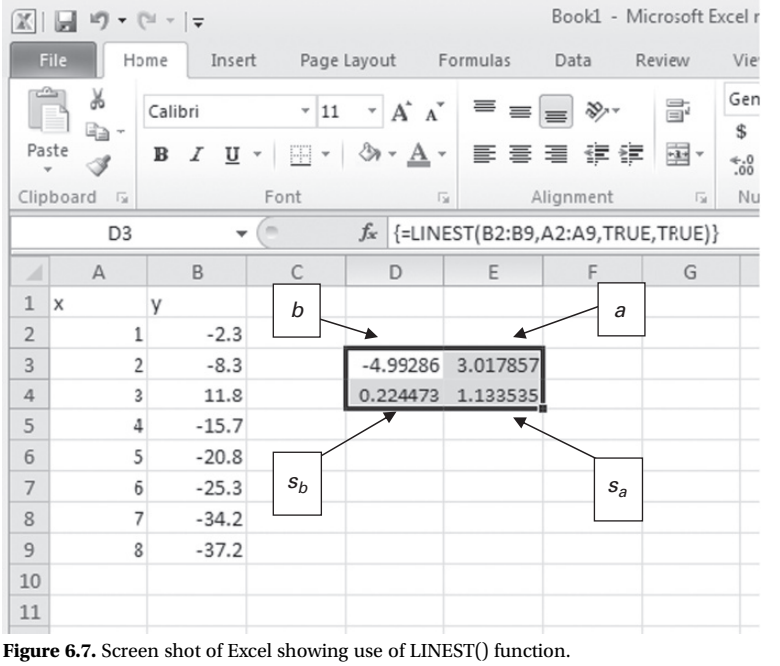

Figure 6.7. Screen shot of Excel showing use of LINEST() function.

Table 6.8. *The x–y data for exercise D.*

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
y	3.7	7.7	8.5	12.1	13.5	15.5	15.3	15.0	18.7

6.4 Using the line of best fit

There are several reasons why a straight line might be fitted to data. These include the following.

(i) A visual inspection of data presented as an x–y graph appears to indicate that there is a linear relationship between the dependent and independent variables. A line through the points helps to confirm this and may reveal anomalies such as points deviating systematically from the line.

(ii) A model or theory predicts a linear relationship between independent and the dependent variables. The best line through the points can provide strong evidence to support or refute the theory.[14]

(iii) The best line may be required for interpolation or extrapolation purposes. That is, a value of y may be found for a particular value of x, x_0, where x_0 lies within the range of x values used in the determination of the line of best fit (interpolation) or outside the range (extrapolation). The equation of the best line may also be used to determine an x value for a given y value.

We consider the use of the line of best fit next, but will return to the important matter of comparing models and data in later sections of this chapter, and again in chapter 7.

6.4.1 Comparing a 'physical' equation to $y = a + bx$

In some situations, a and b derived from fitting a straight line to experimental data can be directly related to a physical constant or parameter. In fact, it may be that something is known of the 'theoretical' relationship between variables prior to performing the experiment and that the main purpose of the experiment is to establish the value of a particular physical constant featuring in an equation derived from theory.

Take, as an example, a thermal expansion experiment in which the length of a rod is measured at different temperatures. Studies on the expansion of solids indicates that the relationship between the length of the rod, l (in metres), at temperature, θ (in degrees Celsius) may be written as

[14] In section 6.7 we will see how the scatter of residuals can provide convincing evidence for the appropriateness, or otherwise, of fitting a straight line to data.

$$l = l_o(1 + \alpha\theta), \tag{6.13}$$

where l_o is the length of the rod at 0 °C, and α is the temperature coefficient of expansion.

The right hand side of equation 6.13 can be expanded to give

$$l = l_o + l_o\alpha\theta. \tag{6.14}$$

Equation 6.14 is of the form $y = a + bx$, where $l = y$ and $\theta = x$. Comparing equation 6.14 to the equation of a straight line reveals,

$$a = l_o \text{ and } b = l_o\alpha.$$

It follows that

$$\alpha = \frac{b}{a}, \tag{6.15}$$

i.e. the ratio, b/a, gives the temperature coefficient of expansion of the material being studied. It would be usual to compare the value of α obtained through analysis of the length–temperature data with values reported by other experimenters who have studied the same, or similar, materials.

Exercise E

The pressure, P, at the bottom of a water tank is related to the depth of water, h, in the tank by the equation

$$P = \rho g h + P_A, \tag{6.16}$$

where P_A is the atmospheric pressure, g is the acceleration due to gravity, and ρ is the density of the water.

In an experiment, P is measured as the depth of water, h, in the tank increases. Let us assume equation 6.16 is valid.

(i) What would you choose to plot on each axis of an x–y graph in order to obtain a straight line?

(ii) How are the intercept and slope of that line related to ρ, g, and P_A in equation 6.16?

6.4.1.1 Uncertainties in parameters which are functions of *a* and *b*

The thermal expansion example discussed in the previous section brings up an important question: How may we establish the uncertainty in the temperature coefficient of expansion, α, given the uncertainties in a and b? α is a function of a and b, so at first sight it seems reasonable to apply the usual relationship for propagation of uncertainties (as derived in appendix 2) to find the standard uncertainty s_α in α using

$$s_\alpha^2 = \left(\frac{\partial \alpha}{\partial a}\right)^2 s_a^2 + \left(\frac{\partial \alpha}{\partial b}\right)^2 s_b^2. \tag{6.17}$$

However, equation 6.17 is only valid if there is no correlation between the errors in a and b. It turns out that the errors in a and b *are* correlated[15] and so we must consider this matter in more detail.

The line of best fit (for an unweighted fit) always passes through the point $(\bar{x}, \bar{y})$, where[16]

$$\bar{x} = \frac{\sum x_i}{n} \quad \text{and} \quad \bar{y} = \frac{\sum y_i}{n},$$

Using these relationships, we can write the intercept, a, as

$$a = \bar{y} - b\bar{x}. \tag{6.18}$$

Any equation which is a function of a and b (such as equation 6.15) can now be written as a function of $\bar{y}$ and b, by replacing a by $\bar{y} - b\bar{x}$. The advantage in this is that the errors in $\bar{y}$ and b are *not* correlated. This allows us to apply equation 5.33 to find the uncertainty in a parameter whose calculation involves both a and b.

As an example, consider the thermal expansion problem described in section 6.4.1. The temperature coefficient of expansion, α, is given by equation 6.15. Replacing a in equation 6.15 by $\bar{y} - b\bar{x}$, gives

$$\alpha = \frac{b}{\bar{y} - b\bar{x}}. \tag{6.19}$$

The standard uncertainty, s_α, is now written

$$s_\alpha^2 = \left(\frac{\partial \alpha}{\partial \bar{y}}\right)^2 s_{\bar{y}}^2 + \left(\frac{\partial \alpha}{\partial b}\right)^2 s_b^2. \tag{6.20}$$

After determining the partial derivatives in equation 6.20 and substituting for the variances in $\bar{y}$ and b we obtain[17]

$$s_\alpha = \frac{s}{a^2}\left[\frac{b^2}{n} + \frac{n\bar{y}^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}\right]^{1/2}, \tag{6.21}$$

where n is the number of data points, s is given by equation 6.10, a is given by equation 6.6, and b is given by equation 6.7.

[15] An estimated slope, b, that is slightly larger than the true slope will consistently coincide with an estimated intercept, a, that is slightly smaller than the true intercept, and vice versa. See Weisberg (2005) for a discussion of correlation between a and b.

[16] See appendix 3.

[17] $s_{\bar{y}} = \frac{s}{\sqrt{n}}$ and s_b is given by equation 6.9.

Table 6.9. *Length–temperature data for an alumina rod.*

T (°C)	100	200	300	400	500	600	700	800	900	1000
l (m)	1.2019	1.2018	1.2042	1.2053	1.2061	1.2064	1.2080	1.2078	1.2102	1.2122

Exercise F

In an experiment to study thermal expansion, the length of an alumina rod is measured at various temperatures. The data are shown in table 6.9.

(i) Calculate the intercept and slope of the best straight line through the length–temperature data and the standard uncertainties in the intercept and slope.
(ii) Determine the temperature coefficient of expansion, α, for the alumina.
(iii) Calculate the standard uncertainty in α.

6.4.2 Estimating *y* for a given *x*

Once the intercept and slope of the best line through the points have been determined, we can predict the value of y, $\hat{y}_0$, at an arbitrary x value, x_0, using the relationship

$$\hat{y}_0 = a + bx_0. \tag{6.22}$$

In the absence of systematic errors, $\hat{y}_0$ is the best estimate of the true value of the y quantity at $x = x_0$.

The true value of y at $x = x_0$ may be written $\mu_{y|x_0}$. Just as the uncertainties in the measured y values contribute to the uncertainties in a and b, so the uncertainties in a and b contribute to the uncertainty in $\hat{y}_0$. As in section 6.4.1.1, we avoid the problem of correlation of errors in a and b by replacing a by $\bar{y} - b\bar{x}$, so that equation 6.22 becomes

$$\hat{y}_0 = \bar{y} + b(x_0 - \bar{x}). \tag{6.23}$$

As the errors in $\bar{y}$ and b are independent, the standard uncertainty in $\hat{y}_0$, written as $s_{\hat{y}_0}$, is given by

$$s_{\hat{y}_0} = s\left(\frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}\right)^{1/2}, \tag{6.24}$$

where s is given by equation 6.10.

We conclude that $s_{\hat{y}_0}$ and hence any coverage interval for $\mu_{y|x_0}$ depends upon the value of x at which the estimate of $\mu_{y|x_0}$ (i.e. $\hat{y}_0$) is calculated. The closer x_0 is to $\bar{x}$, the smaller is the second term inside the brackets of equation 6.24, and therefore the smaller is $s_{\hat{y}_0}$.

In general, the X% coverage interval for $\mu_{y|x_0}$ may be written as

$$\hat{y}_0 \pm t_{X\%,\nu} s_{\hat{y}_0}, \tag{6.25}$$

where $t_{X\%,\nu}$ is the critical t value corresponding to the X% level of confidence evaluated with ν degrees of freedom.[18]

Example 5

Consider the data in table 6.10.

Assuming a linear relationship between the x–y data in table 6.10, determine:

(i) the parameters a and b of the best line through the points;
(ii) $\hat{y}_0$ for $x_0 = 12$ and $x_0 = 22.5$;
(iii) the standard uncertainty in $\hat{y}_0$ when $x_0 = 12$ and $x_0 = 22.5$.
(iv) Plot the data in table 6.10 on an x–y graph showing the line of best fit and the limits of the 95% coverage interval for $\mu_{y|x_0}$ for values of x_0 between 0 and 45.

ANSWER

(i) a and b are found using equations 6.6 and 6.7. We find $a = 42.43$ and $b = -2.527$, so that the equation for $\hat{y}_0$ can be written as

$$\hat{y}_0 = 42.43 - 2.527x_0.$$

(ii) When $x_0 = 12$, $\hat{y}_0 = 12.11$. When $x_0 = 22.5$, $\hat{y}_0 = -14.43$.
(iii) Using the data in table 6.10: $\bar{x} = 22.5, \sum x_i^2 = 5100, \sum x_i = 180, s = 4.513$ (s is calculated using equation 6.10).

Substituting values into equation 6.24 gives for $x_0 = 12$, $s_{\hat{y}_0} = 2.2$. When $x_0 = 22.5$, $s_{\hat{y}_0} = 1.6$.

(iv) When the number of degrees of freedom = 6, the critical t value for the 95% coverage interval, given by table 2 in appendix 1 is $t_{95\%,6} = 2.447$. Equation 6.25 is used to find lines which represent the 95% coverage intervals for $\mu_{y|x_0}$ for values of x_0 between 0 and 45. These are indicated on the graph in figure 6.8.

Table 6.10. *The x–y data for example 5.*

x	5	10	15	20	25	30	35	40
y	28.1	5	−0.5	−7.7	−14.8	−27.7	−48.5	−62.9

[18] $t_{X\%,\nu}$ may be found using table 2 in appendix 1 or the T.INV.2T() function in Excel, as discussed in section 3.9.1.

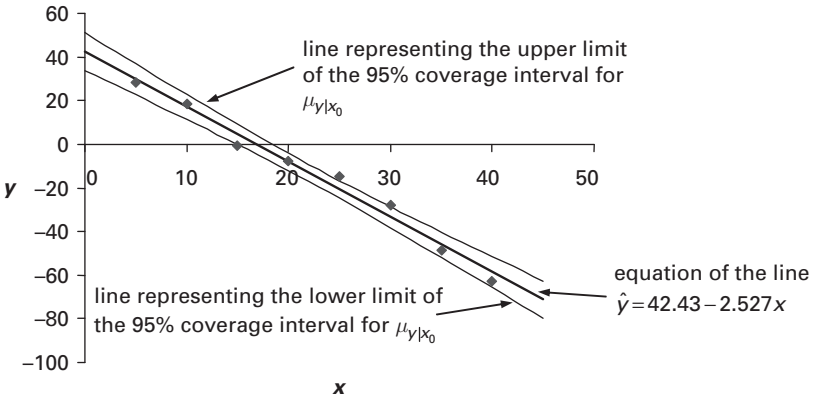

Figure 6.8. Line of best fit and upper and lower limits of the 95% coverage interval for data in table 6.10.

Exercise G

For the x–y data in example 5, calculate the 99% coverage interval for $\mu_{y|x_0}$ when $x_0 = 15$.

6.4.2.1 Uncertainty in prediction of y at a particular value of x

Let us assume that we have fitted the best straight line to a set of x–y data. If we make a measurement of y at $x = x_0$, between what limits would we expect the measured value of y to lie? This is different from considering coverage limits associated with the estimate of the population mean at $x = x_0$ because two factors must be taken into consideration:

(i) the uncertainty in the line of best fit to the data;
(ii) the uncertainty in the measurement of y made at $x = x_0$.

Using addition of uncertainties as discussed in section 5.8.2, we combine the uncertainty in the measurement of y with the uncertainty in the line of best to give[19]

$$s_{P\hat{y}_0} = s\left(1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x_i^2 - (\sum x_i)^2}\right)^{1/2}, \tag{6.26}$$

where $s_{P\hat{y}_0}$ represents that standard uncertainty in the predicted value of y at $x = x_0$.

The X% prediction interval for y at $x = x_0$ is written

$$\hat{y}_0 \pm t_{X\%,\nu} s_{P\hat{y}_0}, \tag{6.27}$$

where $\hat{y}_0$ is the best estimate of the predicted value. Note that $\hat{y}_0$ is the same as the best estimate of the population mean at $x = x_0$ and is given by equation 6.22;

[19] We assume that the error in the line of best fit and the error in the measurement of y made at $x = x_0$, are uncorrelated.

$t_{X\%,\nu}$ is the critical t value corresponding to the $X\%$ level of confidence evaluated with ν degrees of freedom.

Equation 6.26 is very similar to equation 6.24. However, the unity term within the brackets of equation 6.26 dominates over the other terms even when n is small and so the prediction interval for a y value at $x = x_0$ is larger than the coverage interval of the population mean $\mu_{y|x_0}$ at $x = x_0$.

Exercise H

Using information supplied in example 5, calculate the 95% prediction interval for y if a measurement of y is to be made at $x_0 = 12$.

6.4.3 Estimating x for a given y

Using the best straight line through points, we can estimate a value of x for any particular value of y. The equation of the best straight line through points is rearranged to give

$$\hat{x}_0 = \frac{\bar{y}_0 - a}{b}, \tag{6.28}$$

where $\hat{x}_0$ is the value of x when $y = \bar{y}_0$, and where $\bar{y}_0$ is the mean of repeated measurements of the dependent variable. The question arises, as there is uncertainty in a, b and $\bar{y}_0$, what will be the uncertainty in $\hat{x}_0$? As discussed in section 6.4.1.1, the errors in a and b are correlated. Replacing a in equation 6.28 by $\bar{y} - b\bar{x}$, we have

$$\hat{x}_0 = \bar{x} + \frac{\bar{y}_0 - \bar{y}}{b}. \tag{6.29}$$

Assuming the uncertainties in $\bar{y}_0$, $\bar{y}$ and b to be uncorrelated (and that there is no uncertainty in $\bar{x}$), we write

$$s_{\hat{x}_0}^2 = \left(\frac{\partial \hat{x}_0}{\partial \bar{y}_0}\right)^2 s_{\bar{y}_0}^2 + \left(\frac{\partial \hat{x}_0}{\partial \bar{y}}\right)^2 s_{\bar{y}}^2 + \left(\frac{\partial \hat{x}_0}{\partial b}\right)^2 s_b^2, \tag{6.30}$$

where $s_{\bar{y}_0} = \frac{s}{\sqrt{m}}$ (m is the number of repeat measurements made of y; s is given by equation 6.10); $s_{\bar{y}} = \frac{s}{\sqrt{n}}$ and s_b is given by equation 6.9.

Calculating the partial derivatives in equation 6.30 and substituting the expressions for the variances in $\bar{y}_0$, $\bar{y}$ and b, gives

$$s_{\hat{x}_0} = \frac{s}{b}\left[\frac{1}{m} + \frac{1}{n} + \frac{n(\bar{y}_0 - \bar{y})^2}{b^2\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)}\right]^{1/2}. \tag{6.31}$$

The $X\%$ coverage interval for $\hat{x}_0$ can be written,

$$\hat{x}_0 \pm t_{X\%,\nu} s_{\hat{x}_0}, \tag{6.32}$$

where ν is the number of degrees of freedom.

Example 6

A spectrophotometer is used to measure the concentration of arsenic in solution. Table 6.11 shows calibration data of the variation of absorbance[20] with arsenic concentration.

Assuming that the absorbance is linearly related to the arsenic concentration:

(i) plot a calibration graph of absorbance versus concentration;
(ii) find the intercept and the slope of the best straight line through the data.
Three values of absorbance are obtained from repeat measurements on a sample of unknown arsenic concentration. The mean of these values is found to be 0.3520.
(iii) Calculate the concentration of arsenic corresponding to this absorbance and the standard uncertainty in the concentration.

ANSWER

(i) Figure 6.9 shows a plot of the variation of absorbance versus concentration data contained in table 6.11.
(ii) a and b are determined using equations 6.6 and 6.7:

$$a = 0.01258, b = 0.0220\,\text{ppm}^{-1}.$$

(iii) Using the relationship, $\hat{x}_0 = \frac{\bar{y}_0 - a}{b}$,

$$\hat{x}_0 = \left(\frac{0.3520 - 0.01258}{0.02220}\right)\text{ppm} = 15.29\,\text{ppm}.$$

We use equation 6.31 to obtain the standard uncertainty in $\hat{x}_0$. Using the information in the question and the data in table 6.11 we find $n = 5, m = 3, \bar{y}_0 = 0.3520$, $\bar{y} = 0.37842, \sum x_i^2 = 1852.363, \sum x_i = 82.483$, $s = 0.01153$ (s is calculated by using equation 6.10).

Substituting these values into equation 6.31 gives

$$s_{\hat{x}_0} = 0.38\ \text{ppm}.$$

It is worth remarking that the third term in the brackets of equation 6.31 becomes large for y values far from the mean of the y values obtained during the calibration procedure. The third term is zero (and hence the standard uncertainty in $\hat{x}_0$ is smallest) when the y value of the sample under test is equal to the mean y values obtained during calibration.

[20] Absorbance is proportional to the amount of light absorbed by the solution as the light passes from source to detector within the spectrophotometer.

Table 6.11. *Absorbance as a function of arsenic concentration.*

Concentration (ppm)	Absorbance
2.151	0.0660
9.561	0.2108
16.878	0.3917
23.476	0.5441
30.337	0.6795

Table 6.12. *Variation of peak area with nitrite concentration.*

Concentration (ppm)	2.046	3.068	5.114	7.160	10.23
Peak area (arbitrary units)	37 752	47 658	75 847	105 499	154 750

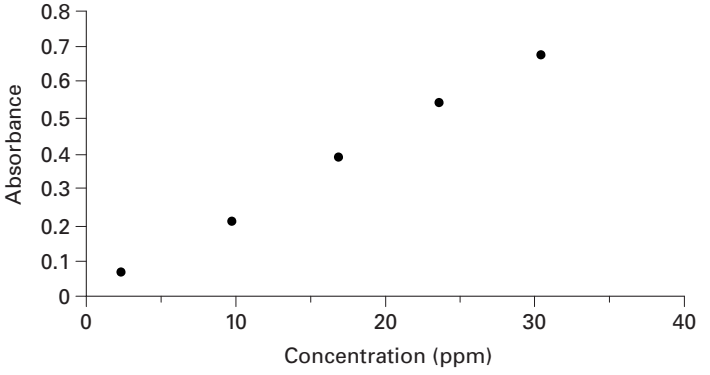

Figure 6.9. Absorbance versus concentration for a solution containing arsenic.

Exercise I

The data in table 6.12 were obtained in an experiment to determine the amount of nitrite in solution using high performance liquid chromatography (HPLC).

(i) Regarding the peak area as the dependent variable, determine the equation of the best straight line through the data.

(ii) Four repeat measurements are made on a solution with unknown nitrite concentration. The mean peak area is found to be 57 156. Use the equation of the best line to find the concentration corresponding to this mean peak area and the standard uncertainty in the concentration.

6.5 Fitting a straight line to data when random errors are confined to the *x* quantity

One of the assumptions we rely upon consistently in least squares analysis is that errors in *x*–*y* data are confined to the *y* quantity. The validity of this assumption must be considered on a case by case basis, but it is reasonable to argue that all measurements have some error so that there will be some error in the *x* values. It is possible to derive equations for slope, intercept and the uncertainties in these quantities in situations in which there are uncertainties in both the *x* and the *y* values.[21] If the errors in the *x* values are constant and errors in the *y* values are negligible, we can use the results already derived in this chapter to find the best line through the data. We write the equation of the line through the data as

$$x = a^* + b^* y, \tag{6.33}$$

where x is regarded as the dependent variable and y as the independent variable, a^* is the intercept (i.e. the value of x when $y = 0$) and b^* is the slope of the line. To find a^* and b^*, we must minimise the sum of squares of the residuals of the observed values of x from the predicted values based on a line drawn through the points. In essence we recreate the argument begun in section 6.2.2, but with y replacing x and x replacing y. The equation for the best line through the x–y data in this case is given when the intercept, a^*, is

$$a^* = \frac{\sum y_i^2 \sum x_i - \sum y_i \sum x_i y_i}{n \sum y_i^2 - \left(\sum y_i\right)^2}, \tag{6.34}$$

and the slope, b^*, is,

$$b^* = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum y_i^2 - \left(\sum y_i\right)^2}. \tag{6.35}$$

Compare these equations with 6.6 and 6.7 for a and b when the sum of squares of residuals in the y values is minimised.

Equation 6.33 can be rewritten as

$$y = \frac{-a^*}{b^*} + \frac{x}{b^*}. \tag{6.36}$$

It is tempting to compare equation 6.36 with $y = a + bx$ and reach the conclusion that

$$a = \frac{-a^*}{b^*} \tag{6.37}$$

[21] This is beyond the scope of this text. For a review of least squares methods when both x and y variables are affected by error, see Cantrell (2008).

Table 6.13. *The x–y data.*

x	y
2.52	2
3.45	4
3.46	6
4.25	8
4.71	10
5.47	12
6.61	14

and

$$b = \frac{1}{b^*}. \tag{6.38}$$

However, a and b given by equations 6.37 and 6.38 are equal to a and b given by equations 6.6 and 6.7 only if both least squares fits (i.e. that which minimises the sum of the squares of the x residuals and that which minimises the sum of the squares of the y residuals) produce the same straight line through the points. The only situation in which this happens is when there are *no* errors in either the x or the y values, i.e. all the data lie exactly on a straight line!

As an example, consider the x–y data in table 6.13.

We can perform a least squares analysis assuming the following.

(i) The errors are in the x values only. Using equations 6.34 and 6.35 we find, $a^* = 1.844$ and $b^* = 0.3136$. Using equations 6.37 and 6.38 we obtain $a = -5.882$ and $b = 3.189$.

(ii) The errors are in the y values only. Using equations 6.6 and 6.7 we find, $a = -5.348$ and $b = 3.067$.

As anticipated, the parameter estimates, a and b, depend on whether minimisation occurs in the sum of the squares of the x residuals or the sum of the squares of the y residuals.

6.6 Linear correlation coefficient, r

When the influence of random errors on values is slight, it is usually easy to establish 'by eye' whether x and y quantities are related. However, when data are scattered it is often quite difficult to be sure of the extent to which y is dependent on x. It is useful to define a parameter directly related to the extent of

[22] During the experiment the current through the diode is held constant.

Table 6.14. *Voltage across a diode, V, as a function of temperature, θ.*

θ (°C)	V (V)
2.0	0.6859
10.0	0.6619
19.0	0.6379
26.4	0.6139
40.9	0.5899
48.8	0.5659
59.7	0.5419
65.0	0.5179
80.0	0.4939
91.0	0.4699
101.3	0.4459

Exercise J

The voltage, V, across a silicon diode is measured as a function of temperature, θ, of the diode.[122] Temperature–voltage data are shown in table 6.14.

Theory suggests that the relationship between V and θ is,

$$V = k_0 + k_1\theta. \tag{6.39}$$

As V is the dependent variable and θ the independent variable, it would be usual to plot V on the y axis and θ on the x axis. However, the experimenter has evidence that the measured values of diode voltage are more precise than those of temperature.

(i) Use least squares to obtain best estimates for k_0 and k_1, where values of V are assumed to have most error.

(ii) Use least squares to obtain best estimates for k_0 and k_1, where values of θ are assumed to have most error.

the correlation between x and y. That parameter is called the linear correlation coefficient, ρ. The estimate of this parameter, obtained from data, is represented by the symbol, r.

When all points lie on the same straight line, i.e. there is perfect correlation between x and y, then equations 6.37 and 6.38 correctly relate a and b (which are determined assuming errors are confined to the y quantity) with a^* and b^* (which are determined assuming errors are confined to the x quantity). Focussing on equation 6.38 we can say that, for perfect correlation,

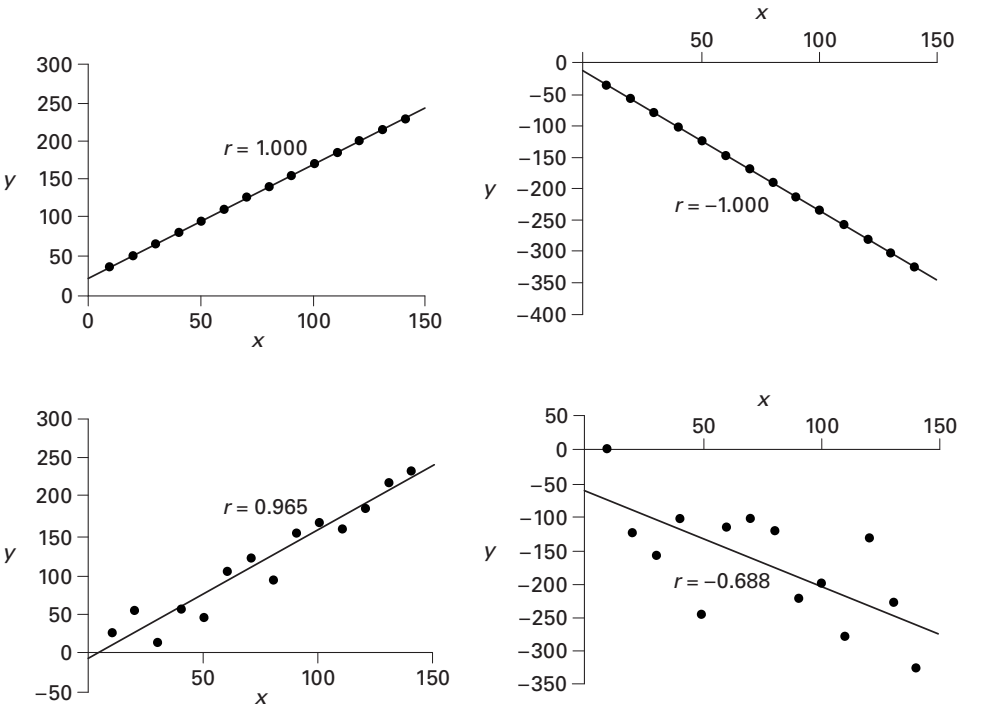

Figure 6.10. Correlation coefficients for x–y data exhibiting various amounts of scatter.

$$bb^* = 1.$$

When there is no correlation between x and y, there is no tendency for y to increase with increasing x nor x to increase with increasing y. Put another way, with no correlation we would expect both b and b^* to be close to zero, so that $bb^* \approx 0$. This suggests that it may be useful to use the product bb^* as a measure of the correlation between x and y. We define the linear correlation coefficient, r, as

$$r = \sqrt{bb^*}. \tag{6.40}$$

Substituting for b and b^* using equations 6.7 and 6.35 respectively, we get

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]^{1/2}\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]^{1/2}}. \tag{6.41}$$

For perfect correlation, r is either +1 or –1. Note that r has the same sign as that of the slope (b or b^*). Figure 6.10 shows graphs of x–y data along with the value of r for each.

As $|r|$ decreases from 1 to 0, the correlation between x and y becomes less and less convincing. Notions of 'goodness' relating to values of r can be misleading. A value of $|r|$ close to unity does indicate good correlation, however it is possible that x and y are not *linearly* related but still result in a value for $|r|$ in excess of 0.99. This is illustrated by the next example.

Example 7

Thermoelectric coolers (TECs) are devices widely used to cool electronic components, such as laser diodes and are also found in portable refrigerators. A TEC consists of a hot and cold surface with the temperature difference between the surfaces maintained by an electric current flowing through the device.

In an experiment, the temperature difference, ΔT, between the hot and the cold surface is measured as a function of the electrical current, I, passing through the TEC. Data gathered are shown in table 6.15.

(i) Calculate the value of the correlation coefficient, r.
(ii) Plot a graph of ΔT versus I and show the line of best fit through the points.

ANSWER

(i) We begin by drawing up a table containing all the values needed to calculate r using equation 6.41.

Summing the values in each column gives $\sum x_i = 4.2$, $\sum y_i = 113.2$, $\sum x_i y_i = 93.02$, $\sum x_i^2 = 3.64$, $\sum y_i^2 = 2402.22$.

Substituting the summations into equation 6.41 gives $r = \dfrac{175.7}{2.8 \times 63.2558} = 0.992$.

A fit of the equation $y = a + bx$ to the x–y data given in table 6.16 gives intercept, $a = 2.725\ ^\circ$C, and slope, $b = 22.41^\circ$C/A.

(ii) Figure 6.11 shows the data points and the line of best fit.

A value of $r = 0.992$ indicates a high degree of correlation between x and y. That x and y are correlated is revealed by the graph, but we might be overlooking something more important. Is the relationship between temperature difference and current *really* linear? We assumed that to be the case when calculating the line of best fit, though a close inspection of the graph above seems to indicate that a curve through the points is more appropriate than a straight line. This is a point worth emphasising: a correlation coefficient with magnitude close to unity is no guarantee that the x–y data are linearly related. We need to inspect an x–y graph of the raw data, the line of best fit and preferably a plot of the residuals (dealt with in section 6.7) in order to be satisfied that the assumption of linearity between x and y is reasonable.

Table 6.15. *Temperature difference, ΔT, of a TEC as a function of current, I.*

I (A)	ΔT (°C)
0.0	0.8
0.2	7.9
0.4	12.5
0.6	17.1
0.8	21.7
1.0	25.1
1.2	28.1

Table 6.16. *Values needed to calculate r using equation 6.41.*

x_i (=I)	y_i (=ΔT)	$x_i y_i$	x_i^2	y_i^2
0.0	0.8	0.0	0.0	0.64
0.2	7.9	1.58	0.04	62.41
0.4	12.5	5.0	0.16	156.25
0.6	17.1	10.26	0.36	292.41
0.8	21.7	17.36	0.64	470.89
1.0	25.1	25.1	1.0	630.01
1.2	28.1	33.72	1.44	789.61

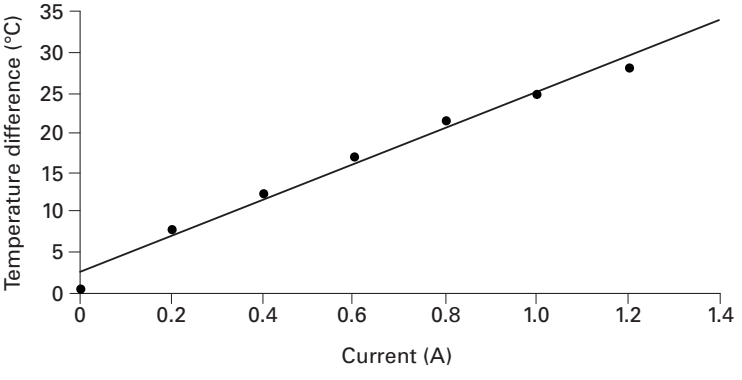

Figure 6.11. Temperature difference for a TEC versus current.

Table 6.17. *Variation of temperature of the cold surface of TEC with the volume of heat sink.*

Volume, V (cm^3)	Temperature, T (°C)
5	37.0
10	25.5
15	17.1
25	11.5
50	6.4

Exercise K

The temperature reached by the cold surface of a thermoelectric cooler depends on the size of the heat sink to which the hot surface of the cooler is attached. Table 6.17 shows the temperature, T, of the cold surface of the TEC for aluminium heat sinks of various volumes, V.

(i) Calculate the linear correlation coefficient for these data.
(ii) Find the equation of the line of best fit through the data assuming that a linear relationship is appropriate (plot T on the y axis and V on the x axis).
(iii) Plot a graph of temperature versus volume and indicate the line of best fit.
(iv) Do you think that the assumption of linearity between T and V is valid?

6.6.1 Calculating *r* using Excel

The CORREL() function in Excel calculates the correlation coefficient, r, and returns the value into a cell. The syntax of the function is as follows.

CORREL(y values, x values)

Consider the calculation of r for the data shown in sheet 6.2.

Sheet 6.2. *The x–y data.*

	A	B
1	x	y
2	42	458
3	56	420
4	56	390
5	78	380
6	69	379
7	92	360
8	102	351
9	120	300
10		

Table 6.18. *The x–y data for exercise L.*

x	2	4	6	8	10	12	14	16	18	20	22	24
y	16.5	49.1	65.2	71.6	101.5	90.1	101.4	113.7	127.7	156.5	203.6	188.4

Table 6.19. *Correlation coefficient for ten sets of randomly generated y values correlated with the x column values.*

x	*y1*	*y2*	*y3*	*y4*	*y5*	*y6*	*y7*	*y8*	*y9*	*y10*
0.2	0.020	0.953	0.508	0.324	0.233	0.872	0.446	0.673	0.912	0.602
0.4	0.965	0.995	0.231	0.501	0.265	0.186	0.790	0.911	0.491	0.186
0.6	0.294	0.159	0.636	0.186	0.227	0.944	0.291	0.153	0.780	0.832
0.8	0.561	0.096	0.905	0.548	0.187	0.002	0.331	0.051	0.862	0.255
1.0	0.680	0.936	0.783	0.034	0.860	0.363	0.745	0.083	0.239	0.363
r	**0.400**	**−0.323**	**0.743**	**−0.392**	**0.655**	**−0.455**	**0.094**	**−0.822**	**−0.541**	**−0.242**

To calculate *r*:

(i) enter the data shown in sheet 6.2 into an Excel spreadsheet;
(ii) type **=CORREL(B2:B9,A2:A9)** into cell B10;
(iii) press the Enter key;
(iv) the value –0.9503 is returned in cell B10.

Exercise L

Use the CORREL() function to determine the correlation coefficient of the *x*–*y* data shown in table 6.18.

6.6.2 Is the value of *r* significant?

We have seen that we need to be cautious when using *r* to infer the extent to which there is a linear relationship between *x*–*y* data, as values of $|r| > 0.8$ may be obtained when the underlying relationship between *x* and *y* is not linear. There is another issue: values of $|r| > 0.8$ may be obtained when *x* and *y* are totally uncorrelated, especially when the number of *x*–*y* values is small. To illustrate this, consider the values in table 6.19.

The first column of table 6.19 contains five values of *x* from 0.2 to 1. The remainder of the columns contain numbers between 0 and 1 that have been randomly generated so that there is *no underlying correlation between x and y*. The bottom row shows the correlation coefficients calculated when each column of *y* is correlated in turn with the column containing *x*.

The column of values headed *y8*, when correlated with the column headed *x*, gives a value of $|r| > 0.8$, which might be considered in some circumstances as 'good' but in this case has just happened 'by chance'. The basic message is that, if there are only a few points (say ≤ 5), it is possible that consecutive numbers will be in ascending or descending order. If this is the case then the magnitude of the correlation coefficient can easily exceed 0.8. How, then, do you know if the correlation coefficient is significant? The question can be put as follows.

On the assumption that a set of *x*–*y* data are uncorrelated, what is the probability of obtaining the value of *r* at least as large as that calculated from the available data?

If that probability is small (say < 0.05) then it is highly unlikely that the *x*–*y* data are uncorrelated, i.e. the data *are* likely to be correlated. Table 6.20 shows the probability of obtaining a particular value for $|r|$ when the *x*–*y* data are uncorrelated. The probability of obtaining a particular value for $|r|$ decreases, when the *x*–*y* values are uncorrelated, as the number of values, *n*, increases.[23]

A correlation coefficient is considered significant if the probability of it occurring, when the data are uncorrelated,[24] is less than 0.05. Referring to table 6.20 we see that, for example, with only four data points ($n = 4$), a value of *r* of 0.9 is not significant as the probability that this could happen with uncorrelated data is 0.10. When the number of data values is 10 or more, then values of *r* of greater than about 0.6 become significant.

Table 6.20. *Probabilities of obtaining calculated r values when x–y data are uncorrelated.*

	$\|r\|$ (calculated from *x*–*y* data)					
n	0.5	0.6	0.7	0.8	0.9	1.0
3	0.67	0.59	0.51	0.41	0.29	0
4	0.50	0.40	0.30	0.20	0.10	0
5	0.39	0.28	0.19	0.10	0.037	0
6	0.31	0.21	0.12	0.056	0.014	0
7	0.25	0.15	0.080	0.031	0.006	0
8	0.21	0.12	0.053	0.017	0.002	0
9	0.17	0.088	0.036	0.010	0.001	0
10	0.14	0.067	0.024	0.005	<0.001	0

[23] See Bevington and Robinson (2002) for a discussion of the calculation of the probabilities in table 6.20.

[24] This value of probability is often used in tests to establish statistical significance, as discussed in chapter 9.

Table 6.21. *The x–y data for exercise M.*

x	y
2.67	1.54
1.56	1.60
0.89	1.56
0.55	1.34
–0.25	1.33

Exercise M

Consider the x–y data in table 6.21.

(i) Calculate the value of the correlation coefficient, r.
(ii) Is the value of r significant?

6.7 Residuals

Another indicator of 'goodness of fit', complementary to the correlation coefficient and often more revealing, is the distribution of residuals, Δy. The residual, $\Delta y = y - \hat{y}$, is plotted on the vertical axis against x. If we assume that a particular equation is 'correct' in that it accurately describes the data that we have collected, and only random errors exist to obscure the true value of y at any value of x, then the residuals would be expected to scattered randomly about the $\Delta y = 0$ axis. Additionally, if the residuals are plotted in the form of a histogram, they would appear to be normally distributed if the errors are normally distributed.

If we can discern a regular pattern in the residuals then one or more factors must be at play to cause the pattern to occur. It could be:

(i) we have chosen an equation of dubious validity to fit to the data;
(ii) there is an outlier in the data which has a large influence on the line of best fit;
(iii) we should be using a weighted fit, that is, the uncertainties in the measured y values are not constant and we need to take this into account when applying least squares to obtain parameter estimates (section 6.10 considers weighted fitting of a straight line to data).

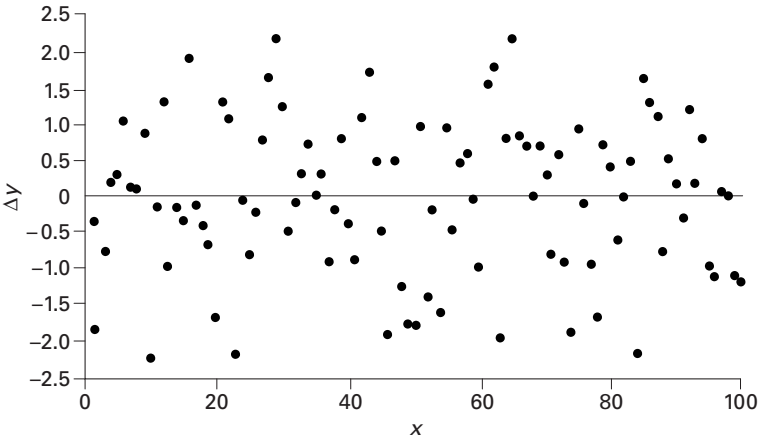

Figure 6.12. Ideal distribution of residuals – no regular pattern discernible.

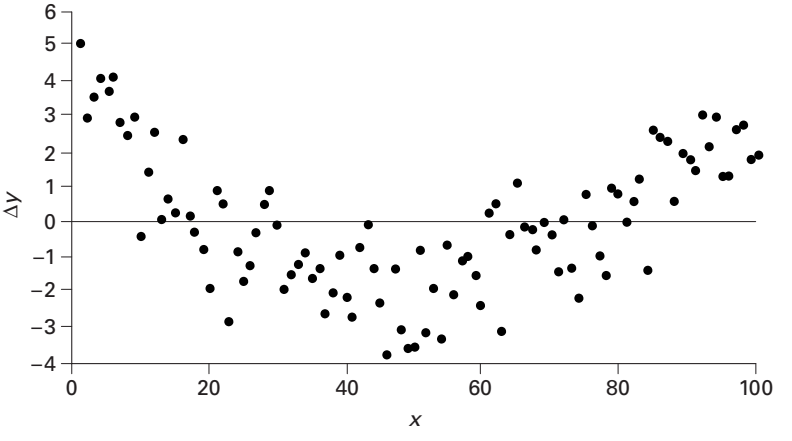

Figure 6.13. Residuals revealing that an inappropriate equation may have been fitted to data.

Typical plots of residuals[25] versus x, which illustrate patterns that can emerge, are shown in figures 6.12 to 6.15.

No discernible regular pattern is observable in the residuals in figure 6.12. This offers strong evidence to support the notion that the fit of the equation to the data is good. If the standard deviation of the residuals is constant as x

[25] Another way to display residuals is to plot Δy_i versus $\hat{y}_i$.

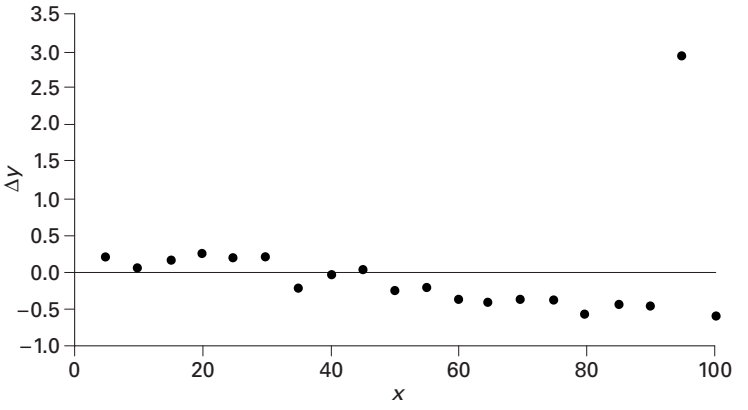

Figure 6.14. Effect of outlier on residuals.

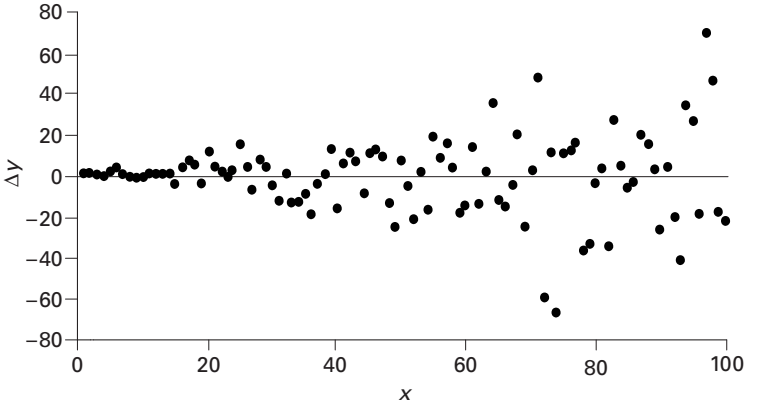

Figure 6.15. Pattern in residuals suggesting weighted fitting should be used.

increases (as in figure 6.12), those residuals are sometimes referred to as being *homoscedastic*.

In figure 6.13 the residuals reveal a systematic variation as x increases. This should lead us to question the appropriateness of the equation we have fitted to the data.

Figure 6.14 shows an extreme example of an outlier existing in x–y data. If there are many data, the intercept and slope, a and b, are little affected by such an outlier. However, the situation is quite different if there are far fewer data points (say 10 or so). In such a situation an outlier can have a dramatic effect on a and b (and the uncertainties in a and b).

In figure 6.15 the residuals increase steadily with increasing in x. This occurs when the scatter in the data (i.e. due to random errors in the y quantity) increases with increasing x. In this situation an unweighted fit of the equation to the data is not appropriate and a *weighted* fit using weighted least squares should be used instead. Weighted least squares is considered in section 6.10. If the standard deviation of the residuals is not constant as x increases (as in figure 6.15), those residuals are sometimes referred to a being *heteroscedastic*.

Though residuals can be very revealing with regard to identifying outliers, inappropriate equations and incorrect weighting, they can be misleading if there are only a few data (say, <10) as a pattern can emerge within the residuals 'by chance'.

Exercise N

The period, T, of oscillation of a body on the end of a spring is measured as the mass, m, of the body increases. Data obtained are shown in table 6.22.

(i) Plot a graph of period versus mass of body.
(ii) Assuming that the period is linearly related to the mass, use least squares to find the equation of the best line through the points.
(iii) Calculate the residuals and plot a graph of residuals versus mass.
(iv) Is there any 'pattern' discernible in the residuals? If so, suggest a possible cause of the pattern.

Table 6.22. *Variation of period, T, of a body of mass, m, on a spring.*

m (kg)	T (s)
0.2	0.39
0.4	0.62
0.6	0.72
0.8	0.87
1.0	0.92
1.2	1.07
1.4	1.13
1.6	1.16
1.8	1.23
2.0	1.32

6.7.1 Standardised residuals

If each residual, Δy_i , is divided by the standard deviation in each y value, s_i, we refer to the quantity

$$\Delta y_{is} = \frac{\Delta y_i}{s_i} \tag{6.42}$$

as the *standardised* residual.[26] In situations in which the standard deviation in each y value is the same, we replace s_i by s, so that

$$\Delta y_{is} = \frac{\Delta y_i}{s}, \tag{6.43}$$

where s is given by equation 6.10.

Scaling residuals in this manner is very useful as the standardised residuals should be normally distributed with a standard deviation of 1 if the errors causing scatter in the data are normally distributed. If we were to plot Δy_s against x, we should find that, not only should the standardised residuals be scattered randomly about the $\Delta y_s = 0$ axis, but that (based on properties of the normal distribution[27]) about 70% of the standardised residuals should lie between $\Delta y_s = \pm 1$ and about 95% should lie between $\Delta y_s = \pm 2$. If this is not the case, it suggests that the scatter of y values does not follow a normal distribution.

Example 8

Consider the x–y data in table 6.23.

(i) Assuming a linear relationship between x and y, determine the equation of the best line through the data in table 6.23.
(ii) Calculate the standard deviation, s, of the data about the fitted line.
(iii) Determine the standardised residuals and plot a graph of standardised residuals versus x.

ANSWER

(i) Applying equations 6.6 and 6.7 to the data in table 6.23, we find $a = 7.849$ and $b = 2.830$, so that the equation of the best line can be written as

$$\hat{y} = 7.849 + 2.830x. \tag{6.44}$$

(ii) Applying equation 6.10 gives $s = 5.376$.
(iii) Table 6.24 includes the predicted y values found using equation 6.44, the residuals and the standardised residuals.

Figure 6.16 shows a plot of the standardised residuals, Δy_s, versus x.
As anticipated, the standardised residuals lie between $\Delta y_s = \pm 2$.

[26] See Devore (2007) for a fuller discussion of standardised residuals.

[27] See section 3.5.

Table 6.23. *The x–y data for example 8.*

x	2	4	6	8	10	12	14	16	18	20	22
y	8.2	23.0	26.3	31.4	39.5	36.7	56.7	46.8	53.4	63.2	74.7

Table 6.24. *Predicted values, residuals and standardised residuals.*

x	2	4	6	8	10	12	14	16	18	20	22
y	8.2	23.0	26.3	31.4	39.5	36.7	56.7	46.8	53.4	63.2	74.7
$\hat{y}$	13.509	19.169	24.829	30.489	36.149	41.809	47.469	53.129	58.789	64.449	70.109
Δy	−5.309	3.831	1.471	0.911	3.351	−5.109	9.231	−6.329	−5.389	−1.249	4.591
Δy_s	−0.988	0.713	0.274	0.169	0.623	−0.950	1.717	−1.177	−1.002	−0.232	0.854

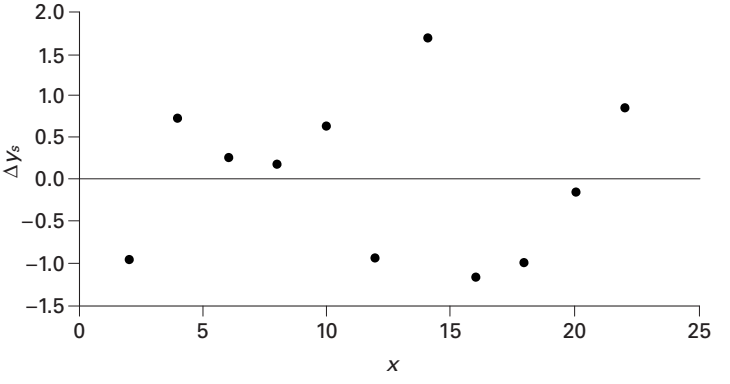

Figure 6.16. Plot of standardised residuals.

6.8 Data rejection

A topic of some importance and controversy is that of data rejection. If we gather x–y data and most of the points lie along a straight line, but one point lies well away from the line, should we reject that point? This is a difficult question to which to give a simple yes or no answer. It may be that a 'slip-up' occurred during the recording of the data and therefore we are justified in ignoring the point. However, it could be that there is a true deviation from straight line

behaviour and that the apparent outlier is not an outlier at all but is revealing something really important – perhaps we have observed a new effect! To reject that point may be to eliminate the most important value you have gathered. If possible, repeating the experiment is always preferable to rejecting data with little justification other than it 'doesn't fit'.

The decision to reject data largely depends on what relationship we believe underlies the data that has been collected. For example, figure 6.17 shows data obtained in an experiment to study the specific heat of a metal at low temperature. Though the underlying trend between x and y may or may not be linear, it *does* appear that there is a spurious value for the specific heat at about 3 K. Performing a statistically based test, such as that described next, would indicate that the point should be removed. However, the 'jump' in specific heat at about 3 K is a real effect and to eliminate the point would mean that a vital piece of evidence giving insight into the behaviour of the material would have been discarded.

There may be situations in which large deviations can be attributed to spurious effects. In these cases it is possible to use the trend exhibited by the majority of the data as a basis for rejecting data that appear to be outliers.

In section 5.11.2 we introduced Chauvenet's criterion as a way of deciding whether a value within a sample is so far from the mean of the sample that it should be considered for rejection. The basic assumption when applying the criterion is that data are normally distributed. If the assumption is valid, the probability can be determined that a value will be obtained that is at least as far from the mean as a 'suspect' value. Multiplying that probability by the number

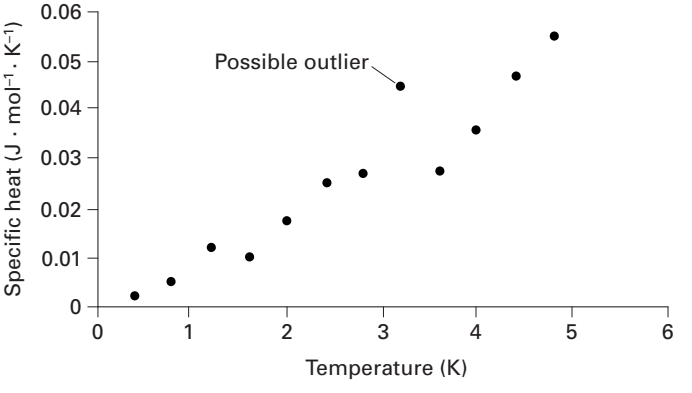

Figure 6.17. Variation of specific heat of tin with temperature. A possible outlier is indicated.

of data gives the number of values *expected* to lie at least as far from the mean as a 'suspect' value. If the expected number is less than or equal to 0.5, the suspect point should be rejected and the sample mean and standard deviation recalculated.

We can apply Chauvenet's criterion to x–y data by considering how far an observed y value is from the line of best fit. The difference between the observed value of y and the predicted value of y based on the line of best fit is expressed in terms of the number of standard deviations between observed and predicted value. Specifically, we write

$$z_{OUT} = \frac{y_{OUT} - \hat{y}}{s}, \tag{6.45}$$

where y_{OUT} is the outlier (i.e. the point furthest from the line of best fit), $\hat{y}$ is the predicted value, found using $\hat{y} = a + bx$. We calculate s using equation 6.10.

Once z_{OUT} has been calculated, we determine the expected number of values, N, at least as far away from $\hat{y}$ as y_{OUT}. To do this:

(i) determine the probability, P, of obtaining a value of $|z| > z_{OUT}$;
(ii) calculate the expected number of values, N, greater than or equal to $|z_{OUT}|$. $N = nP$, where n is the number of data.

If N is less than 0.5, consider rejecting the point. If a point *is* rejected, then a and b should be recalculated (as well as other related quantities, such as s_a and s_b).

Example 9

Consider x–y values in table 6.25.

(i) Plot an x–y graph and use unweighted least squares to fit a line of the form $y = a + bx$ to the data.
(ii) Identify any suspect point(s).
(iii) Calculate the standard deviation of the y values.
(iv) Apply Chauvenet's criterion to the suspect point – should it be rejected?

ANSWER

(i) A plot of data is shown in figure 6.18 with line of best fit attached ($a = -1.610$ and $b = 2.145$).
(ii) A suspect point would appear to be $x = 6$, $y = 8.5$ as this point is furthest from the line of best fit.
(iii) Using the data in table 6.25 and equation 6.10, $s = 1.895$.
(iv) Using equation 6.45, we have

Table 6.25. *The x–y data with a 'suspect' value.*

x	y
2.0	3.5
4.0	7.2
6.0	8.5
8.0	17.1
10.0	20.0

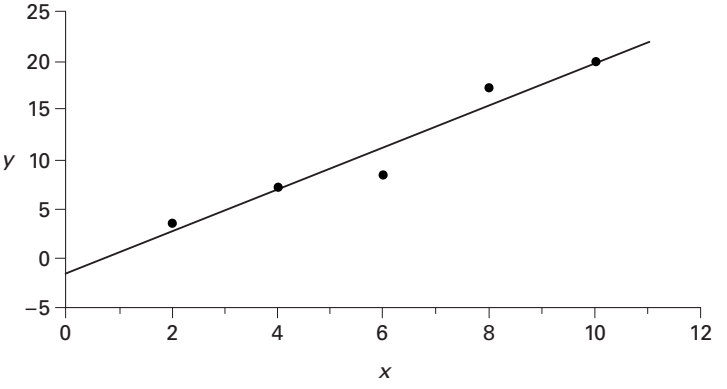

Figure 6.18. The x–y graph for data in table 6.25.

$$z_{OUT} = \frac{8.5 - (-1.610 + 2.145 \times 6)}{1.895} = -1.46.$$

Using table 1 in appendix 1, the probability of an absolute value of $z_{OUT} > 1.46 = 2 \times 0.0722 = 0.144$. The expected number of points at least this far from the line is expected to be $5 \times 0.144 = 0.72$. As this value is greater than 0.5, do not reject the point.

Exercise O

Consider the x–y data in table 6.26.

When a straight line is fitted to the data in table 6.26, it is found that $a = 4.460$ and $b = 0.6111$.

Assuming that the data point $x = 10$, $y = 13$ is 'suspect', apply Chauvenet's criterion to decide whether this point should be rejected.

Table 6.26. *The x–y data for exercise O.*

x	y
5	7
6	8
8	9
10	13
12	11
14	12
15	14

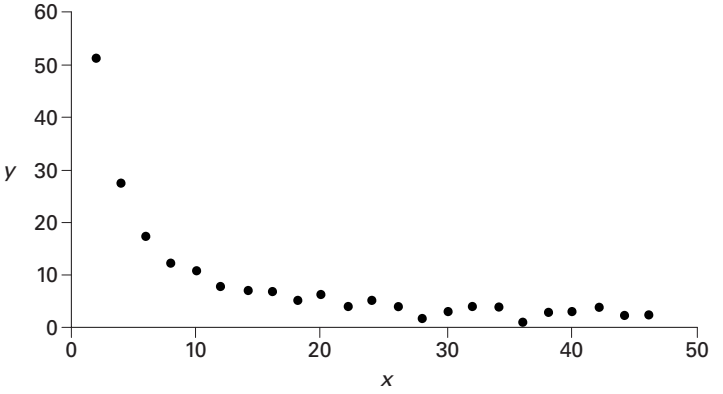

Figure 6.19. Data with non-linear relationship between x and y.

6.9 Transforming data for least squares analysis

Using the technique of least squares we can establish best estimates for the intercept, a, and slope, b, of a line through x–y data through equations 6.6 and 6.7. These equations should only be applied when we are confident that there is a linear relationship between the quantities plotted on the x and y axes. What do we do if x–y data are clearly non-linearly related, such as those in figure 6.19?

It may be possible to apply a mathematical operation to the x or y data (or to both x and y) so that the *transformed* data appear linearly related. The next stage is to fit the equation $y = a + bx$ to the transformed data. How do we choose which mathematical operation to apply? If we have little or no idea what the relationship between the dependent and independent variable is likely to be, we may be

forced into a 'trial and error' approach to transforming data. As an example, the rapid decrease in y with x indicated for $x < 10$ in figure 6.19 suggests that there might be an exponential relationship between x and y, such as

$$y = Ae^{Bx}, \tag{6.46}$$

where A and B are constants.

Assuming this to be the case, taking natural logs of each side of the equation gives

$$\ln y = \ln A + Bx. \tag{6.47}$$

Comparing equation 6.47 with the equation of a straight line predicts that plotting ln y versus x will produce a straight line with intercept, ln A, and slope, B. Figure 6.20 shows the effect of applying the transformation suggested by equation 6.47 to the data in figure 6.19.

The transformation has not been successful in producing a linear x–y graph, indicating that equation 6.46 is not appropriate to these data and that other transformation options should be considered (for example ln y versus ln x). Happily, in many situations in the physical sciences, the work of others (either experimental or theoretical) provides clues as to how the data should be treated in order to produce a linear relationship between data plotted as an x–y graph. Without such clues we must use 'intelligent guesswork'.

As an example, consider a ball falling freely under gravity. The distance, s, through which a ball falls in a time, t, is measured and the values obtained are shown in table 6.27. The data are shown in graphical form in figure 6.21.

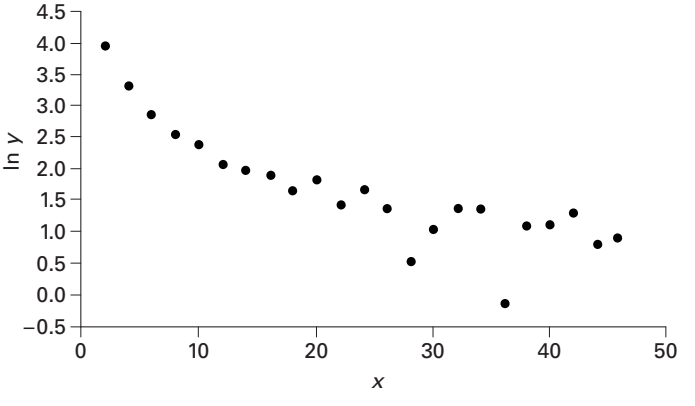

Figure 6.20. Effect of transforming y values by taking natural logarithms.

Table 6.27. *Distance–time data for a freely falling ball.*

t (s)	s (m)
1	7
2	22
3	52
4	84
5	128
6	200

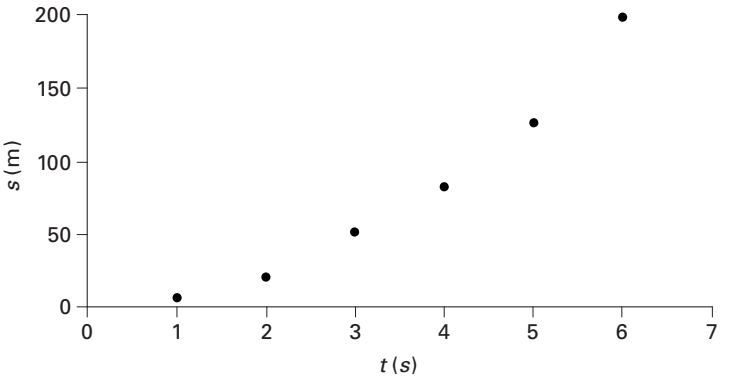

Figure 6.21. Distance–time data for a falling ball.

The relationship between s and t is not linear. Can the data be transformed so that a linear graph is produced? The starting point is to look for an equation which might describe the motion of the ball. When the acceleration, g, is constant, we can write the relationship between s and t, as[28]

$$s = ut + \tfrac{1}{2}gt^2, \tag{6.48}$$

where u is the initial velocity of the body.

To linearise equation 6.48, divide throughout by t to give,

$$\frac{s}{t} = u + \tfrac{1}{2}gt. \tag{6.49}$$

Compare the quantities in equation 6.49 with $y = a + bx$,

$$\underbrace{\frac{s}{t}}_{y} = \underbrace{u}_{a} + \underbrace{\tfrac{1}{2}g}_{b}\ \underbrace{t}_{x}.$$

[28] See Young, Freedman and Ford (2007).

Table 6.28. *Transformation of data given in table 6.27.*

t (s)	s/t (m/s)
1	7
2	11
3	17.3
4	21
5	25.6
6	33.3

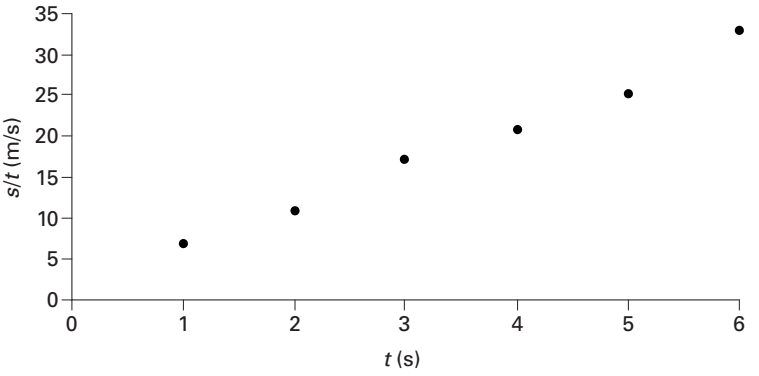

Figure 6.22. Graph of s/t versus t.

If the relationship between the displacement and time can be represented by equation 6.48, then plotting $\frac{s}{t}$ versus t should transform the experimental data so that they lie on (or close to) a straight line which has intercept, u, and slope, ½g.

Table 6.28 contains the transformed data and figure 6.22 the corresponding graph.

It certainly does appear that transforming the data has produced a graph in which the quantities plotted on the x and y axes are linearly related. Fitting a line to the transformed data using least squares gives

$$a = 1.298 \text{ m/s and } b = 5.118 \text{ m/s}^2.$$

When transforming data, the dependent variable should remain on the left hand side of the equation (so that the assumption of errors being restricted to the quantity plotted on the y axis is valid). Usually the independent variable only appears on the right hand side of the equation. However, there are situations, such as in the linearisation of equation 6.48, where this condition must be relaxed.

Exercise P

(1) Transform the equations shown in table 6.29 into the form $y = a + bx$, and indicate how the constants in each equation are related to a and b.

(2) The capacitance of a semiconductor diode decreases as the reverse bias voltage applied to the junction increases. An important diode parameter, namely the contact potential, can be found if the capacitance of the junction is measured as a function of reverse bias voltage. Table 6.30 shows experimental capacitance/voltage data for a particular diode.

Assume that the relationship between C_j and V can be written,

$$\frac{1}{C_j^2} = k(V + \phi), \tag{6.50}$$

where ϕ is the contact potential and k is a constant. Use least squares to find best estimates of k and ϕ.

Table 6.29. *Equations to be transformed into the form $y = a + bx$.*

	Equation	Dependent variable	Independent variable	Constant(s)	Hint
(i)	$R = Ae^{-BT}$	R	T	A, B	Take the natural logs of both sides of the equation.
(ii)	$PV^{\gamma} = C$	P	V	γ, C	Take the logs of both sides of the equation.
(iii)	$H = C(T - T_o)$	H	T	C,T_o	Multiply out the brackets.
(iv)	$T_w = T_c - kR^2$	T_w	R	T_c, k	
(v)	$T = 2\pi\sqrt{\frac{m}{k}}$	T	m	k	Square both sides of the equation.
(vi)	$E = I(R+r)$	I	R	E, r	Move I to the LHS of equation, and E to the RHS.
(vii)	$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$	v	u	f	Move $\frac{1}{u}$ to the RHS of the equation.
(viii)	$N = kC^{1/n}$	N	C	k, n	Take logs of both sides of equation.
(ix)	$n^2 = 1 + \frac{A\lambda^2}{\lambda^2 - B}$	n	λ	A, B	Subtract 1 from both sides of equation then take the reciprocals of both sides of the equation.

Table 6.29. (*cont.*)

	Equation	Dependent variable	Independent variable	Constant(s)	Hint
(x)	$t = \frac{A}{D^2}(1 + BD)$	t	D	A, B	Multiply both sides of equation by D^2 then multiply out the brackets.
(xi)	$D = kE^n$	D	E	k, n	Take logs of both sides of the equation.
(xii)	$v = \sqrt{k(A^2 - l^2)}$	v	l	k, A	Square both sides of the equation.
(xiii)	$g = g_0\left(1 - \frac{h}{2R}\right)$	g	h	g_0, R	Multiply out the brackets.
(xiv)	$v = \sqrt{\frac{2(P - P_A)}{\rho}}$	v	P	P_A, ρ	Square both sides of the equation then multiply out the brackets.

Table 6.30. *Junction capacitance, C_j, of a diode as a function of bias voltage, V.*

V (V)	C_j (pF)
6.0	248
8.1	217
10.1	196
14.1	169
18.5	149
24.6	130
31.7	115
38.1	105
45.6	96.1
50.1	92.1

6.9.1 Consequences of data transformation

Our least squares analysis to this point has assumed that the uncertainties in all the y values are the same. Is this assumption still valid if the data are transformed? Very often the answer is no and to illustrate this let us consider a situation in which data transformation requires that the natural logarithms of the y quantity be calculated.

The intensity of light, I, after it travels a distance x through a transparent medium is given by

$$I = I_0 e^{-kx}, \tag{6.51}$$

where I_0 is the intensity of light incident on the medium (i.e. at $x = 0$) and k is the absorption coefficient.

Measurements of I are made at different values of x. We linearise equation 6.51 by taking natural logarithms of both sides of the equation, giving

$$\ln I = \ln I_0 - kx. \tag{6.52}$$

Assuming that equation 6.51 is valid for the experimental data, plotting $\ln I$ versus x should produce a plot in which the transformed data lie close to a straight line. A straight line fitted to the transformed data would have intercept, $\ln I_0$, and slope $-k$. As $\ln I$ is taken as the 'y quantity' when fitting a line to data using least squares, we must determine the uncertainty in $\ln I$. We write

$$y = \ln I. \tag{6.53}$$

If the standard uncertainty in I, $u(I)$, is small, then the standard uncertainty in y, $u(y)$, is given by[29]

$$u(y) \approx \left|\frac{\partial y}{\partial I}\right| u(I) \tag{6.54}$$

now, $\frac{\partial y}{\partial I} = \frac{1}{I}$, so that

$$u(y) \approx \left|\frac{u(I)}{I}\right|. \tag{6.55}$$

Equation 6.55 indicates that, if $u(I)$ is constant, the uncertainty in $\ln I$ decreases as I increases. The consequence of this is that the assumption of constant uncertainty in the y values used in least squares analysis is no longer valid and unweighted least squares must be abandoned in favour of an approach that takes into account changes in the uncertainties in the y values. There are many situations in which data transformation leads to a similar outcome, i.e. that the uncertainty in y values is not constant and so require a straight line to be fitted using *weighted* least squares. This is dealt with in the next section.

[29] Section 5.8.1 considers propagation of uncertainties.

Exercise Q

For the following equations, determine y and the standard uncertainty in y, $u(y)$, given that $V = 56$ and $u(V) = 2$ in each case. Express $u(y)$ to two significant figures.

(i) $y = V^{1/2}$; (ii) $y = V^2$; (iii) $y = \frac{1}{V}$; (iv) $y = \frac{1}{V^2}$; (v) $y = \log_{10} V$.

6.10 Weighted least squares

When the uncertainties in y values are constant, unweighted least squares analysis is appropriate as discussed in section 6.2.2. However, there are many situations in which the uncertainty in the y values does not remain constant. These include when:

(i) measurements are repeated at a particular value of x, thus reducing the uncertainty in the corresponding value of y;
(ii) data are transformed, for example by 'squaring' or 'taking logs';
(iii) the size of the uncertainty in y is some function of y (such as in counting experiments).

How do you know if you should use a weighted fit? A good starting point is to perform unweighted least squares to find the line of best fit through the data (data should be transformed if necessary, as discussed in section 6.9). A plot of the residuals should reveal whether a weighted fit is required. Figure 6.23 shows a plot of residuals in which the residuals decrease with increasing x (figure 6.23(a)) and increase with increasing x (figure 6.23(b)). Such patterns in residuals are 'tell tale' signs that weighted least squares fitting should be used.

In order to find the best line through the points when weighted fitting is required, we must include explicitly the standard deviation in each y value in the equations for a and b. This is due to the fact that, for weighted fitting, we must minimise the weighted sum of squares of residuals, χ^2, where[30]

$$\chi^2 = \sum \left[\frac{y_i - \hat{y}_i}{s_i}\right]^2. \tag{6.56}$$

Here s_i is the standard deviation in y_i, and a and b are given by

$$a = \frac{\sum \frac{x_i^2}{s_i^2} \sum \frac{y_i}{s_i^2} - \sum \frac{x_i}{s_i^2} \sum \frac{x_i y_i}{s_i^2}}{\Delta} \tag{6.57}$$

[30] Refer to appendix 3 for more details.

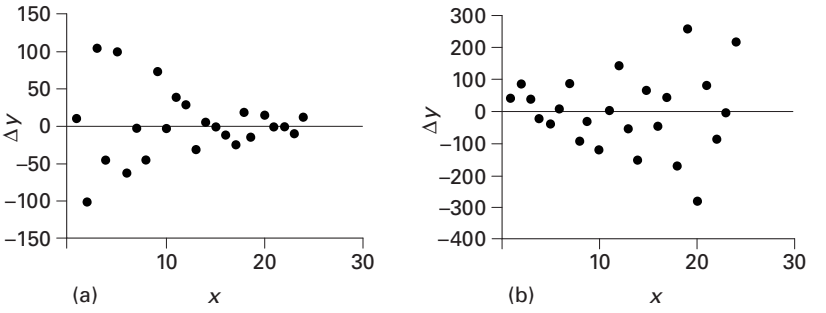

Figure 6.23. Residuals indicating weighted fit is required: (a) indicates the size of the residual decreasing with increasing x; (b) indicates the size of the residual increasing with increasing x.

$$b = \frac{\sum \frac{1}{s_i^2} \sum \frac{x_i y_i}{s_i^2} - \sum \frac{x_i}{s_i^2} \sum \frac{y_i}{s_i^2}}{\Delta}, \tag{6.58}$$

where

$$\Delta = \sum \frac{1}{s_i^2} \sum \frac{x_i^2}{s_i^2} - \left(\sum \frac{x_i}{s_i^2} \right)^2. \tag{6.59}$$

Equations 6.57 and 6.58 give more weight to the points that have smaller uncertainty, thereby ensuring that the fitted line will pass closer to these points than those with large uncertainty.

A word of caution regarding weighted fitting; if the estimates of the standard deviations, s_i, are poor then points will be weighted inappropriately. In such a situation an unweighted fit may be the best option.

With weighted fitting, the best line no longer passes through $(\bar{x}, \bar{y})$ but through the *weighted* centre of gravity of the points, $(\bar{x}_w, \bar{y}_w)$, and $\bar{x}_w$ and $\bar{y}_w$ are given by[31]

$$\bar{x}_w = \frac{\sum \frac{x_i}{s_i^2}}{\sum \frac{1}{s_i^2}} \tag{6.60}$$

and

$$\bar{y}_w = \frac{\sum \frac{y_i}{s_i^2}}{\sum \frac{1}{s_i^2}}. \tag{6.61}$$

[31] For a derivation of equation 6.60 see section A3.1 in appendix 3.

One difficulty with finding a and b is that, in many situations, we do not know s_i. When calculating a and b, a knowledge of the absolute values of s_i is not crucial. As long as we know the relative magnitudes of s_i for all data, the values of a and b are unaffected.

Example 10

Some x–y data along with the standard deviation in the y values, s_i, are shown in table 6.31. Using weighted least squares, find the intercept and slope of the best line through the points.

ANSWER

Table 6.32 contains all the quantities necessary to calculate a and b.

Summing the appropriate columns gives

$$\sum 1/s_i^2 = 0.178\,403, \sum x_i/s_i^2 = 15.009\,86, \sum y_i/s_i^2 = 15.465\,28,$$
$$\sum x_i y_i/s_i^2 = 1242.986, \sum x_i^2/s_i^2 = 1379.559.$$

Substituting the summations into equations 6.57 to 6.59 gives

$$a = 128.6, b = -0.4985.$$

Table 6.31. *x–y data for example 10.*

x_i	y_i	s_i
18	125	10
42	108	8
67	91	6
89	84	4
108	76	4

Table 6.32. *Quantities for weighted least squares calculation.*

x_i	y_i	s_i	$1/s_i^2$	x_i/s_i^2	y_i/s_i^2	$x_i y_i/s_i^2$	x_i^2/s_i^2
18	125	10	0.01	0.18	1.25	22.5	3.24
42	108	8	0.015 625	0.656 25	1.6875	70.875	27.5625
67	91	6	0.027 778	1.861 111	2.527 778	169.3611	124.6944
89	84	4	0.0625	5.5625	5.25	467.25	495.0625
108	76	4	0.0625	6.75	4.75	513	729

Exercise R

Repeat example 10 with every value of s_i multiplied by 5 (so, for example, when $y_i = 125$, $s_i = 50$). Show that the values of a and b remain unchanged (suggestion: use a spreadsheet!).

6.10.1 Weighted uncertainty in *a* and *b*

When a weighted fit is required,[32] equations 6.62 and 6.63 can be used to calculate the standard uncertainties in a and b:

$$s_a = \left(\frac{\sum \frac{x_i^2}{s_i^2}}{\Delta} \right)^{1/2} \tag{6.62}$$

$$s_b = \left(\frac{\sum \frac{1}{s_i^2}}{\Delta} \right)^{1/2}, \tag{6.63}$$

where Δ is given by equation 6.59.

Equations 6.62 and 6.63 are applicable as long as actual values of s_i are known, as relative magnitudes will not do in this case. Although this might seem unduly restrictive, there is one case in which s_i may be estimated fairly accurately and that is in counting experiments (such as those involving radioactivity or X-rays, where the Poisson distribution is valid). If the number of counts recorded is C_i, then the standard deviation in C_i, s_i, is given by

$$s_i = \sqrt{C_i}. \tag{6.64}$$

Table 6.33. *Variation of counts with depth in a solid.*

C (counts)	1.86×10^4	1.18×10^4	4.33×10^3	1.00×10^3	1.36×10^2
d (mm)	10	30	50	70	90

[32] See appendix 4 for more details.

Exercise S

In a diffusion experiment, a radiotracer diffuses into a solid when the solid is heated to a high temperature for a fixed period of time. The solid is sectioned and the number of gamma counts is recorded by a particle counter over a period of 1 minute for each section. Table 6.33 shows the number of counts, C, as a function of depth, d, cut in the material.

Assume that the equation that relates C to d is

$$C = A\exp(\lambda d^2),$$

where A and λ are constants.

(i) Transform this equation into the form $y = a + bx$.
(ii) Assuming the standard deviation in C to be $\sqrt{C}$, perform a weighted fit to find values of A, λ, s_A and s_λ.

6.10.2 Weighted standard deviation, s_w

When only the relative magnitudes of s_i are known, it is still possible to determine s_a and s_b. In this case it is necessary to calculate the weighted standard deviation of y values (i.e. the weighted equivalent to equation 6.10). Writing the weighted standard deviation as s_w, we have[33]

$$s_w = \frac{\left(\frac{n}{n-2}\right)^{1/2}}{\sum\frac{1}{s_i^2}}\left[\sum\frac{1}{s_i^2}\sum\frac{y_i^2}{s_i^2} - \left(\sum\frac{y_i}{s_i^2}\right)^2 - \frac{\left(\sum\frac{1}{s_i^2}\sum\frac{x_i y_i}{s_i^2} - \sum\frac{x_i}{s_i^2}\sum\frac{y_i}{s_i^2}\right)^2}{\Delta}\right]^{1/2} \tag{6.65}$$

where Δ is given by equation 6.59.

We write s_a and s_b as

$$s_a = s_w\left(\frac{\sum\frac{1}{s_i^2}\sum\frac{x_i^2}{s_i^2}}{n\Delta}\right)^{1/2} \tag{6.66}$$

and

$$s_b = \frac{s_w\sum\frac{1}{s_i^2}}{(n\Delta)^{1/2}}. \tag{6.67}$$

For completeness, we include the expression for the weighted linear correlation coefficient, r_w, which should be used whenever weighted least squares fitting occurs; r_w is given as,

[33] For a derivation of equation 6.65 see chapter 10 of Dietrich (1991).

$$r_w = \frac{\sum \frac{1}{s_i^2} \sum \frac{x_i y_i}{s_i^2} - \sum \frac{x_i}{s_i^2} \sum \frac{y_i}{s_i^2}}{\left[\sum \frac{1}{s_i^2} \sum \frac{x_i^2}{s_i^2} - \left(\sum \frac{x_i}{s_i^2}\right)^2\right]^{1/2} \left[\sum \frac{1}{s_i^2} \sum \frac{y_i^2}{s_i^2} - \left(\sum \frac{y_i}{s_i^2}\right)^2\right]^{1/2}}. \tag{6.68}$$

Example 11

The data shown in table 6.34 were obtained from a study of the relationship between current and voltage for a tunnel diode.[34]

For the range of voltage in table 6.34, the relationship between current, I, and voltage, V, for a tunnel diode can be written as

$$I = CV \exp\left(\frac{-V}{B}\right). \tag{6.69}$$

(i) Transform equation 6.69 into the form $y = a + bx$.

(ii) Determine the summations, $\sum \frac{1}{s_i^2}$, $\sum \frac{x_i}{s_i^2}$ etc. required in the calculation of the weighted standard deviation, s_w, given by equation 6.65. Assume that the uncertainties in the values of current are constant.

(iii) Calculate the weighted standard deviation.

ANSWER

(i) Dividing both sides of equation 6.69 by V gives

$$\frac{I}{V} = C \exp\left(\frac{-V}{B}\right). \tag{6.70}$$

Taking the natural logarithms of both sides of equation 6.70,

$$\ln \frac{I}{V} = \ln C - \frac{V}{B}. \tag{6.71}$$

Comparing equation 6.71 with $y = a + bx$, we find that $y = \ln \frac{I}{V}$, $x = V$, $a = \ln C$ and $b = -\frac{1}{B}$. Plotting $\ln \frac{I}{V}$ versus V would be expected to give a straight line with intercept equal to $\ln C$ and slope equal to $-\frac{1}{B}$.

(ii) In order to determine the standard deviation in the ith y value, s_i, we must consider the left hand side of equation 6.71:

$$y = \ln \frac{I}{V}. \tag{6.72}$$

[34] A tunnel diode is a semiconductor device with unusual electrical characteristics. It is sometimes used in high frequency oscillator circuits.

The standard deviation in the ith value y, s_i, is given by

$$s_i \approx \left|\frac{\partial y}{\partial I}\right| s_I, \tag{6.73}$$

now, $\frac{\partial y}{\partial I} = \frac{1}{I}$, so that

$$s_i \approx \frac{s_I}{I}. \tag{6.74}$$

Table 6.35 shows the raw data, the transformed data and the sums of the columns necessary to calculate the weighted standard deviation, s_w. For convenience we take the standard deviation in I, s_I, to be equal to 1, so that using equation 6.74, $s_i = \frac{1}{I}$.

(iii) The weighted standard deviation, calculated using equation 6.65 and the sums of numbers in appearing in the bottom row of table 6.35 is s_w = 0.1064.

Table 6.34. *Current–voltage data for a tunnel diode.*

V (V)	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
I (A)	0.082	0.130	0.136	0.210	0.181	0.180	0.190	0.172

V (V)	0.09	0.10	0.11	0.12	0.13	0.14	0.15
I (A)	0.136	0.150	0.108	0.119	0.110	0.070	0.080

Table 6.35. *Weighted fitting of data in table 6.33.*

V (V) = x	I (A)	$\ln(I/V) = y$	$s_i = \frac{1}{I}$	$\frac{1}{s_i^2}$	$\frac{x_i}{s_i^2}$	$\frac{y_i}{s_i^2}$	$\frac{x_i y_i}{s_i^2}$	$\frac{x_i^2}{s_i^2}$	$\frac{y_i^2}{s_i^2}$
0.01	0.082	2.104134	12.19512	0.006724	6.72E-05	0.014148	0.000141	6.72E-07	0.02977
0.02	0.13	1.871802	7.692308	0.0169	0.000338	0.031633	0.000633	6.76E-06	0.059212
0.03	0.136	1.511458	7.352941	0.018496	0.000555	0.027956	0.000839	1.66E-05	0.042254
0.04	0.21	1.658228	4.761905	0.0441	0.001764	0.073128	0.002925	7.06E-05	0.121263
0.05	0.181	1.286474	5.524862	0.032761	0.001638	0.042146	0.002107	8.19E-05	0.05422
0.06	0.18	1.098612	5.555556	0.0324	0.001944	0.035595	0.002136	0.000117	0.039105
0.07	0.19	0.998529	5.263158	0.0361	0.002527	0.036047	0.002523	0.000177	0.035994
0.08	0.172	0.765468	5.813953	0.029584	0.002367	0.022646	0.001812	0.000189	0.017334
0.09	0.136	0.412845	7.352941	0.018496	0.001665	0.007636	0.000687	0.00015	0.003152
0.1	0.15	0.405465	6.666667	0.0225	0.00225	0.009123	0.000912	0.000225	0.003699
0.11	0.108	−0.01835	9.259259	0.011664	0.001283	−0.00021	−2.4E-05	0.000141	3.93E-06
0.12	0.119	−0.00837	8.403361	0.014161	0.001699	−0.00012	−1.4E-05	0.000204	9.92E-07
0.13	0.11	−0.16705	9.090909	0.0121	0.001573	−0.00202	−0.00026	0.000204	0.000338
0.14	0.07	−0.69315	14.28571	0.0049	0.000686	−0.0034	−0.00048	9.6E-05	0.002354
0.15	0.08	−0.62861	12.5	0.0064	0.00096	−0.00402	−0.0006	0.000144	0.002529
			sums	0.307286	0.021316	0.290285	0.013336	0.001824	0.411229

Exercise T

Determine the intercept and slope of the transformed data in example 11. Calculate also the standard uncertainties in intercept and slope.

6.10.3 Weighted least squares and Excel

If the uncertainties in y values are small, then it matters little whether weighted or unweighted least squares is used to fit a straight line to data, as both will yield very similar values for intercept and slope. Owing to its comparative computational simplicity, it is wise to apply unweighted least squares first. Examining the residuals can help decide whether fitting an equation using weighted least squares is warranted. An added advantage of performing unweighted least squares first is that we have values for the 'unweighted' intercept and slope against which weighted estimates of intercept and slope can be compared. If there is large difference between unweighted and weighted estimates this might indicate a mistake in the weighted analysis.

If weighted fitting is required, it is quite daunting to attempt such an analysis equipped with only a pocket calculator. The evaluation of the many summations required for the determination of a and b (see equations 6.57, 6.58 and 6.59) is enough to deter all but the most tenacious person.

Most 'built in' least squares facilities available with calculators and spreadsheets offer only unweighted least squares. This is also true of the LINEST() feature in Excel which does not provide for the weighting of the fit. Nevertheless, a table, such as that appearing in tables 6.32 and 6.35, can be quite quickly drawn up in Excel. The table should include the 'raw x–y data' the standard deviation in each y value, s_i, and the other quantities, such as $x_i y_i / s_i^2$. Using AutoSum, Σ AutoSum ▾, which can be found in the Editing group on the Home Ribbon, permits the sums to be calculated with the minimum of effort. The calculations required for the examples in this chapter involving weighted least squares were carried out in this manner.

6.11 Review

An important goal in science is to progress from a qualitative understanding of phenomena to a quantitative description of the relationship between physical variables. Least squares is a powerful and widely used technique in the physical sciences (and in many other disciplines) for establishing a quantitative

relationship between two or more variables. In this chapter we have focussed upon fitting an equation representing a straight line to data, as linearly related data frequently emerge from experiments.

Due to the ease with which modern analysis tools such as spreadsheets can fit an equation to data, it is easy to overlook the question 'should we really fit a *straight* line to data?'. To assist in answering this question we introduced the correlation coefficient and residual plots as quantitative and qualitative indicators of 'goodness of fit' and have indicated situations in which each can be misleading. We have also considered situations in which data transformation is required before least squares is applied. Often, after completing the data transformation, we must forsake unweighted least squares in favour of weighted least squares.

The technique of fitting an equation to data using least squares can be extended to situations in which there are more than two parameters to be estimated, for example $y = \alpha + \frac{\beta}{x} + \gamma x$ (where α, β, and γ are the parameters) or where there are more than two independent variables. This will be considered in the next chapter.

End of chapter problems

(1) Table 6.36 contains values of acceleration due to gravity, g, measured at various heights, h, above sea level.

Taking the height as the independent variable (x) and the acceleration due to gravity as the dependent variable (y), do the following.

Table 6.36. *Variation of acceleration due to gravity with height.*

h (km)	g (m/s^2)
10	9.76
20	9.74
30	9.70
40	9.69
50	9.73
60	9.62
70	9.59
80	9.55
90	9.54
100	9.51

(i) Plot a graph of g versus h.
(ii) Assuming g to be linearly related to h, calculate the intercept, a, and slope, b, of the line of best fit.
(iii) Calculate the sum of squares of the residuals and the standard deviation of the y values.
(iv) Calculate the standard uncertainties in a and b.
(v) Determine the correlation coefficient.
(vi) The data pair $h = 50$ km, $g = 9.73$ m/s^2 was incorrectly recorded and should be replaced by $h = 50$ km, $g = 9.65$ m/s^2. Carry out the replacement and repeat parts (i) to (v).

(2) This question uses data from question (1), but with the data pair $h = 50$ km, $g = 9.73$ m/s^2 in table 6.36 replaced by $h = 50$ km, $g = 9.65$ m/s^2.

Assume that the relationship between g and h which is applicable to the data in table 6.36 can be written

$$g = g_0\left(1 - \frac{2h}{R_E}\right), \tag{6.75}$$

where g_0 is the acceleration due to gravity at the Earth's surface, and R_E is the radius of the Earth.

Use least squares to determine best estimates of g_0 and R_E and the standard uncertainty in each.

(3) When a rigid tungsten sphere presses on a flat glass surface, the surface of the glass deforms. The relationship between the mean contact pressure, p_m, and the indentation strain, s, is

$$p_m = ks. \tag{6.76}$$

Table 6.37 shows s–p_m data for indentations made into glass. k is a constant dependent upon the mechanical stiffness of the material.

Using least squares, determine the best estimate of k and the standard uncertainty in the best estimate.

(4) A thermal convection flowmeter measures the local speed of a fluid by measuring the heat loss from a heated element in the path of the flowing fluid. Assume the relationship between the speed of the fluid, u, and the electrical current, I, supplied to the element is

$$u = K_1(I^2 - K_2)^2, \tag{6.77}$$

where K_1 and K_2 are constants. Thirty values for u and I obtained in a water flow experiment are shown in table 6.38.

Assume I to be the dependent variable, and u the independent variable.

Table 6.37. *Variation of contact pressure with indentation strain.*

s	p_m (MPa)
0.0280	1.0334
0.0354	1.2295
0.0383	1.1851
0.0415	1.3574
0.0521	1.6714
0.0568	1.7977
0.0570	1.9303
0.0754	2.3957
0.0764	2.4258
0.0975	3.0208
0.1028	2.8824
0.1149	3.5145
0.1210	3.5072
0.1430	3.8888
0.1703	5.0148

Table 6.38. *Speed/current values in a water flow experiment.*

u (m/s)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1
I (μA) ± 10 μA	270	300	300	330	300	330	350	330	330	360	340
u (m/s)	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2
I (μA) ± 10 μA	340	360	350	370	340	370	360	360	380	380	390
u (m/s)	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0			
I (μA) ± 10 μA	390	400	410	410	400	400	410	410			

(i) Show that equation 6.77 can be rearranged into the form

$$I^2 = K_2 + \frac{1}{K_1^{1/2}} u^{1/2}. \tag{6.78}$$

(ii) Compare equation 6.78 with $y = a + bx$, and use a spreadsheet to perform a weighted least squares fit to data to find best estimates of K_1 and K_2.

(iii) Plot the standardised residuals. Can anything be concluded from the pattern of the residuals (for example, is the weighting you have used appropriate?).

(5) Acetic acid is adsorbed from solution by activated charcoal. The amount of acetic acid adsorbed, Y, is given by

$$Y = kC^{1/n}, \tag{6.79}$$

where C is the concentration of the acetic acid. k and n are constants.

C–Y data are presented in table 6.39.

(i) Transform equation 6.79 into the form $y = a + bx$.
(ii) Use unweighted least squares to find best estimates for k and n and the standard uncertainties in k and n.
(iii) Plot the standardised residuals. Do you think fitting by unweighted least squares is appropriate?
(iv) Determine the 95% coverage interval for Y when $C = 0.085$ mol/L.

Table 6.39. *Variation of adsorbed acetic acid with concentration.*

C (mol/L)	0.017	0.032	0.060	0.125	0.265	0.879
Y (mol)	0.46	0.61	0.79	1.10	1.53	2.45

(6) The variation of intensity, I, of light passing through a lithium fluoride crystal is measured as a function of the angular position of a polariser placed in the path of the light emerging from the crystal. Table 6.40 gives the intensity for various polariser angles, θ.

Assuming that the relationship between I and θ is

$$I = [I_{max} - I_{min}]\cos(2\theta) + I_{min}, \tag{6.80}$$

where I_{max} and I_{min} are constants:

(i) perform an unweighted fit to find the intercept and slope of a graph of I versus $\cos(2\theta)$;
(ii) use the values for the slope and intercept to find best estimates of I_{max} and I_{min};
(iii) Determine the standard uncertainty in the best estimate of I_{max}.

Table 6.40. *Variation of intensity with polariser angle.*

θ (degrees)	0	20	40	60	80	100	120	140
Intensity, I (arbitrary units)	1.86	1.63	1.13	0.52	0.16	0.00	0.57	1.11

(7) The Langmuir adsorption isotherm can be written as

$$\frac{P}{X} = \frac{1}{A} - \frac{B}{A}P, \tag{6.81}$$

where X is the mass of gas adsorbed per unit area of surface, P, is the pressure of the gas, and A and B are constants.

Table 6.41 shows data gathered in an adsorption experiment.

(i) Write equation 6.81 in the form $y = a + bx$. How do the constants in equation 6.81 relate to a and b?
(ii) Using least squares, find values for a and b and standard uncertainties in a and b.
(iii) Determine the standard uncertainties in A and B.

Table 6.41. *Gas adsorbed as a function of pressure.*

P (N/m^2)	X (kg/m^2)
0.27	13.9×10^{-5}
0.39	17.8×10^{-5}
0.62	22.5×10^{-5}
0.93	27.5×10^{-5}
1.72	32.9×10^{-5}
3.43	38.6×10^{-5}

(8) The relationship between mean free path, λ, of electrons moving through a gas and the pressure of the gas, P, can be written

$$\frac{1}{\lambda} = \left(\frac{\pi d^2}{4kT}\right)P, \tag{6.82}$$

where d is the diameter of the gas molecules, k is Boltzmann's constant ($= 1.38 \times 10^{-23}$ J/K) and T is the temperature of the gas in kelvins. Table 6.42 shows data gathered in an experiment in which λ was measured as a function of P.

(i) Fit the equation $y = a + bx$ to equation 6.82 to find b and the standard uncertainty in b.
(ii) If the temperature of the gas is 298 K, use equation 6.82 to estimate the diameter of the gas molecules and the standard uncertainty in the diameter.

(9) The intercept on the x axis x_{INT} of the best straight line through data is found by setting $\hat{y} = 0$ in equation 6.3. x_{INT} is given by

Table 6.42. *Variation of free path, λ, with pressure.*

λ (mm)	P (Pa)
35	1.5
20	5.8
30	5.8
25	6.5
16	8.0
9	13.1
10	14.5
7	18.9
6	24.7
6.5	27.6

$$x_{INT} = -\frac{a}{b}.$$

Given that standard uncertainties in $\bar{y}$ and b are

$$s_{\bar{y}} = \frac{s}{\sqrt{n}} \text{ and } s_b = \frac{sn^{1/2}}{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]^{1/2}},$$

show that the standard uncertainty in the intercept, $s_{x_{INT}}$, can be expressed as

$$s_{x_{INT}} = \frac{s}{b}\left[\frac{1}{n} + \frac{n\bar{y}^2}{b^2\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)}\right]^{1/2}. \tag{6.83}$$

(10) In section 6.4.2, we found that

$$\hat{y}_0 = \bar{y} + b(x_0 - \bar{x}).$$

Show that if the errors in $\bar{y}$ and b are not correlated, the standard uncertainty in $\hat{y}_0$, $s_{\hat{y}_0}$, can be written

$$s_{\hat{y}_0} = s\left[\frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}\right]^{1/2}, \tag{6.84}$$

where expressions for standard uncertainties in $\bar{y}$ and b are given in question 9.
(11) The equation of a straight line passing through the origin can be written, $y = bx$.

Use the method outlined in appendix 3 to show that the slope, b, of the best line to pass through the points and the origin is given by

$$b = \frac{\sum x_i y_i}{\sum x_i^2}.$$

Use the method outlined in appendix 4 to show that the standard uncertainty in b, s_b, is given by

$$s_b = \frac{s}{\left(\sum x_i^2\right)^{1/2}}, \tag{6.85}$$

where s is given by equation 6.10.

A word of caution: in situations involving 'real' data it is advisable not to treat the origin as a special point by forcing a line through it, even if 'theoretical' considerations predict that the line *should* pass through the origin. In most situations, random and systematic errors conspire to move the line so that it does not pass through the origin.

(12) Data obtained in an experiment designed to study whether the amount of CaO in rocks drawn from a geothermal field is correlated to the amount of MgO in those rocks are shown in table 6.43.

(i) Plot a graph of concentration of CaO versus concentration of MgO.
(ii) Calculate the linear correlation coefficient, r.
(iii) Is the value of r obtained in (ii) significant?

Table 6.43. *Concentrations of CaO and MgO in rock specimens.*

CaO (wt%)	MgO (wt%)
2.43	0.72
4.90	2.42
10.43	4.75
15.64	3.99
16.62	5.39
21.12	8.65

(13) A thermoelectric generator (TEG) is a device capable of converting thermal energy to electrical energy. Table 6.44 shows how the voltage, V, at the terminals of the TEG varies with the electrical load, R, to which the TEG is attached.

The relationship between the current, I, through the load and R can be written as

$$\frac{1}{I} = \frac{r}{E} + \frac{R}{E}. \tag{6.86}$$

(i) Use Excel to complete the third column in table 6.44.
(ii) Plot a graph of $1/I$ versus R.
(iii) By comparing $y = a + bx$ with equation 6.86, show that $a = r/E$ and $b = 1/E$.
(iv) Use unweighted least squares to determine the best estimates of r and E and the standard uncertainty in each.
(v) Is fitting by unweighted least squares justified in this situation?

Table 6.44. *Relationship between load attached to a TEG and the voltage across the terminals of the TEG.*

$R\,(\Omega)$	$V\,(\text{mV})$	$1/I = R/V\,(\text{mA}^{-1})$
1	11.2	0.089 286
2	16.9	0.118 343
3	21.1	
4	24.1	
5	26.1	
6	28.8	
7	30.1	
8	30.6	
9	31.5	
10	32.4	

(14) A *Spathiphyllum wallisii* 'Petite' indoor potted plant is sealed in a perspex chamber with field capacity water[35] at 23 °C. The amount of CO_2 in the chamber is measured over a period of 6 hours using an infrared gas analyser. Data gathered are shown in table 6.45.

Assuming that the relationship between CO_2 concentration and time is linear:

(i) fit a straight line to the data using unweighted least squares and determine the slope, intercept and the standard uncertainties in slope and intercept;
(ii) calculate the correlation coefficient, r;
(iii) use the line of best fit to estimate the concentration of CO_2 at $t = 150$ min and $t = 400$ min;
(iv) calculate the 95% coverage interval for the true value of the concentration of CO_2 at $t = 150$ min and $t = 400$ min.

[35] Field capacity water is the amount of water held in the soil after excess water has drained away.

Table 6.45. *Variation of CO_2 concentration with time.*

t (min)	CO_2 concentration (ppm)
0	404
60	478
120	556
180	633
240	716
300	796
360	880

(15) In a study on skin cancer, the concentration of antimony in lymph nodes was established using the technique of Inductivity Coupled Plasma Mass Spectrometry (ICP-MS). An ICP-MS instrument was calibrated using samples containing known amounts of antimony. Table 6.46 shows the output of the instrument (in counts per second, cps) for various antimony concentrations (antimony concentration is expressed in parts per billion, ppb).

(i) Plot a calibration graph of instrument output versus antimony concentration.
(ii) Fit a straight line to the data in table 6.46 using unweighted least squares and determine the slope and intercept, as well as the standard uncertainties in the slope an intercept.
(iii) Three repeat measurements were made on a sample containing unknown antimony concentration. The mean of the three values is 6255 cps. Use this information to find the best estimate of the antimony concentration in the sample and the 95% coverage interval for the true concentration.

Table 6.46. *Calibration data for an ICP-MS.*

Concentration of antimony (ppb)	Output of ICP-MS (cps)
0	71
0.5	1426
1.0	2707
2.0	4974
5.0	12982
10.0	26062
20.0	53135

(16) When a photon is emitted from a nucleus, the photon's energy is shifted due to the recoil of the nucleus. The relationship between the energy of the photon, E, and the angle, θ, between the direction of the recoil of the nucleus and the focal plane containing the photon detectors may be written

$$E = E_{\mathrm{o}}\left(1 + \frac{v}{c}\cos(\theta)\right), \tag{6.87}$$

where E_{o} is the energy of the photon in the absence of recoil, v is the recoil velocity of the nucleus, and c is the speed of light. In this question take c to be equal to 2.998×10^8 m/s. See also the data in Table 6.47.

(i) Plot a graph of E versus $\cos(\theta)$.
(ii) Use unweighted least squares to fit equation 6.87 to the E versus $\cos(\theta)$ data.
(iii) Using the fit to data, estimate E_{o} and v and the standard uncertainties in these.
(iv) Use a plot of residuals to provide evidence as to the goodness of fit of the equation to data.

Table 6.47. *Variation of photon energy, E, with angle, θ.*

angle, θ (degrees)	energy, E (keV)
31.72	468.10
37.38	467.61
50.07	464.95
58.28	463.04
69.82	460.19
79.19	457.60
80.71	457.20
90.00	454.87
99.29	452.40
100.81	451.90
110.18	449.55
121.72	446.90
129.93	445.31
142.62	443.13
148.28	442.17
162.73	440.50

(17) As part of a study on the effect of climate change on the survival of Australian plant species, leaf temperature was recorded for leaves of varying size. Table 6.48 shows how the temperature ΔT, where ΔT = (leaf temperature – ambient temperature) varies with the surface area, A, of leaves as recorded during on a hot summer's day.

Assuming that the relationship between ΔT and A can be written

$$\Delta T = k \log_{10} A + D. \tag{6.88}$$

(i) Plot a graph of ΔT versus $\log_{10} A$.
(ii) Fit a straight line to ΔT versus $\log_{10} A$ data using unweighted least squares and find best estimates for k and D.
(iii) Determine standard uncertainties in best estimates of k and D.
(iv) Plot the residuals. Does it appear as though a weighted or unweighted fit is appropriate?

Table 6.48. *Variation of ΔT with surface area of leaves.*

Leaf area, A, (cm^2)	ΔT (°C)	Leaf area, A (cm^2)	ΔT (°C)
1.19	3.5	54.89	6.3
1.58	2.7	56.09	5.0
1.69	3.3	70.82	4.0
3.50	3.1	72.24	4.5
8.92	4.2	83.80	7.1
11.23	3.8	115.94	4.7
14.48	5.6	133.57	7.0
17.32	5.3	151.90	9.7
18.68	5.4	229.78	5.7
20.56	5.1	238.66	4.8
21.63	3.4	300.01	7.9
21.79	3.7	314.41	8.8
38.93	6.3	360.05	6.4

(18) As the random error of individual y values increases, so do the standard uncertainties in slope and intercept. Figure 6.24 shows an Excel spreadsheet in which 'error-free' y values have been generated in the B column using the relationship $y = 1.54\,x + 4$. Errors, which have a normal distribution with mean of zero and standard deviation of 1, were generated in the C column using the Random Number Generator in the Analysis Toolpak. The D column contains the sum of the errors and the error-free y values. The LINEST() function was used to determine the best estimates of slope and intercept and standard uncertainties in these.

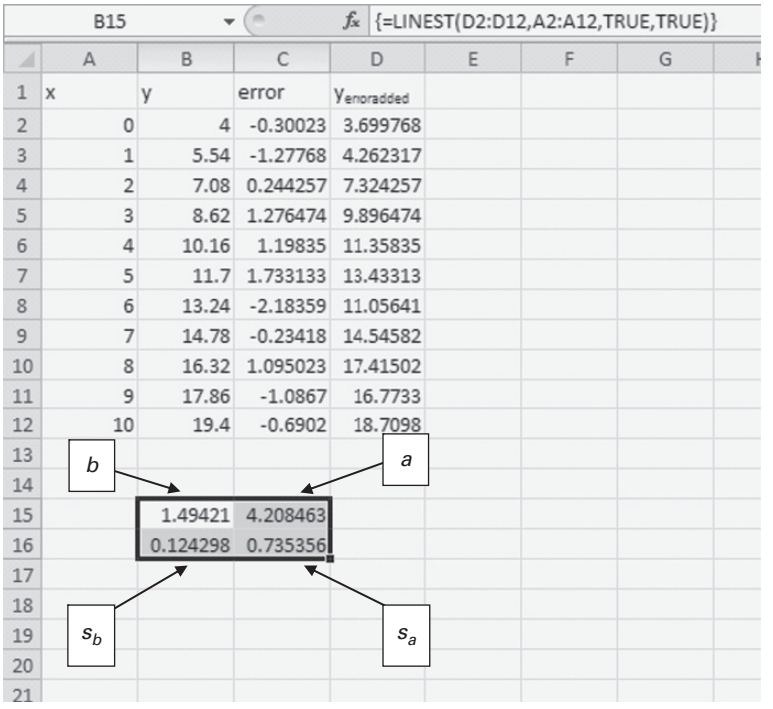

Figure 6.24. Illustration of effect of normally distributed errors on parameter estimates.

(i) Keeping the range of x values the same as in figure 6.24, change the interval between successive x values from 1 to 0.2, so that the total number of x values in the range 0 to 10 increases from 11 to 51. Repeat the procedure described above of adding errors to the y values of mean = 0 and standard deviation = 1.

(ii) What is the effect on the standard uncertainties in the slope and intercept of increasing the number of points?

(iii) By examining the relationship between the standard deviation of sample means and the number of values used to calculate the mean of a sample (see equation 3.21) how might you expect the standard uncertainties in slope and intercept to depend on the number of data points?

(iv) If you increase the number of points from 51 to 501 (while keeping the range and the error distribution as before), how would you expect the standard uncertainties in slope and intercept to change? Try this.