Chapter 7
Least squares II

7.1 Introduction

In chapter 6 we considered fitting a straight line to x–y data in situations in which the equation representing the line is written

$$y = a + bx,$$

where a is the intercept of the line and b is its slope. Other situations that we need to consider, as they occur regularly in the physical sciences, require fitting equations to x–y data where the equations:

- possess more than two parameters that must be estimated, for example, $y = a + bx + cx^2$ or $y = a + \frac{b}{x} + cx$;
- contain more than one independent variable, for example $y = a + bx + cz$ where x and z are the independent variables;
- cannot be written in a form suitable for analysis by linear least squares, for example $y = a + be^{cx}$.

As in chapter 6, we apply the technique of least squares to obtain best estimates of the parameters in equations to be fitted to data. As the number of parameters increases, so does the number of the calculations required to estimate those parameters. Matrices are introduced as a means of solving for parameter estimates efficiently. Matrix manipulation is tedious to carry out 'by hand', but the built in matrix functions in Excel make it well suited to assist in extending the least squares technique. In addition, Excel's dedicated least squares function, LINEST(), is able to fit equations to data where those equations contain more than two parameters.

We also consider in this chapter the important issue of choosing the best equation to fit to data, and what steps to take if two (or more) equations fitted to the same data must be compared.

7.2 Extending linear least squares

In chapter 6 we introduced α and β as (population) parameters which, in the absence of systematic errors, are regarded as the 'true' intercept and slope respectively of a straight line drawn through x–y data. Estimates of these parameters, obtained using least squares, were written as a and b. As we extend the least squares technique, we must deal with an increased number of parameters. Consistent with our earlier choice of symbols, we write estimates of parameters obtained by least squares as a, b, c, d and so on.

Fitting a straight line to data has a central position in the analysis of experimental data in the physical sciences. This is due to the fact that many physical quantities (at least over a limited range) are linearly related to each other. For example, consider the relationship between the output voltage of a thermocouple and the temperature of one junction of a thermocouple[1] as shown in figure 7.1.

There is little departure from linearity exhibited by the data in figure 7.1. However, the linearity of output voltage with temperature for the thermocouple is no longer maintained when a much larger range of temperature is considered, as shown in figure 7.2.

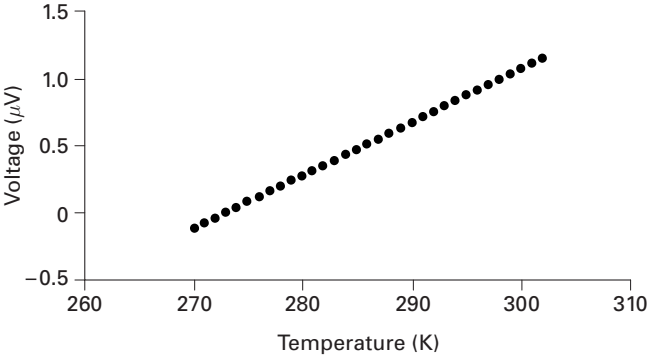


Figure 7.1. Output voltage of a thermocouple between 270 K and 303 K.

[1] The graph shown is for a Type K thermocouple. The reference junction of the thermocouple was held at 273 K.

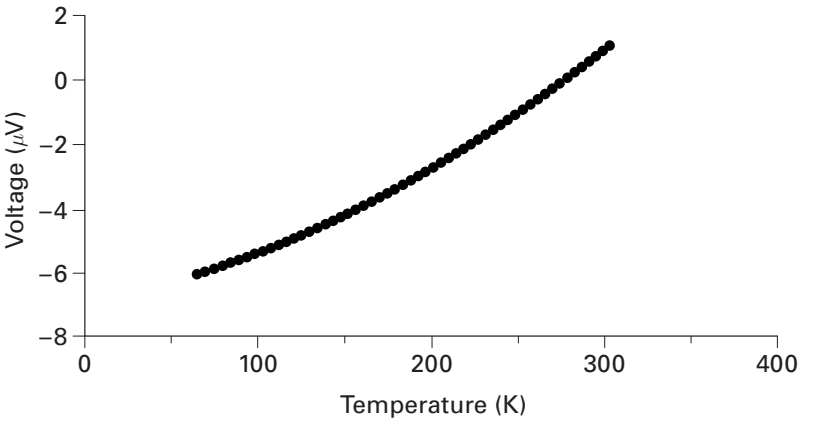


Figure 7.2. Output voltage of a thermocouple between 70 K and 303 K.

If x–y data recorded during an experiment are not linearly related, then they can often be 'linearised' by some straightforward transformation (such as taking the logarithms of the dependent variable, as described in section 6.9). In other situations, such as for the thermocouple voltage–temperature-values shown in figure 7.2, no amount of legitimate mathematical manipulation will linearise these data, and we must consider fitting an equation other than $y = a + bx$ to the data.

The technique of linear least squares is not confined to finding the best straight line through data, but can be extended to any function linear in the parameters appearing in the equation. As examples, a, b, and c in the following equations can be determined using linear least squares:

$$y = a + bx + cx^2;$$
$$y = a + bx + cx \ln x;$$
$$y = a + \frac{b}{x} + cx;$$
$$y = \frac{a}{x} + bx + cx^2.$$

In fitting functions with more than two parameters to data, we continue to assume that:

- random errors affecting y values are normally distributed;
- there are no random errors in the independent (x) values;
- systematic errors in both x and y values are negligible.

The approach to finding an equation which best fits the data, where the equation incorporates more than two parameters follows that described in

appendix 3. Though the complexity of the algebraic manipulations increases as the number of parameters increases, this can be overcome by using matrices to assist in parameter estimation.

7.3 Formulating equations to solve for parameter estimates

To determine best estimates of the parameters in an equation using the technique of least squares, we begin with the weighted sum of squares of residuals, χ^2, given by[2]

$$\chi^2 = \sum \left[\frac{y_i - \hat{y}_i}{s_i}\right]^2, \tag{7.1}$$

where y_i is the ith value of y (obtained through measurement), $\hat{y}_i$ is the corresponding predicted value of y found using an equation relating y to x, and s_i is the standard deviation in the ith value of y.

As an example of fitting an equation with more than two parameters to data, consider the equation

$$y = a + bx + cx^2. \tag{7.2}$$

To determine a, b and c, replace $\hat{y}_i$ in equation 7.1 by $a + bx_i + cx_i^2$, so that

$$\chi^2 = \sum \left[\frac{y_i - (a + bx_i + cx_i^2)}{s_i}\right]^2. \tag{7.3}$$

If the standard deviation in y_i is the same for all x_i then s_i is replaced by s and brought outside the summation, so that

$$\chi^2 = \frac{1}{s^2}\sum (y_i - a - bx_i - cx_i^2)^2. \tag{7.4}$$

To find a, b and c that minimise χ^2, differentiate χ^2 with respect to a, b and c in turn then set the resulting equations equal to zero.

Differentiating equation 7.4 gives

$$\frac{\partial \chi^2}{\partial a} = \frac{-2}{s^2}\sum (y_i - a - bx_i - cx_i^2) = 0 \tag{7.5}$$

$$\frac{\partial \chi^2}{\partial b} = \frac{-2}{s^2}\sum x_i(y_i - a - bx_i - cx_i^2) = 0 \tag{7.6}$$

$$\frac{\partial \chi^2}{\partial c} = \frac{-2}{s^2}\sum x_i^2(y_i - a - bx_i - cx_i^2) = 0. \tag{7.7}$$

[2] Appendix 3 explains the origin of χ^2.

Rearranging equations 7.5 to 7.7 gives[3]

$$na + b\sum x_i + c\sum x_i^2 = \sum y_i \tag{7.8}$$

$$a\sum x_i + b\sum x_i^2 + c\sum x_i^3 = \sum x_i y_i \tag{7.9}$$

$$a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 = \sum x_i^2 y_i. \tag{7.10}$$

We can rearrange equations 7.8 to 7.10 and substitute from one equation into another to solve for a, b and c. However, this approach is labour intensive, time consuming, and the likelihood of a numerical or algebraic mistake is quite high. If the number of parameters to be estimated increases to four or more, then solving for these estimates by 'elimination and substitution' becomes even more formidable. An effective way to proceed is to write the equations in matrix form, as solving linear equations using matrices is efficient, especially if software is available that can manipulate matrices.

Writing equations 7.8 to 7.10 in matrix form gives

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}. \tag{7.11}$$

Equation 7.11 can be written concisely as[4]

$$\mathbf{AB} = \mathbf{P}, \tag{7.12}$$

where

$$\mathbf{A} = \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \quad \mathbf{P} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}.$$

To determine elements, a, b, and c of $\mathbf{B}$ (which are the parameter estimates appearing in equation 7.2), equation 7.12 is manipulated to give[5]

$$\mathbf{B} = \mathbf{A}^{-1}\mathbf{P}, \tag{7.13}$$

where $\mathbf{A}^{-1}$ is the inverse matrix of the matrix, $\mathbf{A}$. Matrix inversion and matrix multiplication are tedious to perform without the aid of a computer, especially if matrices are large. The built in matrix functions in Excel are well suited to estimating parameters in linear least squares problems.

[3] Note that $\sum_{i=1}^{i=n} a = a + a + a + \cdots + a = na$.

[4] By convention, matrices are indicated in bold type.

[5] See appendix 5 for an introduction to matrices.

Exercise A

The relationship between voltage across a semiconductor diode, V, and temperature (in kelvin) of the diode, T, is given by

$$V = \alpha + \beta T + \gamma T \ln T, \tag{7.14}$$

where α, β and γ are constants.

An experiment is performed in which the voltage across the diode is measured as the temperature of the diode changes. Express, in matrix form, the equations that must be solved to obtain best estimates of α, β and γ (assume an unweighted fit is appropriate).

7.4 Matrices and Excel

Among the many built in functions provided by Excel are those related to matrix inversion and matrix multiplication. These functions can be used to solve for the parameter estimates (and the standard uncertainties in the estimates) appearing in an equation fitted to data.

7.4.1 The MINVERSE() function

An important step in parameter estimation using matrices is that of matrix inversion. The Excel function MINVERSE() may be used to invert a matrix such as that appearing in sheet 7.1.

The syntax of the function is

$$= \text{MINVERSE}(\text{array})$$

where array contains the elements of the matrix to be inverted.

Sheet 7.1. *Cells containing a square matrix.*

	A	B	C
1	2.6	7.3	3.4
2	9.5	4.5	5.5
3	6.7	2.3	7.8
4			

Example 1

Use the MINVERSE() function to invert the matrix shown in sheet 7.1.

ANSWER

Inverting a 3 × 3 matrix creates another 3 × 3 matrix. As the MINVERSE() function returns an array with nine elements, we highlight nine cells (cells E1:G3) into which Excel can return those elements. Sheet 7.2 shows the highlighted cells along with the function =MINVERSE(A1:C3) typed into cell E1.

To complete the inversion of the matrix shown in columns A to C in sheet 7.2, it is necessary to hold down the CTRL and SHIFT keys together then press the ENTER key. Sheet 7.3 shows the elements of the inverted matrix in columns E to G.

Sheet 7.2. *Example of matrix inversion using the MINVERSE() function in Excel.*

	A	B	C	D	E	F	G
1	2.6	7.3	3.4		=MINVERSE(A1:C3)		
2	9.5	4.5	5.5				
3	6.7	2.3	7.8				
4							

Sheet 7.3. *Outcome of matrix inversion.*

	A	B	C	D	E	F	G
1	2.6	7.3	3.4		−0.09285	0.203164	−0.10278
2	9.5	4.5	5.5		0.154069	0.01034	−0.07445
3	6.7	2.3	7.8		0.034329	−0.17756	0.238445
4							

Exercise B

Use Excel to invert the following matrices:

$$\text{(i)}\begin{bmatrix} 2.5 & 9.9 & 6.7 \\ 7.8 & 3.0 & 6.9 \\ 8.8 & 4.5 & 3.2 \end{bmatrix}; \quad \text{(ii)}\begin{bmatrix} 62 & 41 & 94 & 38 \\ 42 & 65 & 41 & 87 \\ 102 & 45 & 76 & 32 \\ 91 & 83 & 14 & 21 \end{bmatrix}; \quad \text{(iii)}\begin{bmatrix} 2.3 & 1.8 & 4.4 & 9.4 & 6.9 \\ 6.7 & 9.9 & 3.4 & 3.3 & 7.0 \\ 8.4 & 3.4 & 8.2 & 3.9 & 2.9 \\ 9.4 & 6.6 & 9.0 & 6.6 & 5.6 \\ 10.4 & 1.5 & 1.7 & 5.7 & 4.6 \end{bmatrix}.$$

7.4.2 The MMULT() function

Matrix multiplication is required when performing least squares using matrices. The MMULT() function in Excel is an array function which returns the product of two matrices.

Suppose $\mathbf{P} = \mathbf{A}\,\mathbf{B}$, where

$$\mathbf{A} = \begin{bmatrix} 2.6 & 7.3 & 3.4 \\ 9.5 & 4.5 & 5.5 \\ 6.7 & 2.3 & 7.8 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 34.4 \\ 43.7 \\ 12.3 \end{bmatrix}.$$

The elements of the **P** matrix can be found using the MMULT() function. The syntax of the function is

$$= \text{MMULT}(\text{array1}, \text{array2})$$

where array1 and array2 contain the elements of the matrices to be multiplied together.

Example 2

Use the MMULT() function to determine the product, **P**, of the matrices **A** and **B** shown in sheet 7.4.

ANSWER

Multiplying the 3 × 3 matrix by the 3 × 1 matrix appearing in sheet 7.4, produces another matrix of dimension 3 × 1. As the MMULT() function returns an array containing the elements of the matrix **P**, we highlight cells (shown in column G of sheet 7.5) into which those elements can be returned.

To determine **P**, type **=MMULT(A2:C4,E2:E4)** into cell G2. Holding down the CTRL and SHIFT keys, then pressing the Enter key returns the elements of the **P** matrix into cells G2 to G4 as shown in sheet 7.6.

Sheet 7.4. *Two matrices,* **A** *and* **B** *to be multiplied.*

	A	B	C	D	E	F	G
1	A				B		
2	2.6	7.3	3.4		34.4		
3	9.5	4.5	5.5		43.7		
4	6.7	2.3	7.8		12.4		

Sheet 7.5. *Matrix multiplication using MMULT().*

	A	B	C	D	E	F	G	H
1	A				B		P	
2	2.6	7.3	3.4		34.4		=MMULT(A2:C4,E2:E4)	
3	9.5	4.5	5.5		43.7			
4	6.7	2.3	7.8		12.4			

Sheet 7.6. *Outcome of multiplying* **A** *and* **B** *appearing in sheet 7.4.*

	G
1	P
2	450.61
3	591.65
4	427.71

Exercise C

Use Excel to carry out the following matrix multiplications:

$$\text{(i)}\begin{bmatrix} 56.8 & 123.5 & 67.8 \\ 87.9 & 12.5 & 54.3 \\ 23.6 & 98.5 & 56.7 \end{bmatrix}\begin{bmatrix} 23.1 \\ 34.6 \\ 56.8 \end{bmatrix}; \quad \text{(ii)}\begin{bmatrix} 12 & 45 & 67 & 56 \\ 34 & 54 & 65 & 43 \\ 12 & 54 & 49 & 31 \\ 84 & 97 & 23 & 99 \end{bmatrix}\begin{bmatrix} 32 \\ 19 \\ 54 \\ 12 \end{bmatrix}.$$

7.4.3 Fitting the polynomial $y = a + bx + cx^2$ to data

We bring together the matrix approach to solving for estimated parameters discussed in section 7.3 with the built in matrix functions in Excel to determine a, b, and c in the equation $y = a + bx + cx^2$. Consider the data in table 7.1.

Assuming it is appropriate to fit the equation $y = a + bx + cx^2$ to the data in table 7.1, we can find a, b and c by constructing the matrices

$$\mathbf{A} = \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \tag{7.15}$$

$$\mathbf{B} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \tag{7.16}$$

and

$$\mathbf{P} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}. \tag{7.17}$$

Using the data in table 7.1 (and with the assistance of Excel), we find

$$\sum x_i = 108;\ \sum y_i = 2845.6;\ \sum x_i^2 = 1536;\ \sum x_i^3 = 24192;$$
$$\sum x_i^4 = 405312;\ \sum x_i y_i = 45206.6;\ \sum x_i^2 y_i = 761100.4;\ n = 9.$$

Table 7.1. *x–y data.*

x	y
4	30.4
6	51.2
8	101.6
10	184.4
12	262.6
14	369.6
16	479.4
18	601.5
20	764.9

Matrices **A** and **P** become

$$\mathbf{A} = \begin{bmatrix} 9 & 108 & 1536 \\ 108 & 1536 & 24192 \\ 1536 & 24192 & 405312 \end{bmatrix} \quad \text{and} \quad \mathbf{P} = \begin{bmatrix} 2845.6 \\ 45206.6 \\ 761100.4 \end{bmatrix}.$$

As $\mathbf{B} = \mathbf{A}^{-1}\mathbf{P}$, we find (using Excel for matrix inversion and multiplication)[6]

$$\begin{aligned} \mathbf{B} &= \begin{bmatrix} a \\ b \\ c \end{bmatrix} \\ &= \begin{bmatrix} 3.505 & -6.214 \times 10^{-1} & 2.381 \times 10^{-2} \\ -6.214 \times 10^{-1} & 1.210 \times 10^{-1} & -4.870 \times 10^{-3} \\ 2.381 \times 10^{-2} & -4.870 \times 10^{-3} & 2.029 \times 10^{-4} \end{bmatrix} \begin{bmatrix} 2845.6 \\ 45206.6 \\ 761100.4 \end{bmatrix} \\ &= \begin{bmatrix} 1.9157 \\ -2.7458 \\ 2.0344 \end{bmatrix}, \end{aligned}$$

so that the relationship between x and y can be written,

$$y = 1.916 - 2.746x + 2.034x^2.$$

7.4.4 The X and X^T matrices

The elements of the **A** matrix given by equation 7.15 can be determined by constructing an Excel worksheet to calculate $\sum x$, $\sum x^2$ and so on, column by

[6] For convenience, the elements of $\mathbf{A^{-1}}$ are shown to four figures, but full precision (to 15 figures) is used 'internally' by Excel when calculations are carried out.

column, However there is a more efficient way to determine the elements in equation 7.15. We introduce the matrix, **X**, given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_i & x_i^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}. \tag{7.18}$$

Transposing the matrix, **X**, gives[7]

$$\mathbf{X^T} = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & x_i & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & x_i^2 & \cdots & x_n^2 \end{pmatrix}. \tag{7.19}$$

The elements of **A** matrix emerge by multiplying $\mathbf{X^T}$ and **X** together, i.e.

$$\mathbf{A} = \mathbf{X^T X}. \tag{7.20}$$

The matrix **X** may be formed in Excel and Excel's TRANSPOSE() function may be used to form $\mathbf{X^T}$.

Similarly, if the **Y** matrix is given by

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_i \\ \vdots \\ y_n \end{pmatrix} \tag{7.21}$$

then the **P** matrix given by equation 7.17 may be calculated using

$$\mathbf{P} = \mathbf{X^T Y}. \tag{7.22}$$

As $\mathbf{B} = \mathbf{A}^{-1}\mathbf{P}$ (see section 7.3), then

$$\mathbf{B} = \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T Y}. \tag{7.23}$$

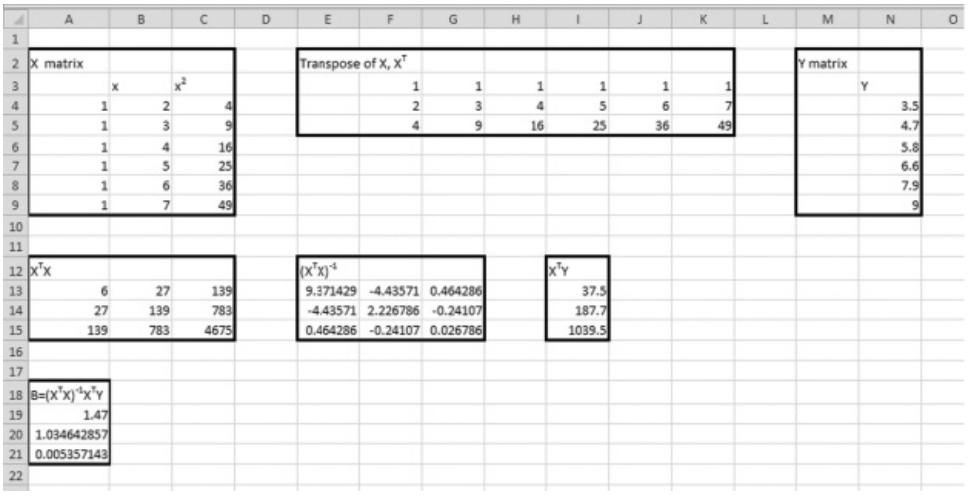
Example 3

Given that

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \end{pmatrix} \text{ and } \mathbf{Y} = \begin{pmatrix} 3.5 \\ 4.7 \\ 5.8 \\ 6.6 \\ 7.9 \\ 9.0 \end{pmatrix},$$

(1) determine the matrices, $\mathbf{X^T X}$ and $\mathbf{X^T Y}$ and $\left(\mathbf{X^T X}\right)^{-1}$;
(2) use equation 7.23 to find the elements of **B**.

[7] $\mathbf{X^T}$ is used to represent the transpose of **X**.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
2	X matrix				Transpose of X, X^T								Y matrix		
3		x	x^2			1	1	1	1	1	1			Y	
4	1	2	4			2	3	4	5	6	7			3.5	
5	1	3	9			4	9	16	25	36	49			4.7	
6	1	4	16											5.8	
7	1	5	25											6.6	
8	1	6	36											7.9	
9	1	7	49											9	
10															
11															
12	X^TX				$(X^TX)^{-1}$				X^TY						
13	6	27	139		9.371429	-4.43571	0.464286		37.5						
14	27	139	783		-4.43571	2.226786	-0.24107		187.7						
15	139	783	4675		0.464286	-0.24107	0.026786		1039.5						
16															
17															
18	$B=(X^TX)^{-1}X^TY$														
19	1.47														
20	1.034642857														
21	0.005357143														
22															

Figure 7.3. Use of Excel to transpose, invert and multiply matrices.

ANSWER

Figure 7.3 shows an Excel worksheet used to calculate $\mathbf{X^TX}$, $\mathbf{X^TY}$, $(\mathbf{X^TX})^{-1}$ and $(\mathbf{X^TX})^{-1}\mathbf{X^TY}$.

Cells A19 to A21 shown in figure 7.3 contain the elements of the **B** matrix.

Exercise D

The variation of the electrical resistance, R, with temperature, T, of a wire made from high purity platinum is shown in table 7.2.

Assuming the relationship between R and T can be written

$$R = A + BT + CT^2, \tag{7.24}$$

use least squares to determine best estimates for A, B and C.

7.5 Fitting equations with more than one independent variable[8]

To this point we have considered situations in which the dependent variable, y, is a function of only one independent variable, x. However, there are situations

[8] Such fitting is often referred to as 'multiple regression'.

Table 7.2. *Variation of resistance with temperature of a platinum wire.*

T (K)	R (Ω)
70	17.1
100	30.0
150	50.8
200	71.0
300	110.5
400	148.6
500	185.0
600	221.5
700	256.2
800	289.8
900	322.2
1000	353.4

in which the effect of two or more independent variables must be accounted for. For example, when vacuum depositing a thin film, the thickness, T, of the film may depend of several variables including:

(1) deposition time, t;
(2) gas pressure in vacuum chamber, P;
(3) distance, d, between deposition source and surface to be coated with the thin film.

To find the relationship between T and the independent variables, we could consider one independent variable at a time and experimentally determine the relationship between that variable and T while holding the other variables fixed. Another approach is to allow all the variables t, P and d to vary simultaneously from one experiment to another and use least squares to establish the relationship between T and all the independent variables.

An example of an equation possessing two independent variables is

$$y = a + bx + cz. \tag{7.25}$$

As usual, y is the dependent variable. Here x and z are independent variables. Can we determine a, b and c by fitting equation 7.25 to experimental data using linear least squares? The answer is yes, and the problem is no more complicated than fitting a polynomial to data.

We begin with the weighted sum of squares of residuals, χ^2, given by

$$\chi^2 = \sum \left[\frac{y_i - \hat{y}_i}{s_i}\right]^2. \tag{7.26}$$

To fit the equation $y = ax + bz + c$ to data, replace $\hat{y}_i$ in equation 7.26 by $a + bx_i + cz_i$ to give

$$\chi^2 = \sum \left[\frac{y_i - (a + bx_i + cz_i)}{s_i}\right]^2. \tag{7.27}$$

If the uncertainty in the y values is constant, s_i is replaced by s, so that

$$\chi^2 = \frac{1}{s^2}\sum (y_i - a - bx_i - cz_i)^2. \tag{7.28}$$

To minimise χ^2, differentiate χ^2 with respect to a, b and c in turn and set the resulting equations equal to zero. Following the approach described in section 7.2, the matrix equation to be solved for a, b and c, is

$$\underbrace{\begin{bmatrix} n & \sum x_i & \sum z_i \\ \sum x_i & \sum x_i^2 & \sum x_i z_i \\ \sum z_i & \sum x_i z_i & \sum z_i \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} a \\ b \\ c \end{bmatrix}}_{\mathbf{B}} = \underbrace{\begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum z_i y_i \end{bmatrix}}_{\mathbf{P}}. \tag{7.29}$$

To solve for a, b and c, we determine $\mathbf{B} = \mathbf{A}^{-1}\mathbf{P}$.

Example 4

In an experiment to study waves on a stretched string, the frequency, f, at which the string resonates is measured as a function of the tension, T, of the string and the length, L, of the string. The data gathered in the experiment are shown in table 7.3.

Assume that the frequency, f, can be written as

$$f = KT^B L^C, \tag{7.30}$$

where K, B and C are constants to be estimated using least squares.

Taking natural logarithms of both sides of equation 7.30 gives

$$\ln f = \ln K + B \ln T + C \ln L; \tag{7.31}$$

use least squares to determine best estimates of K, B and C.

ANSWER

Equation 7.31 can be written as $y = a + bx + cz$, where

$$y = \ln f, \tag{7.32}$$

$$x = \ln T \tag{7.33}$$

Table 7.3. *Variation of resonance frequency with string tension and string length.*

f (Hz)	T (N)	L (m)
28	0.49	1.72
48	0.98	1.43
66	1.47	1.24
91	1.96	1.06
117	2.45	0.93
150	2.94	0.82
198	3.43	0.67

and

$$z = \ln L. \tag{7.34}$$

Also, $a = \ln K$, $b = B$ and $c = C$.

Using the data in table 7.3, and the transformations given by equations 7.32 to 7.34, the matrices **A** and **P** in equation 7.29 become

$$\mathbf{A} = \begin{bmatrix} 7 & 3.532 & 0.5019 \\ 3.532 & 4.596 & -1.045 \\ 0.5019 & -1.045 & 0.6767 \end{bmatrix} \quad \mathbf{P} = \begin{bmatrix} 30.97 \\ 18.38 \\ 0.8981 \end{bmatrix}.$$

The inverse of matrix **A** is

$$\mathbf{A}^{-1} = \begin{bmatrix} 2.433 & -3.512 & -7.226 \\ -3.512 & 5.406 & 10.95 \\ -7.226 & 10.95 & 23.74 \end{bmatrix}.$$

Determining $\mathbf{B} = \mathbf{A}^{-1}\mathbf{P}$, gives

$$\mathbf{B} = \begin{bmatrix} 4.281 \\ 0.4476 \\ -1.157 \end{bmatrix}, \text{ so that}$$

$a = 4.281 = \ln K$, so $K = 72.31$,
$b = 0.4476 = B$, and
$c = -1.157 = C$,

so that $\ln f = 4.281 + 0.4476 \ln T - 1.157 \ln L$.

Finally, we write,

$$f = 72.31 T^{0.4476} L^{-1.157}.$$

Table 7.4. *The y, x and z data for exercise E.*

y	x	z
20.7	1	1
23.3	2	2
28.2	3	4
35.7	4	5
48.2	5	11
56.0	6	14

Exercise E

Consider the data shown in table 7.4.

Assuming that y is a function of x and z such that $y = a + bx + cz$, use linear least squares to solve for a, b and c.

7.6 Standard uncertainties in parameter estimates

We can extend the calculation of standard uncertainties in intercept and slope given in appendix 4 to include situations in which there are more than two parameters. In section 7.3 we found, $\mathbf{B} = \mathbf{A}^{-1}\mathbf{P}$, where the elements of $\mathbf{B}$ are the parameter estimates. The standard uncertainty in each parameter estimate is obtained by forming the product of the diagonal elements of $\mathbf{A^{-1}}$ and the standard deviation, s, in each value, y_i. The diagonal elements of $\mathbf{A^{-1}}$ are written as $A_{11}^{-1}, A_{22}^{-1}, A_{33}^{-1}$. The standard uncertainties of the elements, a, b and c of $\mathbf{B}$, written as s_a, s_b, s_c respectively, are given by[9]

$$s_a = s\left(A_{11}^{-1}\right)^{1/2} \tag{7.35}$$

$$s_b = s\left(A_{22}^{-1}\right)^{1/2} \tag{7.36}$$

$$s_c = s\left(A_{33}^{-1}\right)^{1/2}, \tag{7.37}$$

and s is determined by using

$$s = \left[\frac{1}{n-M}\sum (y_i - \hat{y}_i)^2\right]^{1/2}, \tag{7.38}$$

where n is the number of data and M is the number of parameters in the equation.

[9] See Bevington and Robinson (2002) for a derivation of equations 7.35 to 7.37.

Example 5

Consider the data in table 7.5.

Assuming it is appropriate to fit the function $y = a + bx + cx^2$ to the data in table 7.5, determine, using linear least squares:

(i) a, b and c;
(ii) s;
(iii) standard uncertainties, s_a, s_b and s_c.

ANSWER

(i) Writing equations to solve for a, b and c in matrix form gives (see section 7.3)

$$\underbrace{\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} a \\ b \\ c \end{bmatrix}}_{\mathbf{B}} = \underbrace{\begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}}_{\mathbf{P}}. \tag{7.39}$$

To find the elements of $\mathbf{B}$, write $\mathbf{B} = \mathbf{A}^{-1}\mathbf{P}$. Using MINVERSE() and MMULT() functions in Excel gives,

$$\underbrace{\begin{bmatrix} 1.206 & -0.2091 & 0.007576 \\ -0.2091 & 0.04423 & -0.001748 \\ 0.007576 & -0.001748 & 7.284 \times 10^{-5} \end{bmatrix}}_{\mathbf{A}^{-1}} \underbrace{\begin{bmatrix} 142.6 \\ 4210 \\ 94510 \end{bmatrix}}_{\mathbf{P}} = \underbrace{\begin{bmatrix} 7.847 \\ -8.862 \\ 0.6058 \end{bmatrix}}_{\mathbf{B}} \tag{7.40}$$

so that $a = 7.847$, $b = -8.862$ and $c = 0.6058$.

(ii) s is determined using equation 7.38 with, $n = 11$, $M = 3$ and $\hat{y}_i = a + bx_i + cx_i^2$ so that

Table 7.5. *The x–y data for example 5.*

x	y
2	–7.27
4	–17.44
6	–25.99
8	–23.02
10	–16.23
12	–15.29
14	1.85
16	21.26
18	46.95
20	71.87
22	105.97

$$s = \left[\frac{1}{8}\sum \left(y_i - \left(a + bx_i + cx_i^2\right)\right)^2\right]^{1/2} = 2.423.$$

(iii) s_a, s_b and s_c are found by using equations 7.35 to 7.37:

$$s_a = s\left(A_{11}^{-1}\right)^{1/2} = 2.423 \times (1.2061)^{1/2} = 2.7$$
$$s_b = s\left(A_{22}^{-1}\right)^{1/2} = 2.423 \times (0.04423)^{1/2} = 0.51$$
$$s_c = s\left(A_{33}^{-1}\right)^{1/2} = 2.423 \times \left(7.284 \times 10^{-5}\right)^{1/2} = 0.021.$$

Exercise F

Consider the thermocouple data in table 7.6.

Assuming the relationship between V and T can be written as

$$V = A + BT + CT^2 + DT^3, \tag{7.41}$$

use linear least squares to determine best estimates for A, B, C, D and the standard uncertainties in these estimates.

7.6.1 Coverage intervals for parameters

If a, b and c are estimates of the parameters α, β and γ respectively obtained using least squares, then the X% coverage interval for each parameter can be written as

Table 7.6. *Output voltage of a thermocouple as a function of temperature.*

T (K)	V (μV)
75	–5.936
100	–5.414
125	–4.865
150	–4.221
175	–3.492
200	–2.687
225	–1.817
250	–0.892
275	0.079
300	1.081

$$\alpha = a \pm t_{X\%,\nu} s_a$$
$$\beta = b \pm t_{X\%,\nu} s_b$$
$$\gamma = c \pm t_{X\%,\nu} s_c,$$

where s_a, s_b and s_c are the standard uncertainties in a, b and c respectively, and $t_{X\%,\nu}$ is the critical t value for ν degrees of freedom at the $X\%$ level of confidence.[10] We can find ν by using

$$\nu = n - M, \tag{7.42}$$

where n is the number of data and M is the number of parameters in the equation.

Example 6

In fitting the equation $y = \alpha + \beta x + \gamma x^2 + \delta x^3$ to 15 data points (not shown here), it is found that the best estimates of the parameters α, β, γ and δ (written as a, b, c and d respectively) and their standard uncertainties[11] are

$$a = 14.65,\ s_a = 1.727$$
$$b = 0.4651,\ s_b = 0.02534$$
$$c = 0.1354,\ s_c = 0.02263$$
$$d = 0.04502,\ s_d = 0.01018.$$

Use this information to determine the 95% coverage interval for each parameter.

ANSWER

For the parameter α we write, $\alpha = a \pm t_{X\%,\nu}\, s_a$.

The number of degrees of freedom, $\nu = n - M = 15 - 4 = 11$. Referring to table 2 in appendix 1, we have for a level of confidence of 95%,

$$t_{95\%,11} = 2.201.$$

It follows that

$\alpha = 14.65 \pm 2.201 \times 1.727 = 14.65 \pm 3.801$ (which we normally round to 14.7 ± 3.8).

Similarly for β, γ and δ,

$$\beta = 0.4651 \pm 2.201 \times 0.02534 = 0.465 \pm 0.056$$
$$\gamma = 0.1354 \pm 2.201 \times 0.02263 = 0.135 \pm 0.050$$
$$\delta = 0.04502 \pm 2.201 \times 0.01018 = 0.0450 \pm 0.022.$$

[10] $t_{X\%,\nu}$ is equivalent to the coverage factor, k, at the $X\%$ level of confidence.

[11] Standard uncertainties are given to four significant figures to avoid rounding errors in subsequent calculations.

Exercise G

A force, F, is required to displace the string of an archer's bow by an amount, d. Assume that the relationship between F and d can be written

$$F = \alpha + \beta d + \gamma d^2, \tag{7.43}$$

where α, β and γ are parameters to be estimated using linear least squares.

Estimates of the parameters were determined by fitting equation 7.43 to F versus d data (not shown here) which consisted of 12 data points. Estimates of α, β and γ (written as a, b and c respectively) and their standard uncertainties were found to be:

$a = 0.010\,47,\ s_a = 0.005993\ (\text{units N});$
$b = 3.931,\ s_b = 1.834\ (\text{units N/m});$
$c = 426.8,\ s_c = 13.03\ (\text{units N/m}^2).$

Use this information to determine the 90% coverage interval for α, β and γ .

7.7 Weighting the fit

If a weighted fit is required,[12] we return to the weighted sum of squares, χ^2, given by

$$\chi^2 = \sum \left[\frac{y_i - \hat{y}_i}{s_i}\right]^2. \tag{7.44}$$

To allow for uncertainty that varies from one y value to another, we retain s_i^2 in the subsequent analysis. Equation 7.44 is differentiated with respect to each parameter in turn and the resulting equations are set equal to zero. Next we solve the equations for estimates of the parameters which minimise χ^2. As an example, if the function to be fitted to the data is the polynomial, $y = a + bx + cx^2$, then a, b and c may be determined using matrices by writing $\mathbf{B} = \mathbf{A^{-1}P}$, where

$$\mathbf{B} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}. \tag{7.45}$$

$\mathbf{A^{-1}}$ is the inverse of, $\mathbf{A}$, where $\mathbf{A}$ is written

$$\mathbf{A} = \begin{bmatrix} \sum \frac{1}{s_i^2} & \sum \frac{x_i}{s_i^2} & \sum \frac{x_i^2}{s_i^2} \\ \sum \frac{x_i}{s_i^2} & \sum \frac{x_i^2}{s_i^2} & \sum \frac{x_i^3}{s_i^2} \\ \sum \frac{x_i^2}{s_i^2} & \sum \frac{x_i^3}{s_i^2} & \sum \frac{x_i^4}{s_i^2} \end{bmatrix} \tag{7.46}$$

[12] Weighted fitting is required if the uncertainty in each y value is not constant – see section 6.10.

and $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} \sum \frac{y_i}{s_i^2} \\ \sum \frac{x_i y_i}{s_i^2} \\ \sum \frac{x_i^2 y_i}{s_i^2} \end{bmatrix}. \tag{7.47}$$

The diagonal elements of the $\mathbf{A^{-1}}$ matrix can be used to determine the standard uncertainties in a, b and c. We have[13]

$$s_a = \left(A_{11}^{-1}\right)^{1/2} \tag{7.48}$$

$$s_b = \left(A_{22}^{-1}\right)^{1/2} \tag{7.49}$$

$$s_c = \left(A_{33}^{-1}\right)^{1/2}. \tag{7.50}$$

Exercise H

Consider the data in table 7.7.

Assuming that it is appropriate to fit the equation $y = a + bx + cx^2$ to these data and that the standard deviation, s_i in each y value is given by $s_i = 0.1y_i$:

(i) use weighted least squares to determine a, b and c;
(ii) determine the standard uncertainties in a, b and c.

Table 7.7. *The x–y data for exercise H.*

x	y
−10.5	143
−9.5	104
−8.3	79
−5.3	34
−2.1	12
1.1	19
2.3	29

[13] See Kutner, Nachtsheim and Neter (2003).

7.7.1 More on weighted fitting using matrices

When weighted fitting by least squares is undertaken, best estimates in parameters and standard uncertainties in those best estimates can be established by introducing an n by n 'weight' matrix, $\mathbf{W}$, given by[14]

$$\mathbf{W} = \begin{bmatrix} \frac{1}{s_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{s_2^2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{1}{s_n^2} \end{bmatrix}. \tag{7.51}$$

The matrix given by equation 7.46 can be determined using

$$\mathbf{A} = \mathbf{X^T W X}. \tag{7.52}$$

Similarly, the $\mathbf{P}$ matrix given by equation 7.47 can be determined using

$$\mathbf{P} = \mathbf{X^T W Y}. \tag{7.53}$$

The $\mathbf{B}$ matrix containing the best parameter estimates is found using

$$\mathbf{B} = \mathbf{A^{-1}P} = \left(\mathbf{X^T W X}\right)^{-1}\mathbf{X^T W Y}. \tag{7.54}$$

The matrix containing the standard uncertainties in the best estimates of the parameters (as diagonal elements of the matrix) can be written as $\mathbf{s^2(B)}$, where[15]

$$\mathbf{s^2}(\mathbf{B}) = \left(\mathbf{X^T W X}\right)^{-1}. \tag{7.55}$$

7.8 Coefficients of multiple correlation and multiple determination

In section 6.6 we introduced the linear correlation coefficient, r, which quantifies how well x and y are correlated. r close to 1 or –1 indicates excellent correlation, while r close to zero indicates poor correlation. We can generalise the correlation coefficient to take into account equations fitted to data which incorporate more than one independent variable or where there are more than two parameters in the equation. The coefficient of multiple correlation, R, is written

$$R = \left(1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}\right)^{1/2}, \tag{7.56}$$

where y_i is the ith observed value of y, $\hat{y}_i$ is the ith predicted y value found using the equation representing the best line through the points and $\bar{y}$ is the mean of

[14] n is the number of data points.

[15] $\mathbf{s^2(B)}$ is often referred to as the variance–covariance matrix.

the observed y values.[16] We argue that equation 7.56 is plausible by considering the summation $\sum (y_i - \hat{y}_i)^2$ which is the sum of squares of the residuals, *SSR*. If the line passes near to all the points then *SSR* is close to zero (and is equal to zero if all the points lie on the line) and so R tends to one, as required by perfectly correlated data.

The square of the coefficient of multiple correlation is termed the *coefficient of multiple determination*, R^2, and is written

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}. \tag{7.57}$$

R^2 gives the fraction of the square of the deviations of the observed y values about their mean which can be explained by the equation fitted to the data. So for example, if $R^2 = 0.92$ this indicates that 92% of the scatter of the deviations can be explained by the equation fitted to the data.

Exercise I

Consider the data in table 7.8.

Fitting the equation $y = a + bx + cz$ to the data in table 7.8 using least squares gives

$$\hat{y}_i = 2.253 + 1.978x_i - 0.3476z_i,$$

where $\hat{y}_i$ is the predicted value of y for a given x_i.

Use this information to determine the coefficient of multiple determination, R^2, for the data in table 7.8.

Table 7.8. *Data for exercise I.*

y	x	z
0.84	0.24	5.5
1.47	0.35	4.2
2.23	0.55	3.2
2.82	0.78	2.8
3.60	0.98	1.5
4.25	1.1	0.7

[16] See Walpole, Myers and Myers (1998) for a discussion of equation 7.56.

7.9 Estimating more than two parameters using the LINEST() function

The LINEST() function introduced in section 6.3 can be used to determine:

- estimates of parameters;
- standard uncertainties in the estimates;
- the coefficient of multiple determination, R^2;
- the standard deviation in the y values, s, as given by equation 7.38.

When fitting a straight line to data using LINEST(), a single column of numbers representing the x values is entered into the function whose syntax is as follows.[17]

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

By contrast, when there are two independent variables (such as x and z), an array with two columns of numbers must be entered into LINEST(). That array consists of one column of x values and an adjacent column of z values in the case where $y = a + bx + cz$ is fitted to data. Where the same independent variable appears in two terms of an equation, such as in $y = a + bx + cx^2$, one column in Excel contains x values and an adjacent column contains x^2.

The use of the LINEST() function to fit an equation containing more than two parameters is best illustrated by example. Consider the data in sheet 7.7.

Sheet 7.7. *Using the LINEST() function.*

	A	B	C
1	*y*	*x*	*z*
2	28.8	1.2	12.8
3	42.29	3.4	9.8
4	50.69	4.5	6.7
5	66.22	7.4	4.5
6	73.12	8.4	2.2
7	81.99	9.9	1
8			
9	=LINEST(A2:A7, B2:C7,TRUE,TRUE)		
10			
11			
12			
13			

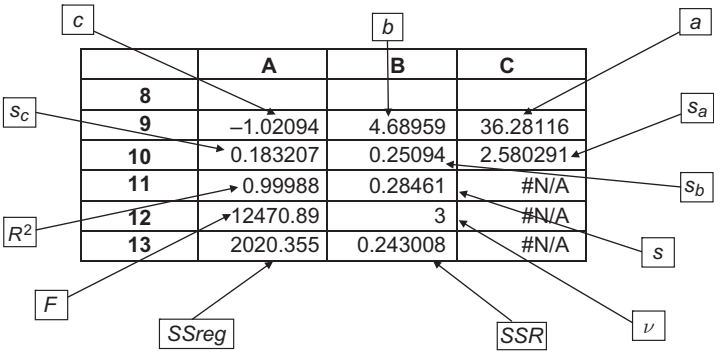
Assuming it is appropriate to fit the equation $y = a + bx + cz$ to the data in sheet 7.7, we proceed as follows.

[17] More details of the syntax of this function appear in section 6.3.

- Highlight a range of cells (cells A9 to C13 in sheet 7.7) into which Excel can return the estimates a, b and c, the standard uncertainties in these estimates, the coefficient of multiple determination and the standard deviation in the y values.
- Type =**LINEST(A2:A7,B2:C7,TRUE,TRUE)** into cell A9.
- Hold down the Ctrl and Shift keys together, then press the Enter key.

The numbers returned in cells A9 to C13 are shown in sheet 7.8.

Sheet 7.8. *Numbers returned by the LINEST() function.*



For completeness we state what each value returned by Excel represents when using the LINEST() function though we do not explain them in detail. The F statistic and its value to fitting using least squares is discussed in chapter 9.

Note that a, b and c are the best estimates of the parameters of the equation fitted to the data and s_a, s_b and s_c are the standard uncertainties in a, b and c respectively; s is the standard deviation in the y values; R^2 is the coefficient of multiple determination, ν is the number of degrees of freedom (which is equal to the number of data points, n, minus the number of parameters estimated using least squares) and SSR is the sum of squares of residuals. $SSReg$ is the regression sum of squares, given by

$$SSreg = \sum (\hat{y}_i - \bar{y}_i)^2. \tag{7.58}$$

F is the F-statistic given by[18]

$$F = \frac{SSreg/(n-\nu-1)}{SSR/\nu}. \tag{7.59}$$

[18] If fitting an equation to data forces the line through the origin (i.e. $a = 0$) then equation 7.59 should be replaced by $F = \frac{SSreg/(n-\nu)}{SSR/\nu}$.

Table 7.9. *Data for exercise J.*

x	y
0.02	−0.0261
0.04	−0.0202
0.06	−0.0169
0.08	−0.0143
0.10	−0.0124
0.12	−0.0105
0.14	−0.0094
0.16	−0.0080
0.18	−0.0067

A limitation of the LINEST() function is that it does not allow for weighted fitting. If weighted fitting of linear functions containing several parameters is required, then the matrix method described in section 7.7 is recommended.

Exercise J

Consider the data in table 7.9.

Use LINEST() to fit the equation $y = a + bx + c\ln x$ to the data in table 7.9 and so determine:

(i) a, b and c;
(ii) the standard uncertainties in a, b and c;
(iii) the coefficient of multiple determination, R^2;
(iv) the standard deviation, s, in the y values.

7.10 Choosing equations to fit to data

In some situations, selecting an equation to fit to data is not difficult. If consideration of the physical principles underlying the x–y data predicts that the relationship between y and x should be linear, and if a visual inspection of the data lends support to that prediction, then fitting the equation $y = a + bx$ is appropriate.

If there is little or no past experience to act as a guide as to which equation to fit to data we may need to rely on an examination of the data presented in graphical form to suggest one or more equations which can be trialled. However, it is worth emphasising that an equation relating the dependent to independent variables which derives from a consideration of physical principles is to be preferred to one that simply gives the smallest sum of squares of residuals.

Another situation in which the choice of equation to fit to data is 'clear cut' is when the purpose of data gathering is to calibrate an instrument. Previous experience, manufacturers guidelines or other published information are determining factors in the choice of the equation. When calibration is the aim of the least squares analysis, the physical interpretation of the parameter estimates that emerge from the fitting process is often not important, as the main aim is to use the fitted function for the purpose of interpolation (and occasionally extrapolation).[19]

Whatever equation is fitted to data, we should be suspicious if standard uncertainties of the parameter estimates are comparable in size (or larger than) the estimates themselves. For example, suppose some parameter estimate, b, is estimated by least squares to be $b = 3.6$ with a standard uncertainty, $s_b = 5.5$. It is likely that the term containing b can be eliminated with little adverse effect on the quality of the fit of the equation to data. In chapter 9 we will discover that it possible to make a judgement, based upon statistical considerations, as to whether a particular parameter estimate is 'significant' and we will leave discussion of this until we have considered 'tests of significance'.

7.10.1 Comparing equations fitted to data

Occasionally a situation arises in which, based on a consideration of the physical processes believed to be underlie the relationship being studied, two or more equations may reasonably be fitted to data. For example, the electrical resistance, r, of a material at a temperature, T, may be described by

$$r = A + BT \tag{7.60}$$

or

$$r = \alpha + \beta T + \gamma T^2, \tag{7.61}$$

where A, B, α, β, and γ are constants.

If we use χ^2 as given by equation 7.1, as a means of determining which of the two equations better fits the data, we will always favour equation 7.61 over equation 7.60. This is due to the fact that as the addition of the term in T^2 in equation 7.61 gives extra flexibility to the line so that predicted values found using the equation lie closer to the data values than if that term were omitted. However, the sum of squares of residuals obtained when fitting equation 7.61 to data may only be marginally less than when equation 7.60 is fitted to the same data.

[19] For details on fitting for calibration purposes, see Kirkup and Mulholland (2004).

Another possibility as an indicator of the equation that better fits the data is the coefficient of multiple determination, R^2, such that the equation yielding the larger value of R^2 is regarded as the better equation. Regrettably, as with the sum of squares of residuals, R^2 favours the equation with the greater number of parameters. For example, the equation $y = a + bx + cx^2 + dx^3$ would always be favoured over the equation $y = a + bx + cx^2$ when both equations are fitted to the same data. To take into account the number of parameters such that any marginal increase in R^2 is offset by the number of parameters used, the adjusted coefficient of multiple determination, R^2_{ADJ} is sometimes used, where[20]

$$R^2_{\text{ADJ}} = \frac{(n-1)R^2 - (M-1)}{n-M}. \tag{7.62}$$

R^2 is given by equation 7.57, n is the number of data and M is the number of parameters.

Once R^2_{ADJ} is calculated for each equation fitted to data, the equation is preferred that has the larger value of R^2_{ADJ}.

Another way of comparing two (or more) equations fitted to data where the equations have different numbers of parameters is to use Akaike's Information Criterion[21] (*AIC*). This criterion takes into account the sum of squares of residuals *SSR*, but also includes a term proportional to the number of parameters used. *AIC* may be written

$$AIC = n \ln\left(\frac{SSR}{n}\right) + 2K, \tag{7.63}$$

where n is the number of data and $K = (M + 1)$ in the equation.[22]

AIC depends on the sum of squares of residuals which appears in the first term of the right hand side of equation 7.63. The second term on the right hand side can be considered as a 'penalty' term. If the addition of another parameter in an equation reduces *SSR* then the first term on the right hand side of equation 7.63 becomes smaller. However, the second term on the right hand side increases by two for every extra parameter used. It follows that a modest decrease in *SSR* which occurs when an extra term is introduced into an equation may be more than offset by the increase in *AIC* by using another parameter. We conclude that, if two or more equations are fitted to data, then the equation producing the *smallest* value for *AIC* is preferred.

[20] See Kutner, M. J., Nachtsheim, C. J. and Neter, J. (2003) for a discussion of equation 7.62.

[21] See Akaike (1974) for a discussion of model identification.

[22] See Motulsky and Christopoulos (2005).

When the number of data is small, such that n/K is less than ≈ 40, a correction should be applied to equation 7.63. The corrected criterion, written as AIC_c, is given by

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}, \tag{7.64}$$

where AIC is given by equation 7.63.

If weighted least squares is required, we replace SSR in equation 7.63 by the weighted sum of squares of residuals, χ^2, where χ^2 is given by equation 7.1.

Example 7

Table 7.10 shows the variation of the resistance of an alloy with temperature.

Using (unweighted) linear least squares, fit equations 7.60 and 7.61 to the data in table 7.10 and determine for each equation:

(i) estimates for the parameters;
(ii) the standard uncertainty in each estimate;
(iii) the standard deviation, s, in each y value;
(iv) the coefficient of multiple determination, R^2;
(v) the adjusted coefficient of multiple determination, R^2_{ADJ};
(vi) the sum of squares of the residuals, SSR;
(vii) the Akaike's information criterion, AIC;
(viii) the corrected Akaike's information criterion, AIC_c.

ANSWER

The extent of computation required in this problem is strong inducement to use the LINEST() function in Excel. As usual, we write estimates of parameters as a, b and c etc. In the case of equation 7.60 we determine a and b where the equation is of the form, $y = a + bx$, and for equation 7.61, we determine a, b and c where $y = a + bx + cx^2$. The numbers shown in table 7.11 are rounded to four significant figures, though intermediate calculations utilise the full precision of Excel.

R^2_{ADJ} for equation 7.60 fitted to data is greater than R^2_{ADJ} for equation 7.61, indicating that equation 7.60 is the better fit. The AIC for equation 7.60 fitted to data is lower than for equation 7.61, further supporting equation 7.60 as the more appropriate equation. Finally, an inspection of the standard uncertainties of the parameter estimates suggests that, for the equation with three parameters, the standard uncertainty in c is so large in comparison to c that γ in equation 7.61 is likely to be 'redundant'.

Table 7.10. *Data for example 7.*

$r\,(\Omega)$	19.5	18.4	20.2	20.1	20.9	20.8	21.2	21.8	21.9	23.6	23.2
$T\,(\mathrm{K})$	150	160	170	180	190	200	210	220	230	240	250
$r\,(\Omega)$	23.9	23.2	24.1	24.2	26.3	25.5	26.1	26.3	27.1	28.0	
$T\,(\mathrm{K})$	260	270	280	290	300	310	320	330	340	350	

Table 7.11. *Parameter estimates found by fitting equations 7.60 and 7.61 to data in table 7.10*[23]

	fitting $y = a + bx$ to data	fitting $y = a + bx + cx^2$ to data
parameter estimates	$a = 12.41,\ b = 4.299 \times 10^{-2}$	$a = 13.97,\ b = 2.970 \times 10^{-2},\ c = 2.658 \times 10^{-5}$
standard uncertainties	$s_a = 0.49,\ s_b = 1.9 \times 10^{-3}$	$s_a = 2.166,\ s_b = 1.8 \times 10^{-2},\ s_c = 3.6 \times 10^{-5}$
s	0.5304	0.5368
R^2	0.9638	0.9649
R^2_{ADJ}	0.9619	0.9610
SSR	5.344	5.186
AIC	−22.74	−21.37
AIC_c	−21.33	−18.87

Exercise K

Using the data in table 7.10 verify that the parameter estimates and the standard uncertainties in the estimates appearing in table 7.11 are correct.

7.10.2 Akaike's weights

When, say, two equations are fitted to data, the equation with the lowest AIC is preferred. If the AICs of the two equations are comparable the extent to which the equation with the smaller AIC is preferred may be quite marginal. We can put it this way: in the situation in which two equations are fitted to the same data, what is the probability that the equation with the lower AIC is the better of the two?

Suppose the AIC value for the first equation fitted to data is AIC_1 and for the second equation is AIC_2. Also assume that $AIC_1 < AIC_2$ (i.e. the first equation is the preferred equation).

The difference between the AIC values for the two equations as $\Delta(AIC)$ is

[23] In the interests of conciseness, units of measurement have been omitted from the table.

$$\Delta(AIC) = AIC_1 - AIC_2. \tag{7.65}$$

The probability, p, that equation 1 is better equation that equation 2 is given by[24]

$$p = \frac{\exp[-1/2\Delta(AIC)]}{1 + \exp[-1/2\Delta(AIC)]}. \tag{7.66}$$

Example 8

Two equations are fitted to data. Fitting equation 1 gives an *AIC* value of –142.1 and fitting equation 2 gives an *AIC* value of –139.7. What is the probability that equation 1 is the better of the two?

ANSWER

Using equation 7.65, $\Delta(AIC) = -142.1 - (-139.7) = -2.40$.

Now, using equation 7.66, $p = \frac{\exp[-1/2 \times -2.40]}{1+\exp[-1/2 \times -2.40]} = 0.769$, or expressed as a percentage, 76.9%.

Exercise L

(i) Two equations are fitted to the same data. The first equation gives an *AIC* value of –92.1; the second equation gives an *AIC* of –88.2. What is the probability that the first equation is a better fit to the data than the second equation?

(ii) Using the relationship between p and $\Delta(AIC)$ given by equation 7.66, plot a graph of p versus $\Delta(AIC)$ for $-5 < \Delta(AIC) < 0$.

7.11 Review

Fitting equations to experimental data is a common occupation of workers in the physical sciences. Whether the aim of the comparison is to establish the values of parameters so that they can be compared with those predicted by theory, or to use the parameter estimates for calibration purposes, finding 'best' estimates of the parameters and the standard uncertainties in the estimates is very important.

In this chapter we considered how the technique of least squares can be extended beyond fitting the equation $y = a + bx$ to data. Increasing the number of parameters that must be estimated increases the number of the calculations that must be carried out in order to determine the best estimates of the parameters

[24] Note that this approach can be extended to the comparison of more than two equations, see Burnham and Anderson (2002) and Wagenmakers and Farrel (2004).

and their standard uncertainties. Applying matrix methods considerably eases the process of estimating parameters. The facility of being able to fit equations with more and more terms to produce a better fit must be applied cautiously as adding more terms reduces the sum of squares of residuals, but the reduction may only be marginal. We introduced the adjusted coefficient of multiple determination and the Akaike's Information Criterion as means of establishing which equation best fits when there are several competing equations.

In the next chapter we will consider fitting equations to data where those equations cannot be fitted using linear least squares. In particular, we will investigate Excel's Solver which is a useful utility that allows for the fitting of equations to data using non-linear least squares.

End of chapter problems

(1) The equation

$$y = a + bx + c \exp x \tag{7.67}$$

is to be fit to x–y data.

(i) Substitute equation 7.67 into equation 7.1. By differentiating χ^2 with respect to each parameter in turn, obtain three equations that may be solved for a, b and c. Assume that an unweighted fit using least squares is appropriate.

(ii) Write the equations obtained in part (i) in matrix form.

Consider the data in table 7.12.

(iii) Assuming that equation 7.67 is appropriate to the data in table 7.12, use matrices to find a, b and c.

Table 7.12. *The x–y data.*

x	y
1	8.37
2	3.45
3	1.70
4	0.92
5	0.53
6	0.62
7	–0.06
8	0.63
9	0.16
10	–0.27

Table 7.13. *Variation of plate height with flow rate of solute.*

v (mL/minute)	H (mm)
3.5	9.52
7.5	5.46
15.7	3.88
20.5	3.48
25.8	3.34
36.7	3.31
41.3	3.13
46.7	3.78
62.7	3.55
78.4	4.24
96.7	4.08
115.7	4.75
125.5	4.89

(2) The movement of a solute through a chromatography column can be described by the van Deemter equation,

$$H = A + \frac{B}{v} + Cv, \tag{7.68}$$

where H is the plate height, and v is the rate at which the mobile phase of the solute flows though the column. A, B, and C are constants. For a particular column, H varies with v as given in table 7.13.

(i) Write the equations in matrix form that must be solved for best estimates of A, B and C, assuming unweighted fitting using least squares is appropriate (hint: follow the steps given in section 7.2).
(ii) Solve for best estimates of A, B and C.
(iii) Determine the standard uncertainties for the estimates obtained in (ii).

(3) An object is thrown off a building and its vertical displacement above the ground at various times after it is released is shown in table 7.14.

The predicted relationship between s and t is

$$s = s_0 + ut + \tfrac{1}{2}gt^2, \tag{7.69}$$

where s_0 is the vertical displacement of the object at $t = 0$, u is its initial vertical velocity and g is the acceleration of the body falling freely under the action of gravity.

Table 7.14. *Vertical displacement of an object with time.*

t (s)	s (m)
0.0	135.2
0.5	141.7
1.0	142.3
1.5	148.2
2.0	143.7
2.5	140.3
3.0	134.1
3.5	126.7
4.0	114.5
4.5	100.8
5.0	85.7
5.5	66.7
6.0	46.5

Use least squares to determine best estimates for s_0, u and g and the standard uncertainties in the estimates.

(4) To illustrate the way in which a real gas deviates from perfect gas behaviour, $\frac{PV}{RT}$ is often plotted against $\frac{1}{V}$, where P is the pressure, V the volume and T the temperature (in kelvins) of the gas. R is the gas constant. Values for $\frac{PV}{RT}$ and V are shown in table 7.15 for argon gas at $T = 150$ K.

Assuming the relationship between $\frac{PV}{RT}$ and $\frac{1}{V}$ can be written as

$$\frac{PV}{RT} = A + \frac{B}{V} + \frac{C}{V^2} + \frac{D}{V^3}, \tag{7.70}$$

use least squares to obtain best estimates for A, B, C and D and standard uncertainties in the estimates.

(5) Consider the x–y data in table 7.16.

Fit equations $y = a + bx$ and $y = a + bx + cx^2$ to these data.

(i) Use the adjusted multiple correlation coefficient and corrected Akaike's Information Criterion to establish which equation better fits the data.

(ii) Calculate the probability that the equation giving the lowest value of AIC is the better equation to fit to the data in table 7.16.

(6) Table 7.17 shows data of the variation of the molar heat capacity (at constant pressure), C_p, of oxygen with temperature, T.

Table 7.15. *Variation of* $\frac{PV}{RT}$ *with V for argon gas.*

$V\,(\mathrm{cm}^3)$	$\frac{PV}{RT}$
35	1.21
40	0.89
45	0.72
50	0.62
60	0.48
70	0.45
80	0.51
90	0.50
100	0.53
120	0.57
150	0.61
200	0.69
300	0.76
500	0.84
700	0.89

Table 7.16. *x–y data.*

x	y
5	361.5
10	182.8
15	768.6
20	822.5
25	1168.2
30	1368.6
35	1723.3
40	1688.7
45	1800.9
50	2124.5
55	2437.9
60	2641.2

Table 7.17. *Measured molar heat capacity of oxygen in the temperature range 300 K to 1000 K.*

T(K)	C_p(J ·mol^{-1} ·K^{-1})
300	29.43
350	30.04
400	30.55
450	30.86
500	31.52
550	31.71
600	32.10
650	32.45
700	32.45
750	32.80
800	33.11
850	33.38
900	33.49
950	33.85
1000	34.00

Assuming that the relationship between C_p and T can be written as

$$C_p = A + BT + \frac{C}{T^2}, \tag{7.71}$$

determine:

(i) best estimates for A, B and C;
(ii) the standard uncertainties in the estimates;
(iii) the 95% coverage intervals for A, B and C.

(7) As part of a study into the behaviour of electrical contacts made to a ceramic conductor, the data in table 7.18 were obtained for the temperature variation of the electrical resistance of the contacts.

It is suggested that there are two possible models that can be used to describe the variation of the contact resistance with temperature.

Model 1

The first model assumes that contacts show semiconducting behaviour, where the relationship between R and T can be written

Table 7.18. *Resistance versus temperature for electrical contacts on a ceramic.*

T (K)	R (Ω)	T (K)	R (Ω)
50	4.41	190	0.69
60	3.14	200	0.85
70	2.33	210	0.94
80	2.08	220	0.78
90	1.79	230	0.74
100	1.45	240	0.77
110	1.36	250	0.68
120	1.20	260	0.66
130	0.86	270	0.84
140	1.12	280	0.77
150	1.05	290	0.75
160	1.05	300	0.86
170	0.74		
180	0.88		

$$R = A \exp\left(\frac{B}{T}\right), \tag{7.72}$$

where A and B are constants.

Model 2

Another equation proposed to describe the data is

$$R = \alpha + \beta T + \gamma T^2, \tag{7.73}$$

where α, β and γ are constants.

Using plots of residuals, the adjusted coefficient of multiple determination, corrected Akaike's Information Criterion and any other indicators of 'goodness of fit', determine whether equation 7.72 or equation 7.73 better fits the data.

Assistance: To allow a straight line based on equation 7.72 to be fitted to data, the data need to be transformed. Transform back to the original units after fitting a line to data before calculating SSR, R^2_{ADJ}, and AIC_{c}. This will allow for direct comparison between SSR, R^2_{ADJ}, and AIC_{c} between equations 7.72 and 7.73.

Table 7.19. *Flow rate dependence on Δp, d, and l.*

Q (ml/s)	Δp (Pa)	d (mm)	l (mm)
0.0092	5488	0.265	43
0.080	5488	0.395	43
0.22	4753	0.82	105
0.37	4753	0.82	80
0.53	4753	0.82	56
0.60	4753	0.82	43
0.67	5488	0.82	43
0.76	2842	1	43
0.90	3626	1	43
1.00	4312	1	43
1.02	4704	1	43
1.14	5488	1	43

(8) The flow of fluid, Q, through a hollow needle (of circular cross-section) depends on:

- the pressure difference, Δp, between the ends of the needle;
- the internal diameter of the needle, d;
- the length of the tube, l.

Table 7.19 shows data obtained for Q for various combinations of d, l and Δp for a range of hollow steel needles.

Assume that the relationship between Q, d, l and Δp can be written as

$$Q = k\Delta p^r d^s l^t, \tag{7.74}$$

where k, r, s and t are constants.

Use least squares to find best estimates for k, r, s and t and standard uncertainties in the best estimates.