The Open University

M248

Analysing data

Book C

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from Student Recruitment, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)300 303 5303; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2017.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed in the United Kingdom by Hobbs the Printers Limited, Brunel Road, Totton, Hampshire SO40 3WX.

ISBN 978 1 4730 2035 1

2.1

Contents

Unit 11

Regression

Introduction

So far in this module, models for variation have been developed that are appropriate for studying a suitably defined underlying population in its entirety. For instance, in Unit 6, you met an example of data collected on the heights of elderly women. (These data formed part of a study into the disease osteoporosis.) There was evident variation in the sample, and it was suggested that the variation in the data could be modelled adequately by a normal distribution with appropriately chosen values for its parameters μ and σ. This model provided useful information about the population of heights of all elderly women. However, it did not provide information about the population of heights of females in general – for example, the variation in heights of teenage girls is likely to be different from that for elderly women. In the wider population of women in general, we would expect height to depend on age at least up to about 15 or 16 years old; manufacturers of children's clothes, for instance, need to be aware of this relationship. The following example illustrates how such a relationship can be modelled.

Example 1 *Heights of schoolboys*

A very early study conducted for the Massachusetts Board of Health in 1877 recorded the age and height of each of 24 500 Boston schoolboys between the ages of 6 and 10 years. A histogram of the heights of the boys (in inches) is shown in Figure 1.

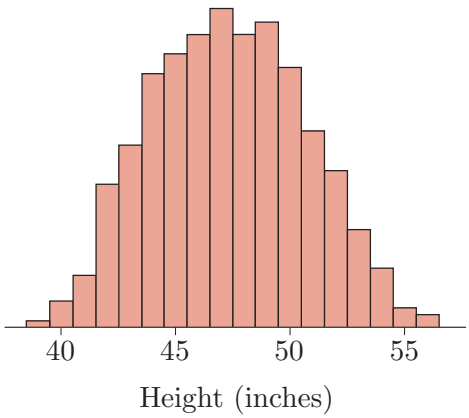


Figure 1 The variation in height for boys between the ages of 6 and 10 years

Nineteenth-century schoolboys (and girls)

Figures 1 and 2 are adapted from Peters, W.S. (1987) *Counting for Something – Statistical Principles and Personalities*, New York, Springer-Verlag, p. 90.

As for the heights of elderly women, we could look for a model for the variation in the heights of these schoolboys; a normal distribution again seems a reasonable possibility, but with different values for its parameters μ and σ. This model would provide some general information about the heights of nineteenth-century Boston schoolboys between the ages of 6 and 10 years, but it would not tell us anything about the relationship between height and age for these boys. Instead, if the boys are divided into five age

groups of a year each (ages 6 to 10 years) and a histogram is drawn separately for each group, then the same data may be represented as in Figure 2. In the figure, height is represented by the vertical axis while age, grouped into 6, 7, 8, 9 or 10 years, is represented by the horizontal axis. Each histogram is plotted on its side rather than in the usual way.

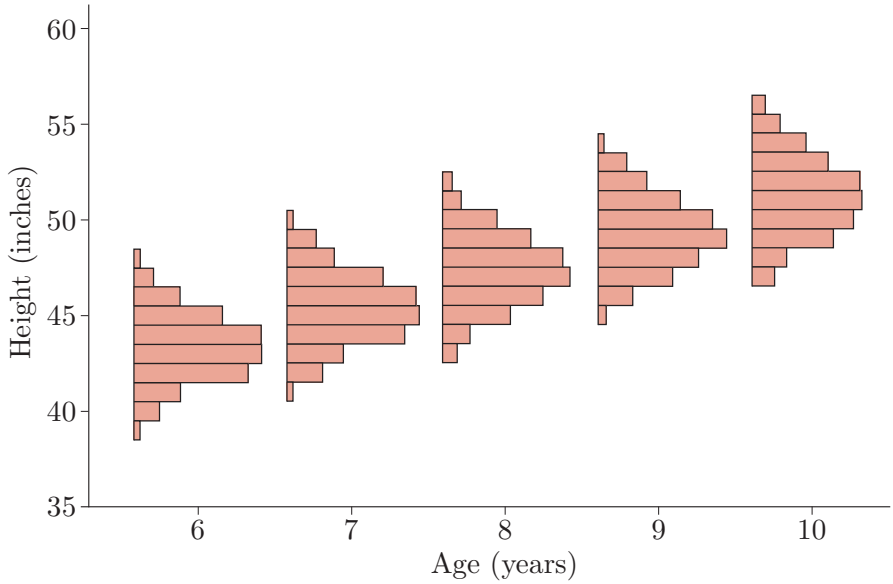


Figure 2 The variation in height for boys in each of the age groups 6 to 10 years

You can now see that, *for each age group*, a normal distribution might provide a good model for the variation in heights. It also appears that the mean height increases roughly linearly with age – at least between the ages of 6 and 10 years. (The increasing linear trend would not continue into all higher ages, of course.) The spread of heights about the mean does not seem to alter much with age, that is, the variance in heights appears to be approximately constant. So perhaps the variation in heights of nineteenth-century Boston schoolboys can be adequately modelled by a collection of normal distributions where the normal distributions differ with respect to their means, μ, but all have the same variance, σ^2. Moreover, rather than being an arbitrary collection of means, it seems that the means of the normal distributions increase linearly with age.

Relationships of this sort between variables are the subject of this unit. As you may know already, statistical models that reflect the way in which variation in an observed variable changes with one or more other variables are called **regression models**. The development and use of these models is known as **regression**. Situations where a regression model might be useful include the following.

- Economists predict future employment rates on the basis of past and current rates, together with various economic variables.
- Farmers wish to know how the yield of crops depends on the amount of fertiliser used.

- Doctors must decide, on the basis of particular measurements, how much of a drug to give to a patient.
- A car owner might be interested in knowing how driving her car at different speeds alters its fuel consumption.

In Section 1, a few more examples are given before the *general regression model* is formally defined. Also, a particularly important regression model, the *linear regression model*, is introduced. In Section 2, a method for fitting linear regression models to data is described. Section 3 is principally concerned with checking the modelling assumptions; at its end, you will see how to fit the linear regression model and how to check the assumptions using Minitab. Statistical inference, such as testing hypotheses and calculating confidence intervals, for linear regression models is discussed quite briefly in Section 4. The unit ends by introducing *multiple regression*, where the relationship between a variable and more than one other variable is of interest.

1 Regression models

1.1 Examples

This section begins with a few more examples of contexts in which regression data arise.

Cruise control

Example 2 *Driving at constant speed*

Consider the following hypothetical situation. For a car driving at a constant speed of 50 mph, the relationship between the distance travelled and the time spent driving can be represented by the straight line in Figure 3(a).

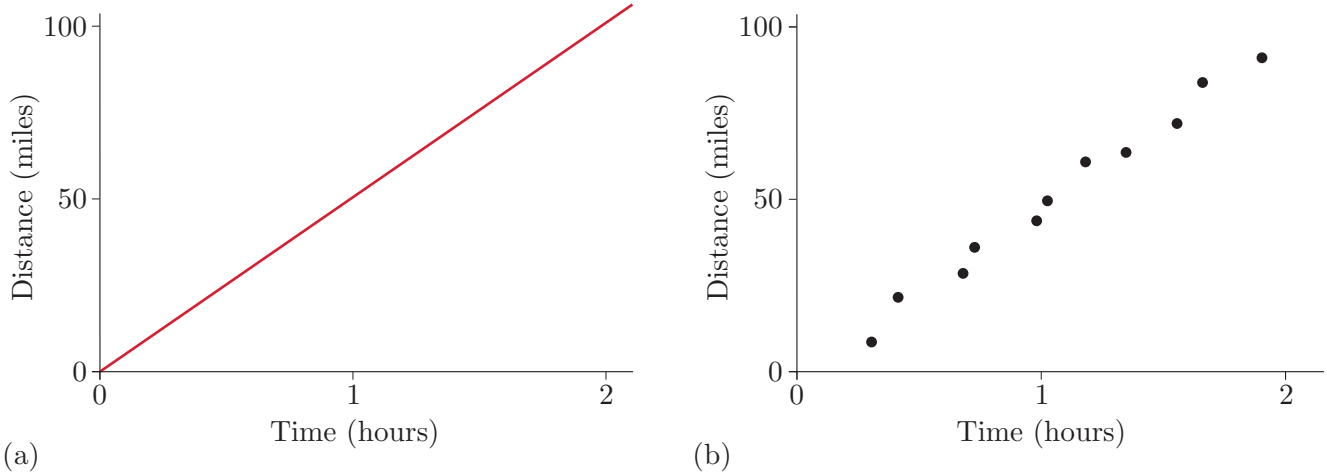


Figure 3 (a) Distance against time. (b) A scatterplot of 'real' observations.

A scatterplot of observations measured without error would consist of dots all lying exactly on the straight line in Figure 3(a). However, in a scatterplot of real observations, the dots are very unlikely to lie exactly along the straight line but would be scattered around the line, perhaps looking something like Figure 3(b). So we need a model that will describe the linear relationship underlying these data while at the same time allowing for some deviation of the data from the line.

Example 3 *Forbes's data on the boiling point of water*

Timby's Mercury Barometer, patented 1857

In the 1840s and 1850s the Scottish physicist James Forbes was interested in developing a method for estimating altitude on a hillside from measurement of the boiling point of water there. The temperature at which water boils is affected by atmospheric pressure which, in turn, is affected by altitude. (You might know that the higher the altitude, the lower the pressure, and the lower the boiling point of water.)

So boiling point depends on atmospheric pressure, and if the details of that relationship were known, Forbes concluded that it should be possible to turn the relationship round so that climbers could estimate their height from the temperature at which water boiled. Carrying barometers – which, at that time, were large instruments which included a long, thin glass tube containing mercury – up and down hills intact was a tricky business; boiling a pan of water and measuring the temperature of the boiling point was less troublesome. Here, however, we will concentrate on the initial question of the way boiling point depends on atmospheric pressure.

The data in Table 1 give the boiling point (in °F) and atmospheric pressure (in inches Hg – that is, inches of mercury) for 17 locations in the Alps and in Scotland.

Table 1 Forbes's data

Boiling point (°F)	194.5	194.3	197.9	198.4	199.4	199.9
Pressure (inches Hg)	20.79	20.79	22.40	22.67	23.15	23.35
Boiling point (°F)	200.9	201.1	201.4	201.3	203.6	204.6
Pressure (inches Hg)	23.89	23.99	24.02	24.01	25.14	26.57
Boiling point (°F)	209.5	208.6	210.7	211.9	212.2	
Pressure (inches Hg)	28.49	27.76	29.04	29.88	30.06	

(Source: Forbes, J.D. (1857) 'Further experiments and remarks on the measurement of heights by the boiling point of water', *Transactions of the Royal Society of Edinburgh*, vol. 21, no. 2, pp. 235–43)

The scatterplot of these data in Figure 4 suggests that there may well be a straight-line relationship between the boiling point of water and atmospheric pressure. A model for the data should exhibit this linear relationship.

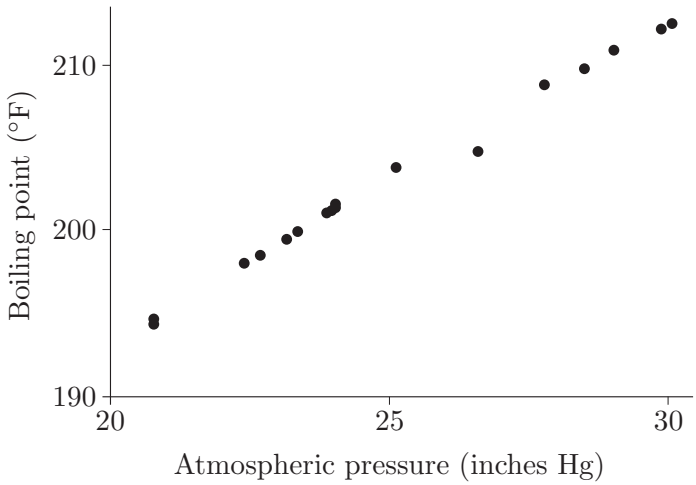


Figure 4 Boiling point of water against atmospheric pressure

Example 4 *The strength of timber beams*

A dataset contains the results of an investigation into how specific gravity (a measure of density) and moisture content might influence the strength of timber beams. Table 2 contains measurements of the three variables for each of ten beams. Unfortunately, units of measurement are not given in the source.

Table 2 Strength of beams

Strength	11.14	12.74	13.13	11.51	12.38
Specific gravity	0.499	0.558	0.604	0.441	0.550
Moisture content	11.1	8.9	8.8	8.9	8.8
Strength	12.60	11.13	11.70	11.02	11.42
Specific gravity	0.528	0.418	0.480	0.406	0.467
Moisture content	9.9	10.7	10.5	10.5	10.7

(Source: Draper, N.R. and Stoneman, D.M. (1966) 'Testing for the inclusion of variables in linear regression by a randomisation technique', *Technometrics*, vol. 8, no. 4, pp. 695–9)

The scatterplot of strength against specific gravity in Figure 5(a) (overleaf) suggests some sort of increasing linear relationship between strength and specific gravity, though possibly there is an outlier at $(0.499, 11.14)$.

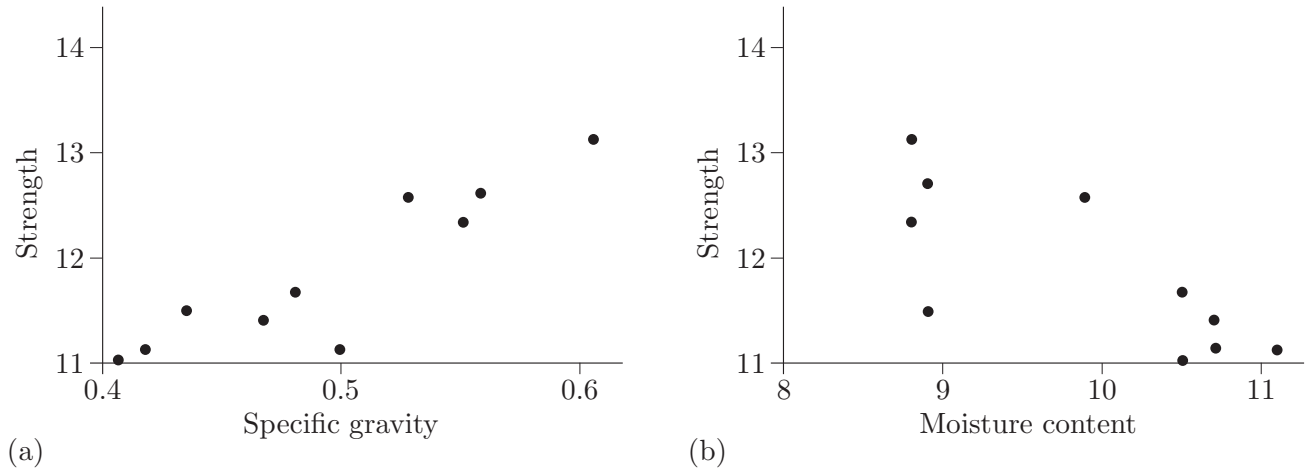


Figure 5 Scatterplots: (a) strength against specific gravity (b) strength against moisture content

Although the scatterplot of strength against moisture content in Figure 5(b) suggests an overall downward trend, it is not all that convincing – it does not seem to strongly suggest any particular form for a relationship. Linearity might, however, be as good as any.

Example 5 *The growth of duckweed*

Ducks in duckweed

In his 1917 book *On Growth and Form*, the Scottish mathematical biologist Sir D'Arcy Wentworth Thompson recounts an experiment into the growth of duckweed, a plant that grows on water. Growth was monitored by counting duckweed fronds at weekly intervals for eight weeks, starting one week after the introduction of a single duckweed plantlet into a growth medium (in this case, pure water). Initially (week 0) there were 20 fronds on the plantlet. The data are given in Table 3.

Table 3 Duckweed growth

Week	1	2	3	4	5	6	7	8
Fronds	30	52	77	135	211	326	550	1052

(Source: Thompson refers to work summarised in Bottomley, W.B. (1914) 'Some accessory factors in plant growth and nutrition', *Proceedings of the Royal Society, Series B*, vol. 88, no. 602, 237–47)

A scatterplot of the data is given in Figure 6.

You can see that there is a very strong suggestion of a relationship between duckweed growth and passing time; but, unlike in the previous examples, the relationship is not linear. Instead, it might be possible to fit a curve to the data – perhaps some sort of power or exponential function.

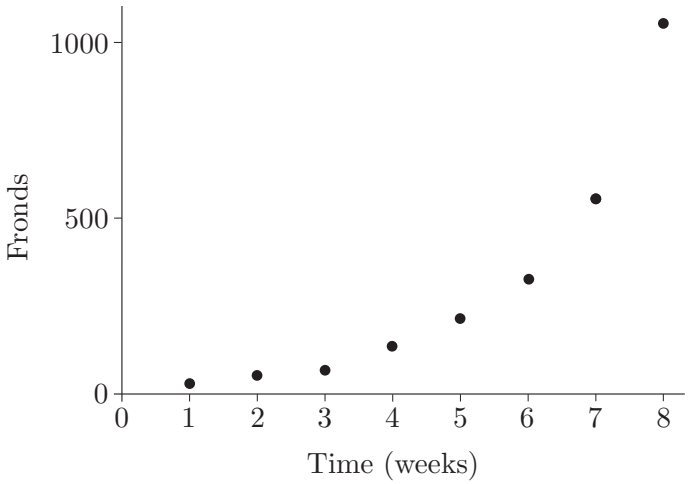


Figure 6 Duckweed growth

Often, data arise as the result of an experiment specifically designed to investigate the effect that changes in one variable (or more than one variable) have on another variable: Forbes investigated the effect of changes in atmospheric pressure on the boiling point of water (see Example 3); the strength of timber was measured for different values of specific gravity and moisture content (see Example 4); the scatterplot in Figure 6 suggests how the growth of duckweed depends on passing time. In all three examples, we could naturally think of one variable (or two variables, in Example 4) having an effect on, or '*explaining*', another variable. And in each of these three examples, it would not be at all natural or even sensible to swap the variables around: we would not speak of an increase or decrease in the boiling point of water 'changing' the atmospheric pressure; or of the strength of a timber beam 'having an effect on' the moisture content; or of the growth of duckweed 'causing' a change in time.

The word '*explain*' should not be taken too literally in this context; it is used only to express that a change in one variable has an effect on another variable.

A variable that 'explains' another variable is called an **explanatory variable**. In Example 3, Forbes used atmospheric pressure to 'explain' the boiling points of water at various altitudes. So atmospheric pressure is an explanatory variable. In Example 4, there are two explanatory variables, namely specific gravity and moisture content, both 'explaining' the strength of the timber beams. In Example 5, time 'explains' the changing number of duckweed fronds, so time is an explanatory variable.

In Example 3, the measured boiling point can be regarded as a *response* to a given atmospheric pressure. For different pressures, the boiling point will be different. The variable that 'responds' to the value of the explanatory variable is called the **response variable**. In Example 4, the response variable is the strength of the timber beam; and in Example 5, the response variable is the number of duckweed fronds.

Other names are sometimes used for the response and explanatory variables. These include *dependent variable* and *independent variable*, respectively. The explanatory variable is also called the *predictor variable*, the *regressor* or the *covariate*.

It can be argued that the name 'independent variable' is misleading because we are interested in relationships between variables which are not independent.

As you saw in Example 4, there can be more than one explanatory variable. However, first we will be concerned with the case where there is only one explanatory variable. In this case, the model is often referred to as a *simple* regression model. The word 'simple' here refers purely to the number of variables involved in the regression; you can be the judge of whether or not the interpretation, properties and application of the simple regression model are what you would call simple! When there are two or more explanatory variables, the model is called a *multiple* regression model. This will be the topic of Section 5.

Activity 1 *Heights of Boston schoolboys*

In the Introduction, data on the age and height of schoolboys from Boston were discussed. Which of the two variables, age and height, would you regard as the response variable and which as the explanatory variable?

Activity 2 *Paper strength*

Table 4 contains data on the strength of kraft paper. ('Kraft' refers to a method of paper production. The paper is of a thick brown type used for wrapping.) The tensile strength (in pounds per square inch (p.s.i.)) of the paper was measured along with the percentage of hardwood in the batch of pulp from which the paper was produced. In Figure 7, tensile strength is plotted against hardwood content.

Table 4 Paper strength

Strength (p.s.i.)	Hardwood content (%)
6.3	1.0
11.1	1.5
20.0	2.0
24.0	3.0
26.1	4.0
30.0	4.5
33.8	5.0
34.0	5.5
38.1	6.0
39.9	6.5
42.0	7.0
46.1	8.0
53.1	9.0
52.0	10.0
52.5	11.0
48.0	12.0
42.8	13.0
27.8	14.0
21.9	15.0

(Source: Joglekar, G., Schuenemeyer, J.H. and LaRiccia, V. (1989) 'Lack-of-fit testing when replicates are not available', *American Statistician*, vol. 43, no. 3, pp. 135–43)

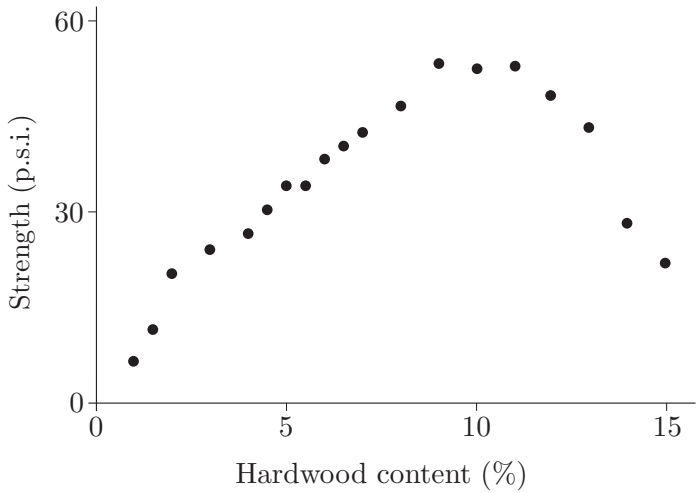


Figure 7 Tensile strength of kraft paper against hardwood content

(a) Which of the two variables, hardwood content and tensile strength, is the response variable and which is the explanatory variable?

(b) What can you say from the scatterplot about the nature of the relationship between the variables?

Notice that in all these examples and activities, the explanatory variable has been plotted along the x-axis in the scatterplot, and the response variable along the y-axis. This is standard practice.

1.2 The general regression model

In regression, it is customary to regard the explanatory variable as non-random and the response variable as a random variable. That is, the values of the explanatory variable are considered 'exact' and hence all the scatter observed in a scatterplot is ascribed to variability in the response. This set-up is directly applicable to the sort of *designed experiments* in which the experimenter is able to choose specific values for the explanatory variable and is interested in the values of the response variable which result. A particular example of this is the duckweed experiment of Example 5; there, the experimenter decided to count the numbers of duckweed fronds (the response variable) at a selected number of values of the explanatory variable, namely after one week, after two weeks, and after each week up to eight weeks. In other regression situations, the values of the explanatory variable might have arisen via some chance mechanism. However, for modelling purposes, interest remains centred on how values of the response variable arise, given those values of the explanatory variables (and not on how the explanatory variables themselves are distributed).

Since the explanatory variable is regarded as non-random, it is always denoted by a lower-case letter, usually x. The response variable is denoted by an upper-case letter, usually Y, to indicate that it is a random variable, whenever it is appropriate to do so. So the points in a general sample of size n are then denoted (x_1, Y_1), (x_2, Y_2), $\ldots$, (x_n, Y_n). That said, the *observed* values of such a sample are usually denoted using lower-case y_is: (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n).

In Subsection 1.1, you saw examples where the relationship between two variables appears to be linear and other examples where the relationship might be better modelled by a curve. In each case, there was some scatter about the line or curve – a little in some cases, but a lot in others. The general regression model is made up of two parts:

- (the 'systematic' or deterministic part) a function $h(x)$ that defines the line or curve about which the points in a scatterplot are scattered; $h(x)$ is called the **regression function**
- (the random part) a term which models the scatter, that is, the variation in the response variable about the regression function. This term is itself a random variable, W say. An important property of W is that $E(W) = 0$, that is, that the random part of the model, W, has zero mean.

The general regression model is defined formally as follows.

The general regression model

If the response variable is denoted by Y and the explanatory variable by x, then the **general regression model** for the collection of points (x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n) can be written

$$Y_i = h(x_i) + W_i, \quad i = 1, 2, \ldots, n.$$

Here h represents some function and the W_is are independent random variables with zero mean.

The function h may be linear, but it can also represent a curve – perhaps polynomial or logarithmic or exponential or trigonometric.

Note that $h(x_i)$ is not random (since x_i is not random) but is an additive constant, so the assumptions for the W_is are equivalent to the response variables Y_i being independent with mean $h(x_i)$. To see the latter, note that

$$\begin{aligned} E(Y_i) &= E\{h(x_i) + W_i\} = E\{h(x_i)\} + E(W_i) = h(x_i) + E(W_i) \\ &= h(x_i) + 0 = h(x_i). \end{aligned}$$

A schematic example of the general regression model is given in Figure 8. There, $h(x) = x^2$ is the regression function which represents the main trend in the model. For each of a number of values of x, the distribution of $Y = h(x) + W = x^2 + W$ is shown; in particular, notice how the distribution is 'centred on' the value $h(x) = x^2$.

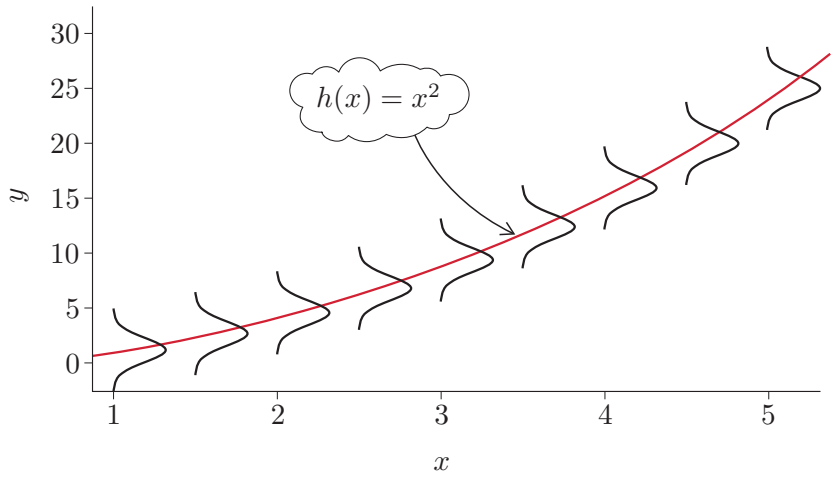


Figure 8 The regression function $h(x) = x^2$ and the distribution of $Y = x^2 + W$ at $x = 1, 1.5, 2, 2.5, \ldots, 5$

Example 6 *A model for Forbes's data*

For Forbes's data in Table 1, the response variable Y is the boiling temperature of water and the explanatory variable x is the atmospheric pressure. From Figure 4, you can see that there appears to be a straight-line relationship between boiling temperature and atmospheric pressure. So a suitable model might be

$$Y_i = \alpha + \beta x_i + W_i.$$

Here α and β are the intercept and slope, respectively, of the straight line relating the boiling temperature to the atmospheric pressure. The random terms W_i account for the scatter around the straight line.

Activity 3 *A model for the heights of Boston schoolboys*

In the Introduction, you saw that the heights of nineteenth-century Boston schoolboys of different ages seemed to be adequately modelled by normal distributions with means linearly related to age and with roughly equal variances. What might be the form of an appropriate regression model for these data? Can you say anything more about the distribution of the random terms in this model?

A little caution is needed here. Sometimes a list of data pairs may appear to suggest a linear relationship between the variables, but when further measurements are taken outside the range investigated, it becomes clear that a more complex model is required. We have already alluded to this in the case of height measurements in the Introduction. There (and in Activity 3 above) a linear relationship was suggested for the mean height of boys between the ages of 6 and 10 years; however, it was noted that such an increasing linear trend would not provide a suitable model for males of older ages. The case of atmospheric pressure and altitude provides another example of this. The scatterplots in Figure 9 show atmospheric pressure (as a percentage of pressure at sea level) plotted against altitude (in metres, at various points on the Earth's surface).

Figure 9 is taken from The Open University (1992) MS284 *An Introduction to Calculus*, Unit 7, *Numbers from Nature*, Milton Keynes, The Open University.

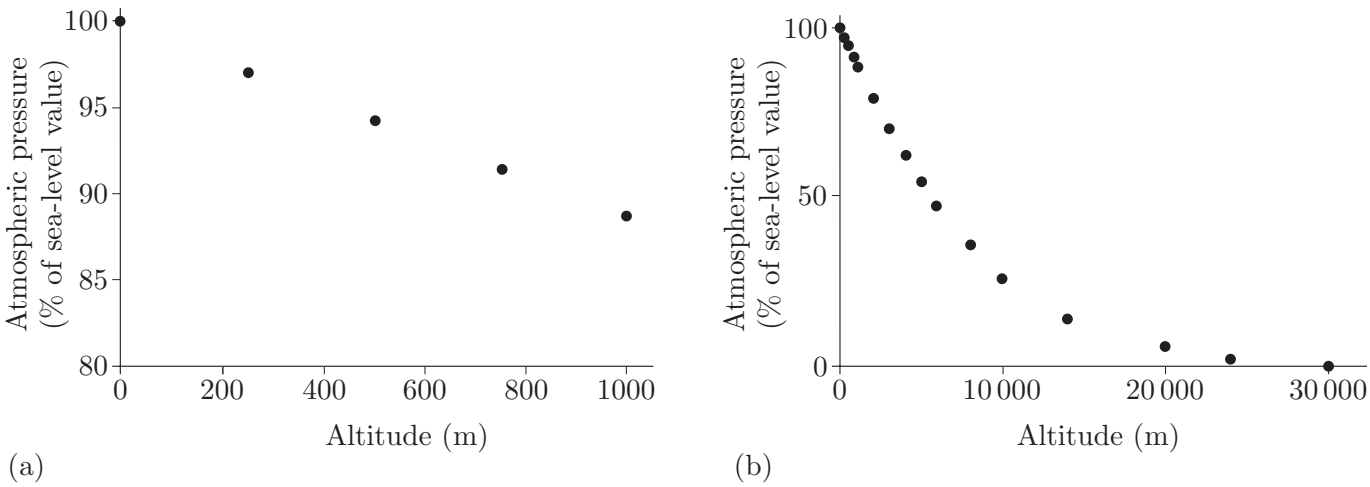


Figure 9 Pressure at different altitudes: (a) up to 1000 metres; (b) up to 30 000 metres

You can see from both panels of Figure 9 that pressure decreases with increasing altitude. Over the range of altitudes considered in Figure 9(a), which was from sea level up to 1000 metres, the relationship appears to be linear. If, however, you were to explore what happens when further

measurements are taken outside this range, you would find that the relationship is no longer linear. This is clear from the scatterplot in Figure 9(b), which shows measured values of atmospheric pressure at altitudes up to 30 000 metres. For this wider range, a more sophisticated mathematical model than a simple straight-line regression model is needed to describe the relationship between the variables.

Example 7 *A model for the duckweed data*

It is clear from Figure 6 that there is a relationship between duckweed growth and passing time, but this relationship is not linear. A possible regression model might be a formula expressing exponential growth, say,

$$Y_i = 20e^{\lambda x_i} + W_i$$

for some parameter value λ. The regression function $h(x) = 20e^{\lambda x}$ is shown for the case $\lambda = 0.5$ in Figure 10. (The value 20 occurs because there were 20 fronds at time 0.) The random term W_i accounts for the scatter. Notice that the regression model of exponential growth cannot persist for all values of the explanatory variable time, else we would now all be covered in duckweed!

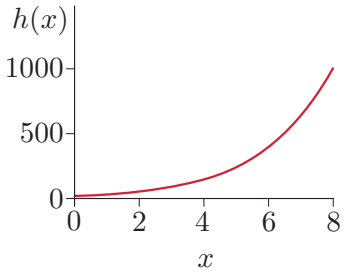


Figure 10 The function $h(x) = 20e^{0.5x}$

1.3 The linear regression model

The most important case of the general regression model is the **linear regression model**. A linear regression model is a regression model where the relationship between Y and x is linear.

The importance of the linear regression model is that not only is it very common (you have already met it in Example 6 and Activity 3), but it is also, as you will see in Section 4, relatively simple to use for statistical inference, such as testing hypotheses and obtaining confidence intervals. In addition, as you will see in Unit 12, particular apparently non-linear regression models can be reduced to linear regression models. A formal definition of the linear regression model is given in the box below.

The linear regression model

If Y is the response variable and x is the explanatory variable, then the **linear regression model** for the collection of points (x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n) can be written

$$Y_i = \alpha + \beta x_i + W_i, \quad i = 1, 2, \ldots, n.$$

The parameters α and β are the intercept and the slope, respectively, of the straight line relating Y to x. The terms W_i are independent random variables with zero mean and constant variance.

In other texts, the random terms W_i are often called random 'errors'. We also use the phrase 'random terms' solely to refer to the W_is in this model, not the Y_is (though they are random, too).

The linear regression model is the special case of the general regression model with the regression function taken to be of the linear form $h(x) = \alpha + \beta x$, together with one additional assumption that it is quite standard to include in the basic linear regression model: the variance of the random term is a constant, $V(W_i) = \sigma^2$ say, for all $i = 1, 2, \ldots, n$.

Activity 4 *The mean and variance of Y_i*

If $E(W_i) = 0$ and $V(W_i) = \sigma^2$, what are $E(Y_i)$ and $V(Y_i)$?

In addition to the results of Activity 4, the response variables Y_i are independent (because the W_is are).

A schematic example of the linear regression model is given in Figure 11 (overleaf). There, $h(x) = 6x - 5$ (the case $\alpha = -5$, $\beta = 6$) is the regression function which represents the main, linear trend in the model. As in Figure 8, the distribution of $Y = h(x) + W$ is also shown for each of a number of values of x.

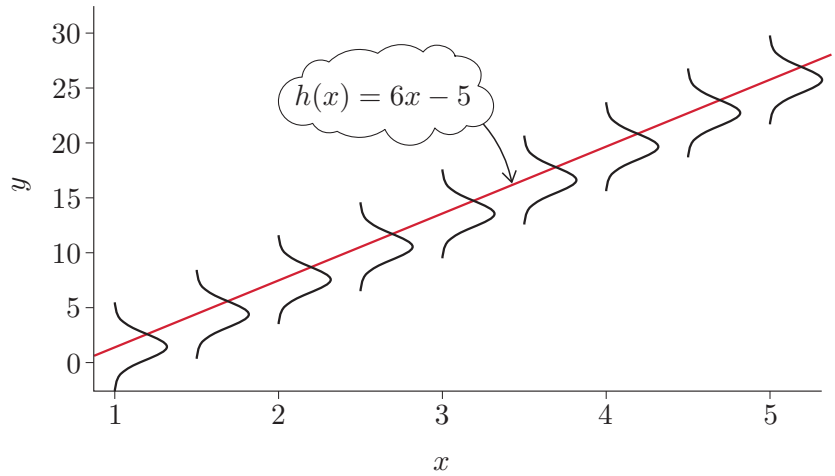


Figure 11 The regression function $h(x) = 6x - 5$ and the distribution of $Y = 6x - 5 + W$ at $x = 1, 1.5, 2, 2.5, \ldots, 5$

The line $y = \alpha + \beta x$ is called the **regression line**. As already mentioned in Example 6 and Activity 3, the parameters α and β can be interpreted as the intercept and slope of the regression line. This interpretation is the same as the usual mathematical one for a straight line. In case you need a reminder:

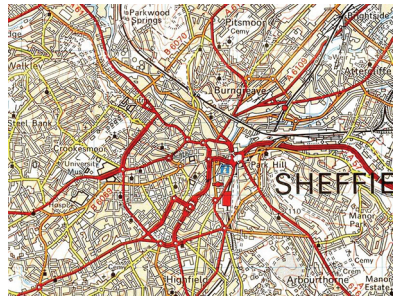
- the intercept α is the value taken by the line when $x = 0$
- the slope β gives how much y changes for every unit change in x. It is also the derivative of the line (for all x). If $\beta > 0$, the line is increasing; if $\beta = 0$, the line is the constant α; and if $\beta < 0$, the line is decreasing.

In the definition of the linear regression model, no assumption has been made about the entire distribution of the random variables W_i or Y_i, just assumptions about their means and variances. However, in order to make inferences, such as testing hypotheses, producing confidence intervals, and so on, it is necessary to assume some distribution for the W_is (or equivalently, for the Y_is). Later in the unit, when inference for linear regression models is discussed, normality of the W_is will be assumed. (You will also learn how to check whether this assumption is reasonable.) If you study statistics further, you will learn about regression models where other distributions are assumed for the W_is or Y_is.

In Example 6 and Activity 3, you saw situations where linear regression models might be useful to describe the relationships between the variables, while a non-linear regression model might be more appropriate for the data in Example 7. Example 8 illustrates a special case of linear regression: the straight line relating Y to x is constrained to go through the origin, that is, the point $x = 0$, $y = 0$.

Example 8 *Distance by road*

Road maps can sometimes be deceptive in the impression they give of distances between two locations. The data in Table 5 are the map distance (that is, the straight-line distance) and the distance by road (both in miles) between twenty different pairs of locations in and around Sheffield. The data raise the following questions. What is the relationship between the two variables? How well can the road distance be predicted from the map distance?



It is clear from the table that the road distance exceeds the map distance in every case. This is hardly surprising: roads tend to have bends, adding to the distance between two points. A scatterplot of road distance against map distance is given in Figure 12.

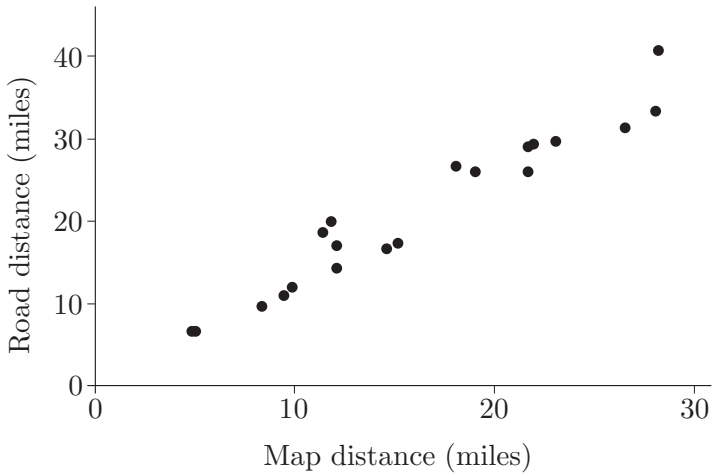


Figure 12 Road distance against map distance between pairs of locations in and around Sheffield

Table 5 Distances in and around Sheffield

Road distance (miles)	Map distance (miles)
10.7	9.5
11.7	9.8
6.5	5.0
25.6	19.0
29.4	23.0
16.3	14.6
17.2	15.2
9.5	8.3
18.4	11.4
28.8	21.6
19.7	11.8
31.2	26.5
16.6	12.1
6.5	4.8
29.0	22.0
25.7	21.7
40.5	28.2
26.5	18.0
14.2	12.1
33.1	28.0

(Source: Gilchrist, W. (1984) *Statistical Modelling*, Chichester, John Wiley and Sons, p. 5)

The plot suggests a roughly linear relationship between the two measures. However, the appropriate model here is a little different from that considered in the previous examples. If the map distance between two points is zero (if the two points are the same), then the road distance will also be zero. Therefore the line fitted to the data should go through the origin. That is, the model relating Y (road distance) to x (map distance) should have zero intercept and, since a straight line appears to continue to be a good model all the way to the origin, take the form

$$Y_i = \gamma x_i + W_i.$$

In this model, the parameter γ represents the factor by which a map distance needs to be multiplied to give an estimate of the road distance. The random term W_i again accounts for the scatter identified in the data. Assuming constant variance of the W_is, a linear regression model may be used for the relationship between the variables.

Notice that the letter γ was used for the slope parameter in the model constrained to go through the origin. It is useful to distinguish in this way between the slopes of the two straight-line models $Y_i = \alpha + \beta x_i + W_i$ and $Y_i = \gamma x_i + W_i$. The constrained model goes by the natural name of **regression through the origin**.

This section concludes with a few points worth noting about linear regression models. First, it is important to realise that it is not necessary to formulate any reason why the relationship between the response variable and the explanatory variable is linear. It is sufficient to argue on the basis of the scatterplot that the relationship *appears* to be linear. Remember also that linearity has been assumed only *within the range* of the data (or just outside the range in Example 8); as mentioned before, you should be cautious about extrapolating outside the range of the data, that is, about assuming that the linearity continues outside the range of the observed data. Finally, you should be aware that statisticians often fit a straight line to data even when there are reasons to believe that something more elaborate is really appropriate. (If you know about Taylor series expansions, you might know that some very complicated curves can be approximated over limited domains by straight lines.)

2 Fitting a linear regression model

In Section 1, you saw several examples of scatterplots where it looked as though a straight-line model would fit the scattered data points (x_i, Y_i) moderately well (in some cases, very well). A practical problem now arises: which straight line fits the data best? In this section, you will see how a technique called the method of least squares can be used to fit the 'best' straight line to the data. The fitted line is called the least squares line.

The method of least squares is discussed in Subsection 2.1. This subsection includes some work using your computer. The special case of a linear regression model where the line is constrained to go through the origin is considered in Subsection 2.2; how to obtain the least squares line for this simple model is described in some detail. In Subsection 2.3, the formula for the least squares line for an 'unconstrained' linear regression model is given without proof.

Fitting a straight line to data by least squares is a method for estimating the parameters α and β of that line. As you will see, the method is quite simple, general and 'natural'. However, you know from Unit 7, in particular, that maximum likelihood is often used to estimate parameters of models from data. So why not use that? Well, in Subsection 2.4, it is shown that if the random terms W_i are from normal distributions, then the slope and the intercept of the least squares line are, in fact, the same as those of the line obtained using the method of maximum likelihood.

2.1 The method of least squares

We begin by looking at a small, illustrative dataset on the way cholesterol level changes with age.

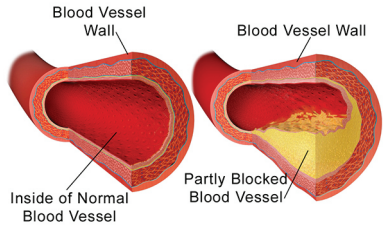
Example 9 *Cholesterol and age*

The data given in Table 6 are the plasma levels (in mg/ml) of total cholesterol in 11 patients aged over 40 who were admitted to a clinic with hyperlipoproteinaemia, a disorder characterised by high levels of lipoproteins in the blood.

Table 6 Cholesterol levels (in mg/ml) and ages (in years)

Age	43	46	48	49	50	52	52	57	57	58	63
Cholesterol	3.8	3.5	4.2	4.0	3.3	4.0	4.3	4.5	4.1	3.9	4.6

(Source: data extracted from a dataset in Krzanowski, W.J. (1998) *An Introduction to Statistical Modelling*, London, Arnold, Chapter 3)



Normal and Partly Blocked Blood Vessel

So-called bad cholesterol contributes to plaque, which narrows arteries and can lead to heart disease

The scatterplot of the data in Figure 13 suggests a roughly linear upward trend but, of course, there is some scatter about the trend due to random variation.

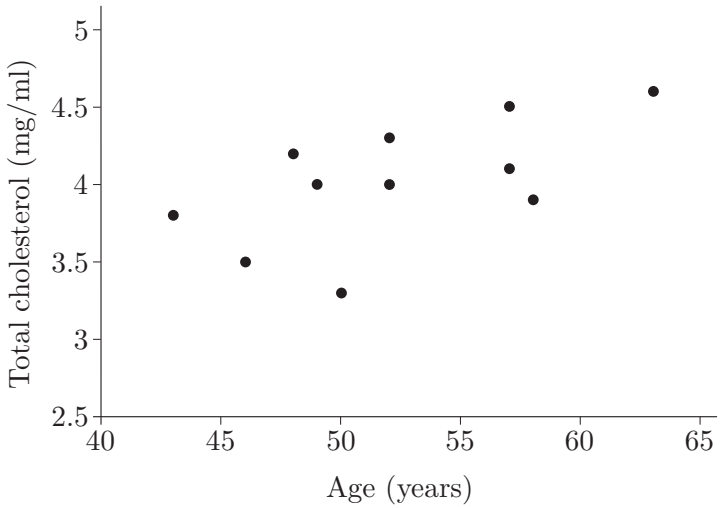


Figure 13 Total cholesterol against age

It seems that a straight-line model of the form

$$Y_i = \alpha + \beta x_i + W_i$$

might describe the data moderately well. Here x_i denotes age (in years), Y_i denotes total cholesterol level (in mg/ml) and W_i is a random term accounting for the scatter. How do we determine the equation of the line which is 'better' than any other line?

This is sometimes more grandiosely called the 'principle of least squares'.

The traditional criterion underlying the estimation of the line that best fits data is the minimisation of a sum of squares of quantities called *residuals*; the resulting method is called the *method of least squares*.

In general, if a line of the form $y = \alpha + \beta x$ is to be fitted to data points (x_i, y_i), then the **residual** w_i for the point (x_i, y_i) is the difference between the observed value y_i and the value of $\alpha + \beta x_i$:

$$w_i = y_i - (\alpha + \beta x_i).$$

The residuals are illustrated for the cholesterol data and one particular choice of line in Figure 14. Notice that the size of each residual is equal to the length of the dashed line joining the data point to the fitted line *vertically* (rather than horizontally or at an angle of 90° to the fitted line).

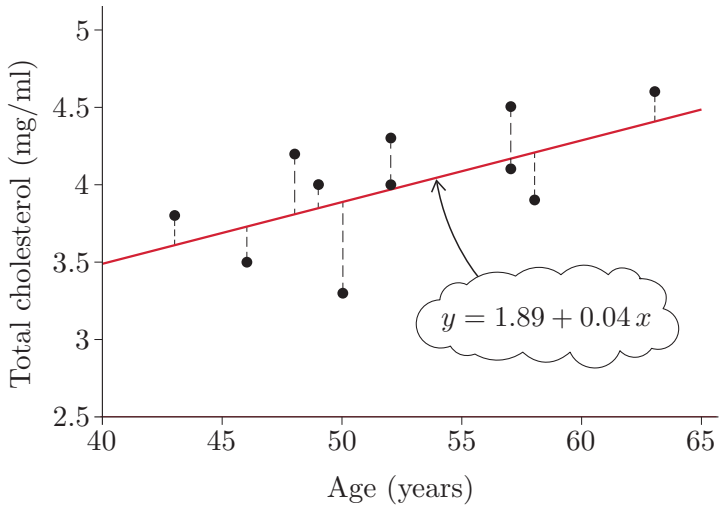


Figure 14 The residuals $w_i = y_i - (\alpha + \beta x_i)$ for one choice of α and β

If the line fits the data well, then the residuals will be small (in absolute value); if not, then at least some of the residuals will be large (in absolute value). When using least squares to choose a best-fitting line, the sum of the squares of the residuals is minimised. The reasoning behind using the sum of *squares* of residuals is the same as that behind summing (and then averaging) squared deviations to form the sample variance (as in Unit 1). You can remind yourself of that reasoning in the following activity.

Activity 5 *Method of least what?*

(a) What is the main reason for choosing the parameters of a regression model to minimise the sum of squared residuals rather than the sum of residuals?

(b) Can you suggest another quantity of the form 'sum of function of residuals' which would have a similar effect to using the function 'square'?

The sum of squared residuals is more often called the **residual sum of squares** and is given by

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2. \quad (1)$$

A small sum indicates a good fit of the line to the data, while a large sum indicates a poor fit. Note that there are two unknown quantities in the expression on the right-hand side of Equation (1): the parameters α and β. The residual sum of squares varies for different values of these parameters. We are interested in the values of α and β that minimise the residual sum of squares (that is, the values that minimise the deviations between the data and the fitted model). The minimising values of α and β are called the **least squares estimates** of the parameters of the regression line, and are denoted by $\widehat{\alpha}$ and $\widehat{\beta}$.

The rest of the work in this subsection consists of a chapter in Computer Book C, in which you can explore the ideas behind the method of least squares.

Refer to Chapter 1 of Computer Book C for the rest of the work in this subsection.

2.2 The least squares line through the origin

Before the formulas for the least squares estimates for the linear regression model are given, a slightly simpler model will be looked at more closely. In Example 8, it was suggested that a good model for the data considered there might be a straight line with the constraint that the line passes through the origin. In this subsection, you will see how to derive the least squares line for this constrained model.

In Example 8, actual road distances between locations in and around Sheffield were compared with direct distances taken from a map. It was decided to fit a straight line passing through the origin to the data. The proposed model was

$$Y_i = \gamma x_i + W_i.$$

A line of the form $y = \gamma x$ has been drawn on the scatterplot of the data in Figure 15 (overleaf) for illustrative purposes only: the value of the slope γ that corresponds to the best straight line through the data is not yet known. The observed residuals based on this line, which are also shown in Figure 15, are in this case given by

$$w_i = y_i - \gamma x_i.$$

Road map or satnav?

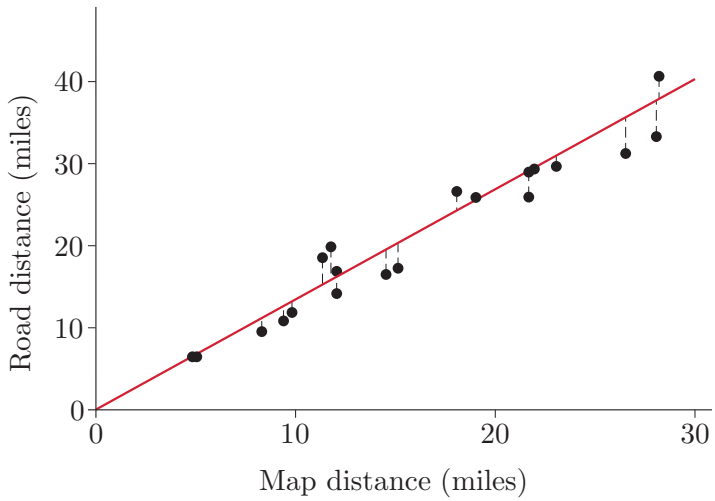


Figure 15 The residuals $w_i = y_i - \gamma x_i$ for one choice of γ

For this model (with the constraint that the straight line goes through the origin), the residual sum of squares is given by

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} (y_i - \gamma x_i)^2 = R(\gamma), \tag{2}$$

say. Here, since $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the observed data and therefore are known, there is only one unknown quantity in the residual sum of squares: the slope parameter γ. So the residual sum of squares can be thought of as a function of γ, which we have called $R(\gamma)$. We wish to estimate γ by the value that minimises $R(\gamma)$, that is, that minimises the residual sum of squares. The minimising value of γ is called the *least squares estimate* of the slope of the regression line, and is denoted $\widehat{\gamma}$.

Let us start by taking a look at a graph of $R(\gamma)$. This graph is shown for the Sheffield distance data in Figure 16. ($R(\gamma)$ is plotted only for a limited range of values of γ; for other values of γ, $R(\gamma)$ is even larger and off the scale.) A clear minimum at a value of γ a bit less than 1.3 can be observed.

Now, it turns out that a graph of $R(\gamma)$ always looks very much like the graph of $R(\gamma)$ in Figure 16, whatever the data on which it is based. This is because $R(\gamma)$ is a *quadratic* function of γ, that is, $R(\gamma)$ is of the form $a\gamma^2 + b\gamma + c$ for some coefficients a, b and c.

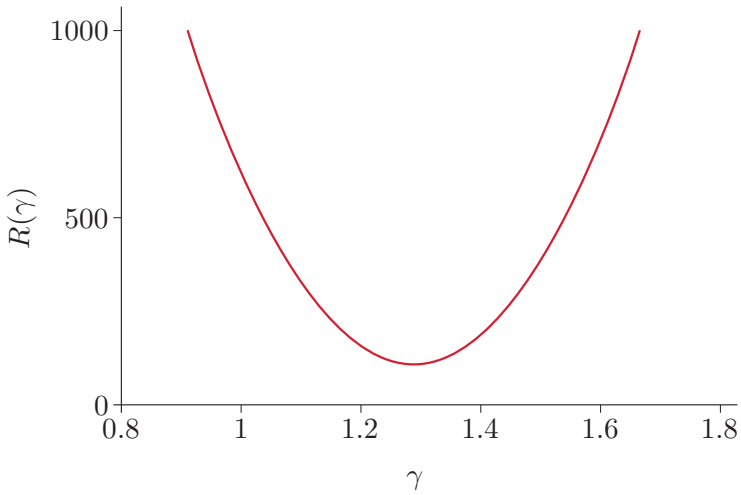


Figure 16 The residual sum of squares

Activity 6 *$R(\gamma)$ as a quadratic function of γ*

By expanding the squared bracket in Equation (2), identify expressions for a, b and c in the representation $R(\gamma) = a\gamma^2 + b\gamma + c$.

In fact, quadratic functions in general look either like that in Figure 16 – 'down-then-up', with a clear minimum – or else like upside-down versions of the function in Figure 16 – 'up-then-down', with a clear maximum. The determinant of shape of a quadratic function is the sign of a. In particular, if $a > 0$, the quadratic function is of the 'down-then-up' variety. (To see this, notice that $a\gamma^2 + b\gamma + c$ necessarily becomes very large as γ becomes very large in absolute value, when $a > 0$.) And it is always the case that $a > 0$ for $R(\gamma)$ because, as you showed in Activity 6, then $a = \sum_{i=1}^{n} x_i^2$ is a sum of squared quantities.

Quadratic curves, or parabolas, abound in the built environment. This one – with $a < 0$! – is the Memorial Cenotaph in Hiroshima Peace Memorial Park, Japan.

Activity 7 *Minimising a quadratic function and hence $R(\gamma)$*

(a) Consider the general quadratic function $ax^2 + bx + c$ with $a > 0$.

(i) Confirm that

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a} + c. \tag{3}$$

(ii) Hence argue that the minimum of $ax^2 + bx + c$ when $a > 0$ is given by

$$x = -\frac{b}{2a}.$$

(b) By combining the results of Activity 6 and part (a)(ii) above, give an expression for the value of γ that minimises $R(\gamma)$.

For simplicity, the limits $i = 1$ and $i = n$ on the summation symbols have been omitted, and you can do the same from here on.

From the solution to Activity 7(b), the value of γ which minimises the residual sum of squares is

$$\widehat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2};$$

and $\widehat{\gamma}$ is the slope of the best straight line through the scattered points that passes through the origin. The equation of the least squares line can be written

$$y = \widehat{\gamma}\, x.$$

These results are summarised in the following box.

The least squares line through the origin

Suppose that it is desired to fit a regression line through the origin and that a scatterplot of data points (x_i, y_i), $i = 1, 2, \ldots, n$, suggests that an appropriate regression model is of the form

$$Y_i = \gamma\, x_i + W_i,$$

where the W_is are independent with zero mean and constant variance. Then the least squares estimate $\widehat{\gamma}$ of γ is given by

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

The equation of the least squares line through the origin is

$$y = \widehat{\gamma}\, x.$$

Example 10 *Road distances*

In Example 8, Sheffield map and road distances were given. As you have just seen, the least squares estimate of γ in the regression model through the origin depends on two summary statistics, $\sum x_i y_i$ and $\sum x_i^2$. The values of these quantities for the distance data are as follows:

$$\begin{aligned}\sum x_i y_i &= (9.5 \times 10.7) + (9.8 \times 11.7) + \cdots + (28.0 \times 33.1)\\ &= 101.65 + 114.66 + \cdots + 926.80 = 8026.25,\end{aligned}$$

$$\begin{aligned}\sum x_i^2 &= 9.5^2 + 9.8^2 + \cdots + 28.0^2\\ &= 90.25 + 96.04 + \cdots + 784.00 = 6226.38.\end{aligned}$$

So the least squares estimate of the slope γ is

$$\widehat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{8026.25}{6226.38} \simeq 1.289.$$

This value corresponds to the exact minimum of $R(\gamma)$ as shown in Figure 16. The equation of the least squares line through the scattered data points is

$$y = 1.289\,x,$$

or, perhaps more intelligibly,

$$\text{road distance} = 1.289 \times \text{map distance}.$$

The least squares line is shown in Figure 17. You can see that the fit is really quite good; the residuals are not large. It seems that the road distance can be predicted quite well from the map distance by multiplying the latter by a factor of 1.289 (that is, by inflating the map distance by a little less than 30%).

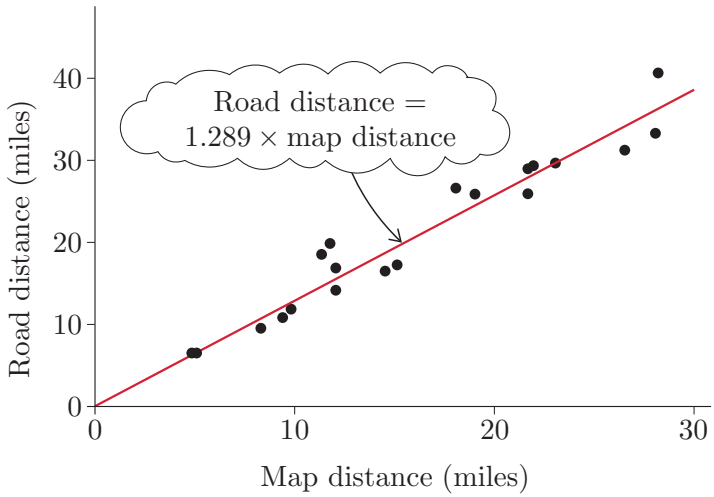


Figure 17 Road distance against map distance, and the least squares line

Activity 8 *Beetles in brackets*

In a botanical experiment, a researcher wanted to estimate the number of a particular species of beetle (*Diaperis maculata*) within fruiting bodies (called brackets) of the birch bracket fungus *Polyporus betulinus*. (This is a shelf fungus that grows on the trunks of dead birch trees.) When the brackets are stored in a laboratory, the beetle larvae within them mature over several weeks. The adults then emerge and can be removed and counted. The bracket weight (in grams) and the number of beetles in each bracket were recorded for a sample of 25 brackets. (Source: Pielou, E.C. (1974) *Population and Community Ecology – Principles and Methods*, New York, Gordon and Breach, pp. 117–21.)

A *Diaperis maculata* fungus beetle

It is suggested that a straight line through the origin might provide an adequate model for the data. The relevant summary statistics for the bracket weight x and the count of beetles y are:

$$\sum x_i^2 = 796\,253, \quad \sum x_i y_i = 219\,817.$$

Calculate the equation of the least squares line through the origin for the data.

2.3 The least squares line

Now consider the 'unconstrained' linear regression model

$$Y_i = \alpha + \beta x_i + W_i.$$

In Subsection 2.1, you saw that the least squares estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of the parameters α and β are the values that minimise the residual sum of squares,

$$\sum (y_i - (\alpha + \beta x_i))^2.$$

There are other routes to the same answer too.

This sum can be minimised using an extension to two parameters of the technique that was used when fitting the least squares line through the origin in Subsection 2.2. However, you will be spared the details. For present purposes it is sufficient simply to write the estimates down. However, before writing them down, it is useful to introduce the following standard shorthand notation.

$$S_{xx} = \sum (x_i - \overline{x})^2 \tag{4}$$

$$S_{yy} = \sum (y_i - \overline{y})^2 \tag{5}$$

$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) \tag{6}$$

The expression $(x_i - \overline{x})$ is the deviation of x_i from the mean $\overline{x}$ of the x values, and $(y_i - \overline{y})$ is the deviation of y_i from $\overline{y}$. Thus each term in the sums S_{xx}, S_{yy} and S_{xy} consists of two deviations multiplied together. For this reason S_{xx} and S_{yy} are sometimes called sums of squared deviations, while S_{xy} is a sum of products of deviations. Note that $S_{xx}/(n-1)$ and $S_{yy}/(n-1)$ are the sample variances of the x values and y values, respectively, where n is the sample size.

The easiest way to calculate S_{xx}, S_{yy} and S_{xy} is usually by using the alternative formulas in Equations (7), (8) and (9) below.

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = \sum x_i^2 - n\overline{x}^2 \tag{7}$$

$$S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = \sum y_i^2 - n\overline{y}^2 \tag{8}$$

$$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = \sum x_i y_i - n\overline{x}\,\overline{y} \tag{9}$$

That the two versions of each formula within the box immediately above are equal to one another is a simple consequence of recalling that $\overline{x} = \sum x_i/n$ and $\overline{y} = \sum y_i/n$. That Equations (7), (8) and (9) are equivalent to Equations (4), (5) and (6), respectively, takes a bit more algebraic manipulation that you can do for yourself in the next activity.

Activity 9 *Equivalence of formulas*

(a) Check that Equation (7) is equivalent to Equation (4) by manipulating Equation (4).

(b) Why can you now claim that Equation (8) is equivalent to Equation (5) without further mathematical manipulation?

(c) Check that Equation (9) is equivalent to Equation (6) by manipulating Equation (6).

Activity 10 *Calculating S_{xx}, S_{yy} and S_{xy}*

The summary statistics for the cholesterol data from Example 9 are given by

$$n = 11, \quad \sum x_i = 575, \quad \sum y_i = 44.2,$$

$$\sum x_i^2 = 30\,409, \quad \sum y_i^2 = 179.14, \quad \sum x_i y_i = 2324.8.$$

Use Equations (7), (8) and (9) to calculate S_{xx}, S_{yy} and S_{xy}.

We are now ready to write down the formulas for the least squares estimates of the parameters of the linear regression model. First, the least squares estimate $\widehat{\beta}$ of β is given by

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}.$$

A similar expression can be written down for $\widehat{\alpha}$, the least squares estimate of α, but it is easier to use the value of $\widehat{\beta}$ to calculate $\widehat{\alpha}$ as follows:

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}.$$

Then the equation of the least squares line can be written as

$$y = \widehat{\alpha} + \widehat{\beta}\,x.$$

This can be rewritten in various equivalent ways, a popular one being in terms of $\overline{x}$, $\overline{y}$ and $\widehat{\beta}$:

$$y = (\overline{y} - \widehat{\beta}\,\overline{x}) + \widehat{\beta}\,x = \overline{y} + \widehat{\beta}\,(x - \overline{x}).$$

These results may be summarised as follows.

The least squares line

Suppose that the scatterplot of the data points (x_i, y_i), $i = 1, 2, \ldots, n$, suggests that an appropriate regression model might be of the form

$$Y_i = \alpha + \beta x_i + W_i,$$

where the random terms W_i are independent with zero mean and constant variance. Then the least squares estimate $\widehat{\beta}$ of the slope of the regression line is

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}},$$

where $S_{xx} = \sum(x_i - \overline{x})^2$ and $S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y})$. The least squares estimate $\widehat{\alpha}$ of the intercept α is given by

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}.$$

The equation of the least squares line is

$$y = \widehat{\alpha} + \widehat{\beta}\,x = \overline{y} + \widehat{\beta}(x - \overline{x}).$$

An interesting property of the least squares regression line is given in the next activity.

Activity 11 *Passing through the centroid*

The point on the scatterplot $(\overline{x}, \overline{y})$ is known as the *centroid* of the data. Show that the least squares line passes through the centroid.

Example 11 *Fitting a line to the cholesterol data*

We are now in a position to use least squares to produce a best-fitting line to the cholesterol data discussed in Example 9 and Activity 10.

The summary statistics for the cholesterol data were given in Activity 10. In that activity, you found that $S_{xx} \simeq 352.182$ and $S_{xy} \simeq 14.345$. Hence the least squares estimate of the slope β is

$$\widehat{\beta} \simeq \frac{14.345}{352.182} \simeq 0.04.$$

Using $\widehat{\beta}$ and the summary statistics $n = 11$, $\sum x_i = 575$ and $\sum y_i = 44.2$, the least squares estimate of the intercept α is

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = \frac{44.2}{11} - \widehat{\beta} \times \frac{575}{11} \simeq 1.89.$$

The value of $\widehat{\beta}$ prior to final rounding has been used in this calculation.

So the equation of the fitted least squares line is

$$y = 1.89 + 0.04\,x.$$

Alternatively, the model can be written as

$$\text{total cholesterol} = 1.89 + 0.04 \times \text{age},$$

where total cholesterol is measured in mg/ml and age is given in years. The least squares line is shown in Figure 18. (It is the same line as was shown in Figure 14.) The line appears to fit the data reasonably well.

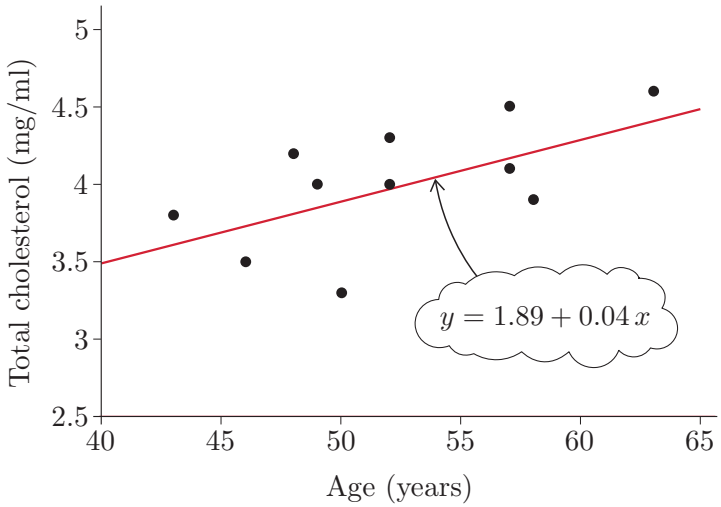


Figure 18 Total cholesterol against age, and the least squares line

In terms of interpretation, the estimated value of the intercept, $\widehat{\alpha}$, is of little interest in this particular context because it refers to a person of age 0 years, whereas the linear model is fitted – and assumed appropriate – to data on people over the age of 40 years. The estimated value of the slope, $\widehat{\beta} \simeq 0.04$, is of interest, however. It tells us that, for patients with hyperlipoproteinaemia aged over 40 years, an increase in age of one year is expected to lead, on average, to an increase in total cholesterol of about 0.04 mg/ml.

Another use of the least squares fitted regression line is for *prediction*. Suppose that another individual of the same type as those to whom the line was fitted has a value x_0 say, of the explanatory variable; however, we do not yet know the value of the response variable, y_0 say, for this individual. The least squares line allows us to predict what we think that value might be, by setting $x = x_0$ in the equation of the least squares line:

$$y_0 = \widehat{\alpha} + \widehat{\beta}\, x_0.$$

Prediction via statistics or crystal ball?

Example 12 *Predicting total cholesterol*

As an example, the least squares line obtained in Example 11, $y = 1.89 + 0.04\,x$, could be used to predict the total cholesterol level of a person with hyperlipoproteinaemia aged over 40. For example, for someone aged 45, the fitted line predicts a total cholesterol level of

$$1.89 + 0.04 \times 45 = 3.69 \text{ mg/ml},$$

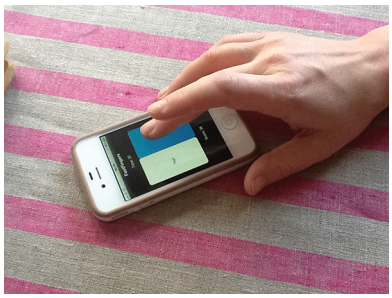
while for someone aged 60, the fitted line predicts a total cholesterol level of

$$1.89 + 0.04 \times 60 = 4.29 \text{ mg/ml}.$$

Notice, however, that these are single-value or 'point' predictions, without any indication of uncertainty concerning that prediction. We are not claiming that, for example, in Example 12, everyone with hyperlipoproteinaemia aged 60 should have a cholesterol value of exactly 4.29 mg/ml, just that 4.29 mg/ml seems to be a reasonable prediction of the average value of cholesterol for people with this condition aged 60. Indeed, any prediction of the form $y_0 = \widehat{\alpha} + \widehat{\beta}\,x_0$ is actually an estimate of the average value, $\alpha + \beta\,x_0$, of the response for an individual with $x = x_0$. As point estimates have corresponding interval estimates (Unit 8), so point predictions have corresponding interval predictions; these will be considered briefly in Subsection 4.3 to follow.

Activity 12 *Finger-tapping*

Finger-tapping is a fairly standard psychological task performed by subjects to assess alertness through manual dexterity.



Modern finger-tap testing

An experiment was carried out to investigate the effect of the stimulant caffeine on performance on a simple physical task. Thirty male college students were trained in finger-tapping; it is the speed of finger-tapping that is of interest. They were then randomly divided into three groups of ten, and the students in each group received different doses of caffeine (0 mg, 100 mg and 200 mg). Two hours after treatment, each student was required to do finger-tapping, and the number of taps achieved per minute was recorded. The recorded figures for each of the 30 students are given in Table 7.

Table 7 Finger-tapping

Caffeine dose (mg)	Taps per minute									
0	242	245	244	248	247	248	242	244	246	242
100	248	246	245	247	248	250	247	246	243	244
200	246	248	250	252	248	250	246	248	245	250

(Source: Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd edn, New York, John Wiley and Sons, p. 425)

It is not possible to deduce very much about the shape of the variation in tapping performances at each dose level from the scatterplot shown in Figure 19. However, there is some evidence of a linear upward trend.

For clarity, coincident points are shown slightly displaced, or 'jittered', vertically in Figure 19.

Suppose that we wish to model the relationship between tapping performance Y and caffeine dose x by a linear regression model. The summary statistics for the data in Table 7 are given by

$$n = 30, \quad \sum x_i = 3000, \quad \sum y_i = 7395,$$

$$\sum x_i^2 = 500\,000, \quad \sum x_i y_i = 743\,000.$$

(a) Use the summary statistics to calculate S_{xx} and S_{xy}.

(b) Calculate the equation of the least squares line for the data.

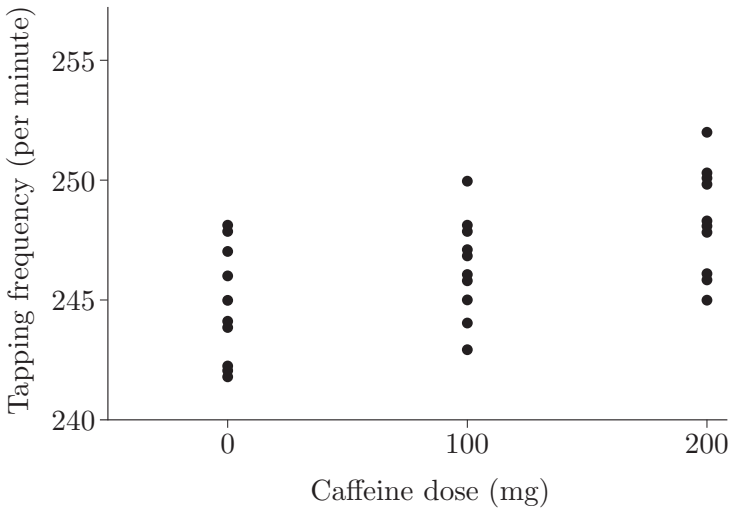


Figure 19 Tapping performance against caffeine dose

(c) Interpret what the values of the least squares estimates of the parameters of the regression line tell us.

(d) Use the equation of the fitted least squares line to predict the number of taps per minute of a student treated with 50 mg of caffeine.

In practice, a computer is almost always used to fit least squares lines. You will do this using Minitab in Section 3.

2.4 Maximum likelihood estimation in regression

In this subsection, you will see that if normality is assumed for the random terms W_i in a linear regression model, then the least squares estimates of the parameters of the line are also the maximum likelihood estimates of those parameters. This argument further justifies the use of least squares estimation in regression. The argument is given in full for completeness and worked through in Screencast 11.1. If you cannot follow all the details, don't worry: the result is worth knowing but you won't need to be able to reproduce the argument leading to it.

Suppose that the random terms W_i, $i = 1, 2, \ldots, n$, come from independent normal distributions, that is,

$$W_i \sim N(0, \sigma^2), \quad i = 1, 2, \ldots, n.$$

Notice that each of these normal distributions has zero mean and the same variance, σ^2, these being general properties of the random terms in the linear regression model. Equivalently, since $Y_i = \alpha + \beta x_i + W_i$ and the $\alpha + \beta x_i$ terms can be treated as constants,

See Activity 4 in this unit and Distributional Result (3) of Unit 6.

This is just for simplicity: the least squares estimates of α and β are the maximum likelihood estimates of α and β when σ^2 is not known too.

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i = 1, 2, \ldots, n,$$

and these normal distributions are independent also.

Suppose for the remainder of this subsection that the value of σ^2 is known. Then, following the discussion of likelihood estimation for continuous distributions in Unit 7 (with two unknown parameters α and β replacing the single unknown parameter θ there), the likelihood in this case is

$$L(\alpha, \beta) = f(y_1; \alpha, \beta) \times f(y_2; \alpha, \beta) \times \cdots \times f(y_n; \alpha, \beta),$$

where $f(y_i; \alpha, \beta)$ is the p.d.f. of Y_i when $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$. Using the p.d.f. of the normal distribution from Unit 6, the likelihood is therefore

$$\begin{aligned} L(\alpha, \beta) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_1 - (\alpha + \beta x_1)}{\sigma}\right)^2\right] \\ &\quad \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_2 - (\alpha + \beta x_2)}{\sigma}\right)^2\right] \\ &\qquad \vdots \\ &\quad \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_n - (\alpha + \beta x_n)}{\sigma}\right)^2\right] \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2\right]. \end{aligned}$$

Now, the maximum likelihood estimates of α and β (when σ^2 is known) are the values that maximise the likelihood. The first term in the likelihood is a constant (since σ^2 is known) and the second term is of the form

$$\exp\{-kR(\alpha, \beta)\},$$

where $k = 1/(2\sigma^2)$ is a positive constant and

$$R(\alpha, \beta) = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2.$$

Since the function e^{-kx}, $k > 0$, is decreasing (see, for example, Figure 2(a) of Unit 5 or Figure 13 of Unit 8), the second term in the likelihood (and hence the whole product) is maximised with respect to α and β when the quantity $R(\alpha, \beta)$ is minimised. However, $R(\alpha, \beta)$ can be recognised as precisely the residual sum of squares, given in Equation (1), that is minimised to find the least squares estimates.

So, under the assumption of normality with known variance, the least squares estimates of α and β are the same as the maximum likelihood estimates of α and β.

The above argument is reviewed in Screencast 11.1.

Screencast 11.1 Maximum likelihood estimation of regression parameters assuming normality is the same as least squares estimation

Exercise on Section 2

Exercise 1 *The least squares line for Forbes's data*

For Forbes's data, which are given in Table 1, the summary statistics are as follows:

$$n = 17, \quad \sum x_i = 426, \quad \sum y_i = 3450.2,$$

$$\sum x_i^2 = 10\,820.9966, \quad \sum x_i y_i = 86\,735.495.$$

(a) Use the summary statistics to calculate the equation of the least squares line for the data.

(b) Interpret what the values of the least squares estimates of the parameters of the regression line tell us.

(c) Use the fitted line to obtain a point prediction of the boiling point of water at an atmospheric pressure of 25 inches Hg.

3 Checking the assumptions

You will shortly be using Minitab to fit linear regression models to some of the datasets described in Sections 1 and 2. Before doing that, however, there is an important question to ask that was neglected in Section 2: how can we check that a fitted model is reasonable? For the linear regression model

$$Y_i = \alpha + \beta x_i + W_i,$$

the basic assumptions are as follows.

1 The random terms W_i are independent.

2 The W_is have zero mean and constant variance σ^2.

Remember that, as you showed in Activity 4, the assumption of zero mean for the W_is is equivalent to the assumption that a line of the form $\alpha + \beta x$ is appropriate for the mean of the Y_is:

$$E(Y_i) = \alpha + \beta x_i.$$

So, together, the two basic assumptions of linear regression are that the W_is are independent random variables with zero mean and variance σ^2. Or, equivalently, the two basic assumptions of linear regression are that the Y_is are independent random variables with mean $\alpha + \beta x_i$ and variance σ^2.

Assumption 1 on independence of the $W_i s$ can be checked, although it will not be done here: independence has to do with the design of the experiment and how the data were collected. It will usually be clear whether or not the independence assumption is justifiable.

It's always worth checking your assumptions about the British weather

Assumption 2, that the W_is have zero mean and constant variance, can be checked using a diagram called a *residual plot*. This is the topic of the next subsection.

3.1 Residual plots

We wish to check the properties of the W_is. Well, the linear regression model can be rearranged as

$$W_i = Y_i - (\alpha + \beta x_i).$$

However, because α and β are unknown, we immediately come up against the problem that the W_is cannot be observed. The W_is can, though, be estimated in the natural way via the estimated values of $\alpha + \beta x_i$. The latter are

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i,$$

Fitted values differ from predicted values by their values of x: fitted values are for observed values $x = x_i$ while predicted values are for other values $x = x_0$.

which we will now refer to by the standard nomenclature of **fitted values**. The required estimates of the W_is are therefore the differences between the observed values, y_i, and the fitted values, $\widehat{y}_i$, namely the quantities

$$w_i = y_i - \widehat{y}_i = y_i - (\widehat{\alpha} + \widehat{\beta} x_i).$$

And these you have seen before: the w_is are, in the terminology of Subsection 2.1, residuals – specifically residuals from the least squares fitted line (but they will be referred to just as residuals from now on). So the modelling assumptions can be checked by looking to see whether the residuals w_i, used as estimates of the random terms W_i, might have come from some distribution with zero mean and variance σ^2 (for some value of σ^2).

Elsewhere, a residual plot is sometimes defined to be a scatterplot of the observed residuals w_i against the values x_i of the explanatory variable. For the linear regression model discussed in this section, the two plots are the same after rescaling.

In this module, a **residual plot** is defined to be a scatterplot of the observed residuals w_i against the fitted values $\widehat{y}_i$. If Assumption 2 is satisfied, then the residuals should be scattered about zero in a random, unpatterned fashion. Note that the residuals are the deviations from the fitted model: a pattern in the residual plot would suggest a dependence between the residuals and the corresponding fitted values, indicating a breach of the assumption that the random terms W_i, which the residuals w_i are estimating, have zero mean and constant variance. The key here is actually that the mean and variance of the random terms should both be *constant*. Patterns in the residuals when plotted against the fitted values suggest that either the mean or the variance or both are not constant. Figure 20 shows four typical shapes of residual plots.

Figure 20(a) is a residual plot with no apparent pattern of any kind in the residuals: this is the type of plot that you might expect to obtain when the assumptions are justified. There is a definite pattern in each of the other panels of Figure 20.

In Figure 20(b), when moving from left to right, from smaller to larger fitted values, the residuals are negative at first, then positive, then negative again. In general, a residual plot displaying a pattern such as this is an indication that the assumption of constant, zero mean may not be

valid – that is, that the relationship between the response and explanatory variables is not linear. (The pattern in the residuals gives an indication of what the regression function should have been – perhaps, in this instance, quadratic rather than linear.)

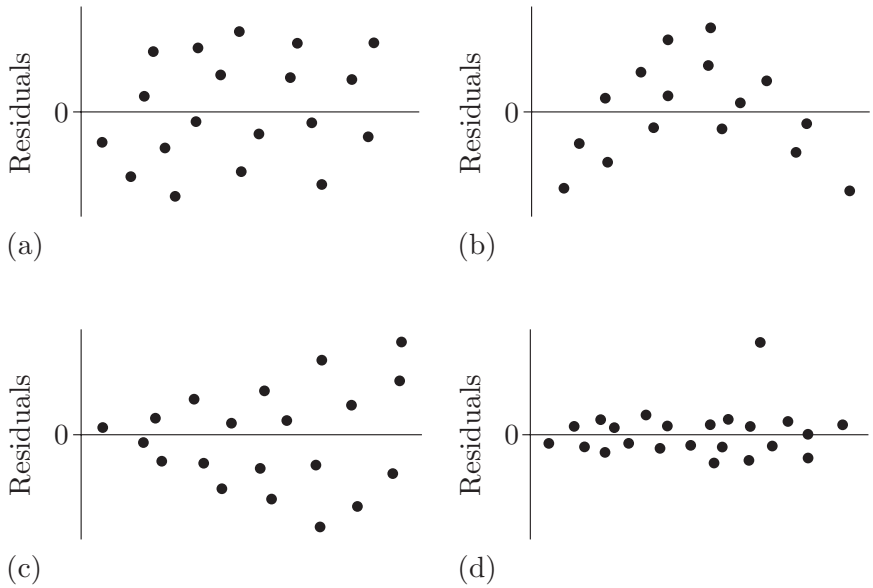


Figure 20 Residual patterns, when plotted against fitted values: (a) unpatterned, (b) a systematic discrepancy, (c) variance not constant, (d) an outlier

In Figure 20(c), the pattern is indicative of the variance of the random terms not being constant: the variability of the residuals (and hence presumably the variance of the random terms) increases as the fitted values increase. (In variations on this theme, the variance of the residuals might decrease as the fitted values increase, or even exhibit some other pattern, such as small variance then larger variance then smaller variance again.)

Finally, the residual plot in Figure 20(d) has a single residual that is considerably larger in magnitude than any of the others. The plotted point may correspond to an outlier.

Let's see what residual plots can tell us in an example and a couple of activities.

Example 13 *Checking residuals for the cholesterol data*

In Example 11, the least squares line was fitted to the cholesterol data. The equation of the line is

$$\text{total cholesterol} = 1.89 + 0.04 \times \text{age}.$$

This line was shown on a scatterplot of the data in Figure 18. The figure is repeated in Figure 21 (overleaf).

This figure is not a residual plot!

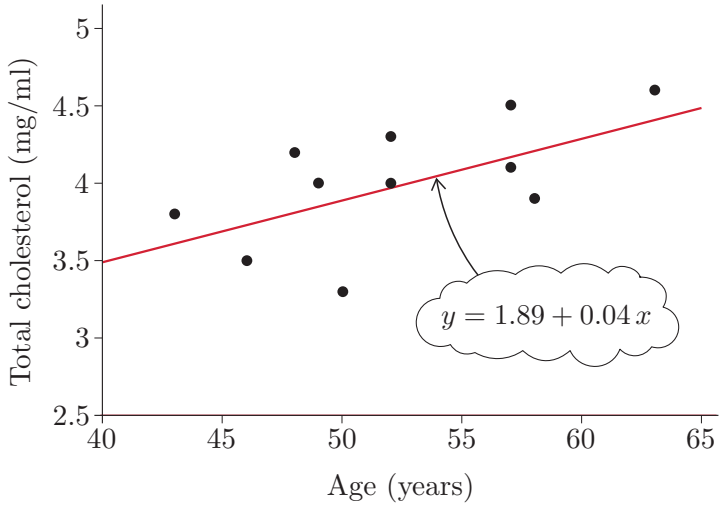


Figure 21 Total cholesterol against age, and the least squares line

In order to check Assumption 2, it is necessary to calculate the fitted values and the residuals. In this case, for each age $x_1, x_2, \ldots, x_n$, the fitted values $\widehat{y}_i$ are given by

$$\widehat{y}_i = 1.89 + 0.04\,x_i$$

and the residuals w_i are then found from

$$w_i = y_i - \widehat{y}_i = y_i - 1.89 - 0.04\,x_i.$$

A residual plot for the cholesterol data, that is, a scatterplot of the residuals w_i against the fitted values $\widehat{y}_i$, is shown in Figure 22.

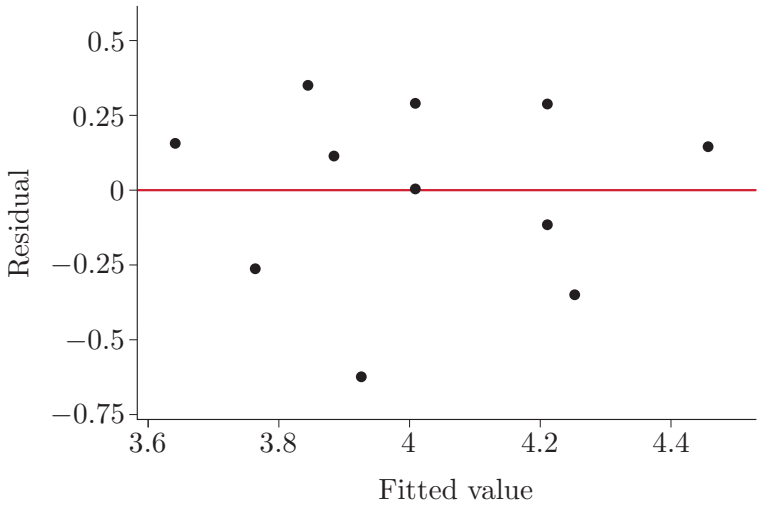


Figure 22 A residual plot for the cholesterol data

This plot shows no particular pattern, and there are approximately the same number of points below the line as above it; the points seem to be randomly scattered around zero. That is, Assumption 2 seems to be satisfied. So a linear regression model might provide an adequate model for these data.

Activity 13 *Checking residuals for Forbes's data*

Forbes's data on the way in which the boiling point of water depends on pressure were introduced in Example 3. The data were shown in Figure 4, and the linear regression model was fitted to the data by least squares in Exercise 1. The equation of the least squares fitted line is

$$y = 155.30 + 1.90\,x,$$

where y is the boiling point of water in °F and x is the pressure in inches of mercury.

If the modelling assumptions are reasonable for these data, then the W_is are observations with a constant, zero mean and an unknown but constant variance. A residual plot is given for this fitted regression line in Figure 23.

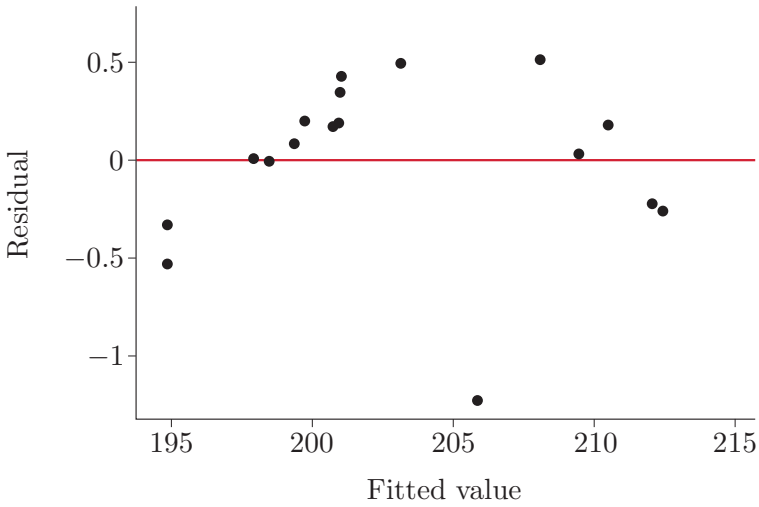


Figure 23 A residual plot for Forbes's data

Comment on what this plot tells you. Is the linear regression model a good one for these data?

Activity 14 *Defects in the Trans-Alaska oil pipeline*

In Subsection 5.4 of Unit 1, we looked at a dataset of size $n = 107$ concerning the measurement of defects in the Trans-Alaska oil pipeline. Depths of defects were measured in the field using ultrasonic measuring equipment and again, potentially more accurately, in the laboratory later. Interest now centres on how well calibrated in-field measurements of pipeline defects were, in the sense of how closely they depend on their corresponding laboratory measurements.

In Example 26 of Unit 1, with the help of a scatterplot interpretation checklist, we decided that the data exhibit a 'moderately strong' positive linear relationship with no obvious outliers. The data therefore seem ripe

Ultrasonic pipeline inspection

for modelling by linear regression. This was done and the fitted least squares line turned out to be

$$\text{field defect depth} = 4.99 + 0.731 \times \text{laboratory defect depth}.$$

The data together with the least squares line are plotted in Figure 24. The line appears to fit the data pretty well.

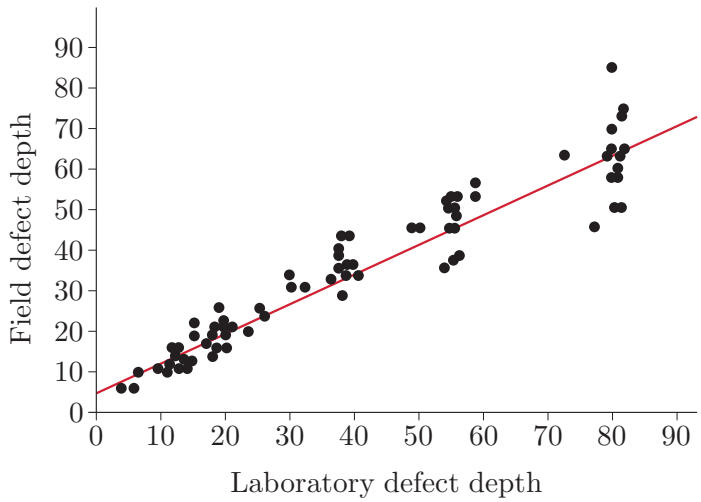


Figure 24 The Trans-Alaska oil pipeline data and least squares line

But what of the wider linear regression model with its assumption of constant, zero mean and constant variance of the random terms W_i? Are these assumptions (collectively Assumption 2) justified by the residual plot provided in Figure 25? If not, why not?

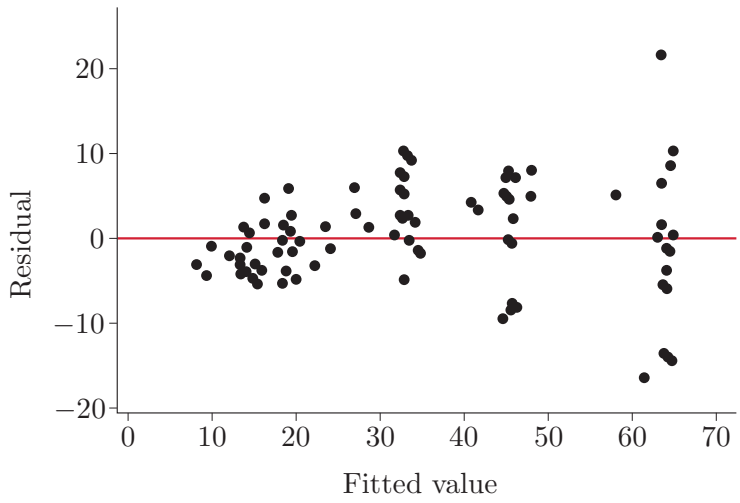


Figure 25 Residual plot for the Trans-Alaska oil pipeline data

A small point to note is that for the linear regression model including both slope and intercept terms (i.e. not regression through the origin), the sum

of the residuals is always zero. If a plot purported to be a residual plot clearly violates this property, then something has gone wrong in producing that residual plot. You can prove this property of residuals for yourself in the next activity.

Activity 15 *Summing the residuals*

The residuals w_i can be written $w_i = y_i - \widehat{\alpha} - \widehat{\beta}\,x_i$ where $\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}$. Use these facts to show that $\sum_{i=1}^{n} w_i = 0$.

3.2 Checking normality of residuals

In order to use the fitted regression model to make inferences, test hypotheses, produce confidence intervals, and so on, it is necessary to assume some distribution for the W_is. The most common assumption to make is the one made in Subsection 2.4: that the random terms are normally distributed. Sometimes other distributions are used – for example, the Poisson distribution or the Bernoulli distribution, where appropriate; you may well come across some of these in further statistical studies. However, for inferential work on the linear regression model in this module, the following assumption will be made.

Using these distributions requires further modifications to the linear regression model, however.

3 The W_is are normally distributed.

If Assumptions 1 to 3 are satisfied, then the W_is are independent normal random variables with zero mean and some variance σ^2. That is, the W_is can be regarded as a random sample from an $N(0, \sigma^2)$ distribution, and the Y_is can be regarded as a random sample from an $N(\alpha + \beta\,x_i, \sigma^2)$ distribution (as in Subsection 2.4).

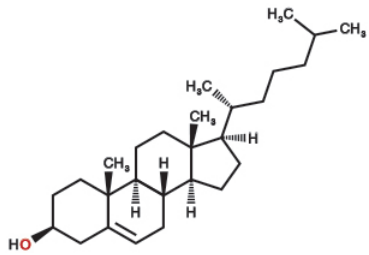
A normal probability plot can be used to check whether it is reasonable to assume that a sample of data comes from a normal distribution (Section 5 of Unit 6). So, if a residual plot has confirmed that a linear regression model is appropriate for the data at hand, in the sense that we can assume that the W_is have zero mean and constant variance, then Assumption 3 can be checked using a normal probability plot of the residuals. This is illustrated in Example 14 and Activity 16 to follow.

Example 14 *Normality of residuals for the cholesterol data*

The cholesterol data introduced in Example 9 were fitted in Example 11 by the least squares regression line

$$\text{total cholesterol} = 1.89 + 0.04 \times \text{age}.$$

In Example 13, it was concluded, on the basis of the residual plot in Figure 22, that it was reasonable to make Assumption 2 (zero mean, constant variance of W_is).



The structure of a cholesterol molecule

In order to now check Assumption 3 (normality of W_is), a normal probability plot of the residuals w_i is shown in Figure 26.

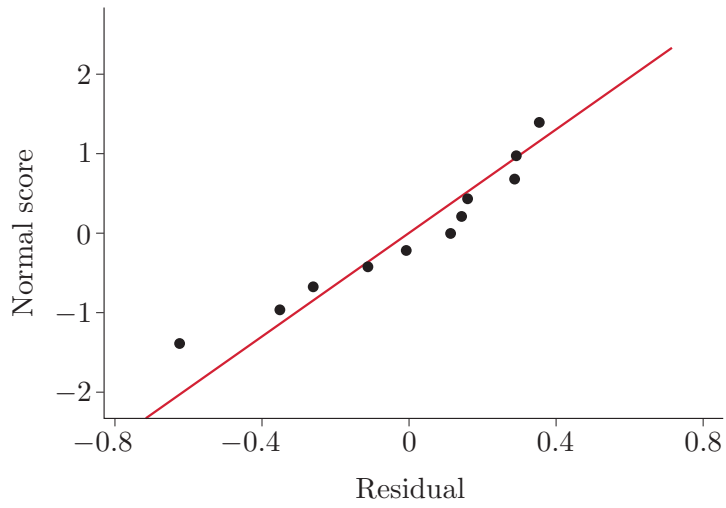


Figure 26 A normal probability plot of the residuals for the cholesterol data

The residuals lie reasonably close to a straight line, so it seems plausible that the random terms come from a normal distribution. That is, it can be argued that Assumption 3 seems to be reasonable for these data. (You might, however, perceive a curve in the probability plot, but with so few data points it seems insufficiently strong to rule out the normality of random terms in the model.)

Overall, the linear regression model with normally distributed random terms appears to be a reasonable one to explain the dependence of total cholesterol on age for patients aged over 40 years with hyperlipoproteinaemia.

Activity 16 *Checking the assumptions for the tapping data*

The tapping data introduced in Activity 12 were fitted there by the least squares regression line

$$\text{taps} = 244.75 + 0.0175 \times \text{caffeine dose}.$$

Here, the response variable is the number of taps per minute and the explanatory variable, caffeine dose, is measured in mg. In order to check the assumptions of the linear regression model, a residual plot and a normal probability plot of residuals are given in Figure 27(a) and Figure 27(b), respectively. Similarly to the scatterplot of the original data in Figure 19, coincident points in the residual plot are shown jittered slightly, in the vertical direction.

Comment on what these plots tell you. Is the linear regression model a good one for these data?

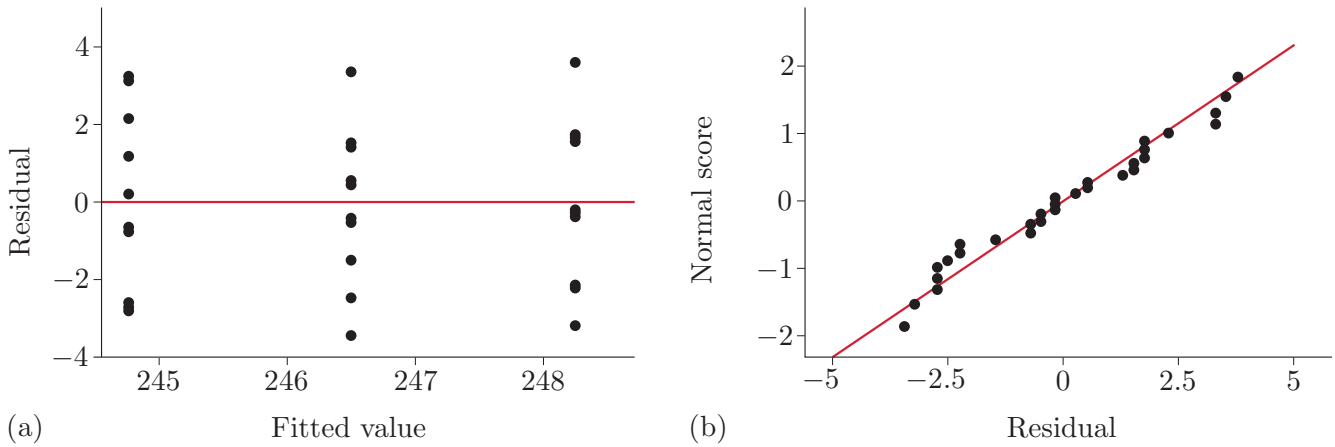


Figure 27 Plotting the residuals for the tapping data: (a) residual plot; (b) normal probability plot

A computer software package is usually used to produce residual plots and normal probability plots of residuals. However, although a computer can do the calculations and draw the plots, it is up to you to interpret the plots and assess whether the assumptions are reasonable! So now go to Computer Book C. There you will use Minitab to fit linear regression models to data and to check the assumptions of the fitted models.

Refer to Chapter 2 of Computer Book C for the rest of the work in this section.

It should be added that all is not lost if the assumptions of the linear regression model are not met. Further regression modelling of the data can still be performed. One way of accounting for failures in the assumptions will be investigated in Unit 12.

Exercise on Section 3

Exercise 2 *Cholesterol and a wider range of ages*

The data given in Table 8 are the plasma levels (in mg/ml) of total cholesterol in 24 adults with hyperlipoproteinaemia. This is the full dataset from which the smaller dataset studied so far in this unit was extracted. The smaller dataset concerned only those individuals aged over 40 years; the full dataset adds 13 other patients who were aged 40 years or under.

Table 8 Cholesterol levels (in mg/ml) and ages (in years)

Age	20	22	22	24	25	28	28	29	30	33	34	36
Cholesterol	1.9	2.1	2.5	2.5	3.0	2.3	2.9	3.3	2.6	3.0	3.2	3.8
Age	40	43	46	48	49	50	52	52	57	57	58	63
Cholesterol	3.2	3.8	3.5	4.2	4.0	3.3	4.0	4.3	4.5	4.1	3.9	4.6

(Source: full dataset from Krzanowski, W.J. (1998) *An Introduction to Statistical Modelling*, London, Arnold, Chapter 3)

For the data concerning the over-40s only, the least squares line

$$\text{total cholesterol} = 1.89 + 0.04 \times \text{age}$$

was fitted in Example 11. In Examples 13 and 14, Assumptions 2 and 3 were checked. In Example 14, it was concluded that: 'Overall, the linear regression model with normally distributed random terms appears to be a reasonable one to explain the dependence of total cholesterol on age for patients aged over 40 years with hyperlipoproteinaemia.' This exercise concerns the question of whether or not the linear regression model with normally distributed random terms remains appropriate to explain the dependence of total cholesterol on age for all adult patients with hyperlipoproteinaemia.

(a) A scatterplot of the full dataset is provided in Figure 28 along with the least squares line fitted to these data. On the basis of this plot, does there seem to be a case for use of the linear regression model?

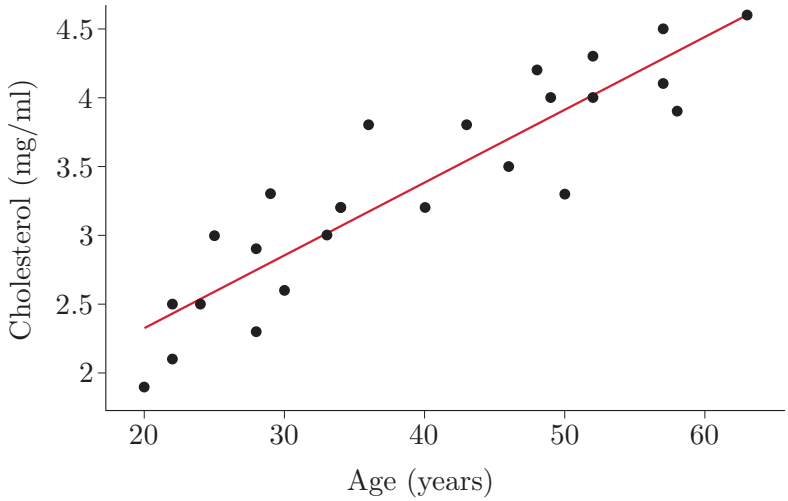


Figure 28 Total cholesterol against age, and the least squares line, for the full dataset

(b) A residual plot for the data in Figure 28 is provided in Figure 29. Does Assumption 2, that the random terms have constant, zero mean and constant variance, seem to be satisfied?

(c) A normal probability plot of the residuals for the data in Figure 28 is provided in Figure 30. Does Assumption 3, that the random terms are normally distributed, seem to be satisfied?

(d) The least squares line in Figure 28 has the formula

$$\text{total cholesterol} = 1.28 + 0.05 \times \text{age}.$$

How, numerically, do the least squares lines fitted to the different versions of the dataset differ? Sketch the lines on a graph as functions of age (in years): for the whole dataset on the range 20 to 63, and for the cut-down dataset on the range 43 to 63. Does the line appear to have changed much since the younger patients' data were included?

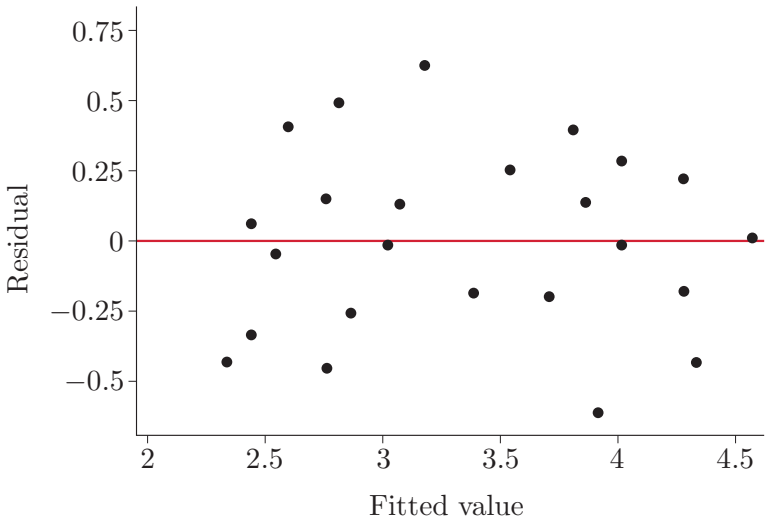


Figure 29 A residual plot for the full cholesterol dataset

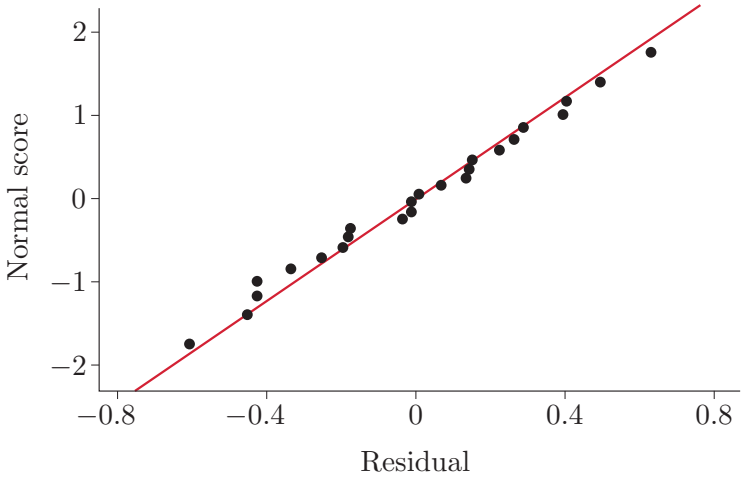


Figure 30 A normal probability plot of the residuals for the full cholesterol dataset

4 Sampling properties and statistical inference

In Section 2, you learned how to calculate the least squares estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of the parameters α and β of the regression line. However, a repeated experiment would almost certainly result in different responses and hence different estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of α and β. The estimates $\widehat{\alpha}$ and $\widehat{\beta}$ vary from one experiment to the next, so they are observations of random

Results similar to the results in this section are available for the estimator $\widehat{\gamma}$ of the constrained model. They will not be given in this module.

variables. It is standard to use the same notation $\widehat{\alpha}$ and $\widehat{\beta}$ for these random variables as for the individual estimates. These random variables are called the **least squares estimators** of α and β. The formula for the least squares estimator of β is

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}.$$

In this case, remembering that the explanatory variable is regarded as non-random in regression, S_{xy} is the random variable given by

$$S_{xy} = \sum (x_i - \overline{x})(Y_i - \overline{Y}).$$

Here, $\overline{Y}$ denotes the mean of the random variables $Y_1, Y_2, \ldots, Y_n$, that is, $\overline{Y} = \sum Y_i/n$.

Notice that the same notation S_{xy} is used for this sum, which involves random variables, as was used in Subsection 2.3 for the sum of products of deviations $\sum (x_i - \overline{x})(y_i - \overline{y})$. This duality of notation is standard.

Was the Roman who had the job of estimating the least number of square tiles required for this mosaic the least squares estimator?

The least squares estimator of α is given by

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\overline{x}.$$

The sampling distributions and some properties of the estimators $\widehat{\alpha}$ and $\widehat{\beta}$ will be given in Subsection 4.1. There too you will find the estimator of σ^2, the final unknown parameter in linear regression (which we haven't addressed yet), and its sampling properties. The properties of Subsection 4.1 are used to provide a particular hypothesis test in Subsection 4.2. Other aspects of statistical inference in regression are reviewed briefly in Subsection 4.3. Note that it is not the intention of this section to provide anything like an exhaustive list of results, or to offer illustrations of many of the questions that might arise in a regression context.

4.1 The sampling distributions of the estimators

By assuming only that the random terms W_i are independent with zero mean and variance σ^2, it is a cumbersome task, although not a difficult one, to show that the two estimators $\widehat{\alpha}$ and $\widehat{\beta}$ are unbiased estimators of α and β, respectively, that is,

$$E(\widehat{\alpha}) = \alpha, \quad E(\widehat{\beta}) = \beta.$$

Furthermore, it can be shown that the variances of the estimators are given by the formulas

$$V(\widehat{\alpha}) = \left(\frac{\overline{x}^2}{S_{xx}} + \frac{1}{n}\right)\sigma^2, \quad V(\widehat{\beta}) = \frac{\sigma^2}{S_{xx}}.$$

The calculations involved are not entirely straightforward.

Notice that if the x values are widely dispersed, that is, if the sum of squared deviations S_{xx} is large, then the variances of both estimators are smaller than if the x values are close together (so that S_{xx} is small). This makes sense because there is more information about the parameters of the regression line when the explanatory variable takes on a wide range of values than there is if it is confined to a narrow range. As the estimators are unbiased for any values of the explanatory variable, it is possible to choose values $x_1, x_2, \ldots, x_n$ such that the variances of the estimators $\widehat{\alpha}$ and

$\widehat{\beta}$ are small, and thus the precision of the results is improved. In particular, it is helpful to obtain data for a wide spread of x values rather than to concentrate on only a narrow range. This is important when designing a statistical experiment.

Until now, the parameter σ^2 has been treated as if its value was known. In general, though, its value is not known, and it has to be replaced with an estimate. Write $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}\, x_i$ for the fitted values in random variable form. The unbiased estimator for σ^2 that is used is

$$S^2 = \frac{\sum (Y_i - \widehat{Y}_i)^2}{n-2}. \tag{10}$$

Thus the numerator in S^2 is simply the residual sum of squares for the least squares line: $\sum (Y_i - \widehat{Y}_i)^2 = \sum (Y_i - (\widehat{\alpha} + \widehat{\beta}\, x_i))^2$. Convention then dictates that the unbiased estimate of σ^2 is used rather than the maximum likelihood estimate of σ^2 under the assumption of normality for the random terms W_i. As for estimation of the variance in a single sample (Unit 7), maximum likelihood estimation of the variance, σ^2, in the regression model would result in a divisor of n in Equation (10), not $n-2$. But it turns out that as unbiasedness is achieved in the one-sample case by subtracting 1 from n because there is one other parameter – the mean, μ – also being estimated, so in regression unbiasedness is achieved by subtracting 2 from n, there being two other parameters, α and β, also being estimated.

The results given so far in this subsection hold whatever form is taken by the distribution of the random terms. The results that are given in the box that follows hold when the random terms W_i are assumed to be normally distributed. This assumption is made throughout the rest of this section. The results are stated without proof.

Distributions of the least squares estimators

Assuming that, independently, $W_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$, it can be shown that

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right), \quad \frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2),$$

and these two random variables are independent. These results can be combined to give the following result, which proves to be very useful for statistical inference when σ^2 is unknown (which is usually the case):

$$\frac{\widehat{\beta} - \beta}{S/\sqrt{S_{xx}}} \sim t(n-2). \tag{11}$$

4.2 Testing whether a relationship exists

When $\beta = 0$, the linear regression model simplifies to

$$Y_i = \alpha + W_i.$$

In this case, the value of Y_i does not depend on the value of x_i – that is, the response variable and the explanatory variable are unrelated: it is often said that no regression relationship exists. Equivalently, the regression line is flat; see Figure 31.

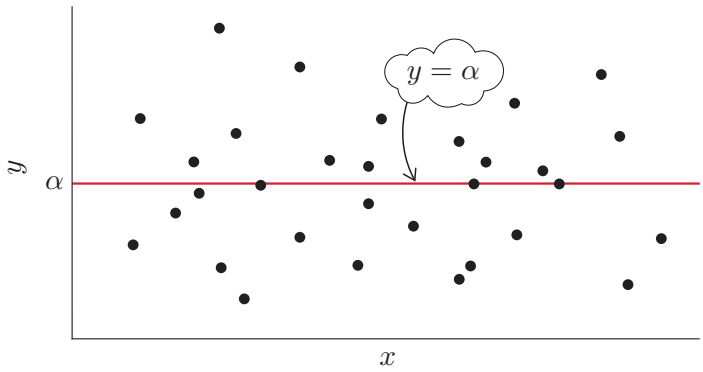


Figure 31 Artificial data from the regression $Y_i = \alpha + W_i$

In fact, if $\beta = 0$, the responses are just a random sample from a normal distribution with mean α and variance σ^2. So researchers are often interested in testing whether the slope parameter β in the linear regression model is zero, that is, testing the null hypothesis $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. This can be done using Distributional Result (11).

Example 15 *Does caffeine have an effect on tapping performance?*

Tap dancers give a different type of tapping performance; do they need caffeine?

In Activity 12, you found that the equation of the least squares regression line for the finger-tapping data is

$$y = 244.75 + 0.0175\,x,$$

that is,

$$\text{taps} = 244.75 + 0.0175 \times \text{caffeine dose},$$

where taps are counted per minute and the caffeine dose is measured in mg.

An interesting question to consider is 'Does caffeine really have any effect on tapping frequency?' When $\beta = 0$, there is no relationship between the explanatory variable and the response variable. So one approach to answering the question is to carry out a two-sided test of the null hypothesis

$$H_0 : \beta = 0,$$

against

$$H_1 : \beta \neq 0.$$

Certainly, the estimated value $\widehat{\beta} = 0.0175$ seems to be quite a small number in absolute terms, but it needs to be assessed in the context of the overall variation in the data.

The following summary statistics are required in order to perform the test:

$$n = 30, \quad \sum (y_i - \widehat{y}_i)^2 = 134.25, \quad S_{xx} = 200\,000.$$

First, we need an estimate of σ^2. Using the sample version of Equation (10), the estimate of σ^2 is given by

$$s^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2} = \frac{134.25}{28} \simeq 4.7946.$$

$n - 2 = 30 - 2 = 28$

Using Distributional Result (11), the null distribution of the test statistic is $t(n-2) = t(28)$. Then, when H_0 is true, $\beta = 0$ and Distributional Result (11) means that the observed value of the test statistic is

$$\frac{\widehat{\beta} - 0}{s/\sqrt{S_{xx}}} = \frac{0.0175 - 0}{\sqrt{4.7946}/\sqrt{200\,000}} \simeq 3.574.$$

From the table of quantiles of the t-distribution in the Handbook, the 0.999-quantile of $t(28)$ is 3.408, so the p-value for this two-sided test is less than 0.002. This p-value is extremely small, so the null hypothesis $H_0 : \beta = 0$ is rejected. That is, despite the seemingly small value of $\widehat{\beta}$, there is strong evidence against the hypothesis that caffeine dose has no effect on tapping performance.

A computer gave 0.0013 for the p-value.

Activity 17 *Does cholesterol really change with age for older ages?*

Consider again the cholesterol data for the 11 patients aged over 40 that have been much studied in previous sections. The equation of the least squares line for the cholesterol data, which was found in Example 11, is

$$y = 1.89 + 0.04\,x,$$

where y represents the total cholesterol in mg/ml, and x represents the patient's age in years. As in Example 15, the value of $\widehat{\beta} = 0.04$ seems rather close to zero, but the presence or otherwise of a non-zero slope needs to be tested taking into account the variability in the data. In order to test $H_0 : \beta = 0$ you will need the following summary statistics:

$$n = 11, \quad \sum(y_i - \widehat{y}_i)^2 = 0.952, \quad S_{xx} = 352.18.$$

(a) What is the value of s^2, the estimate of σ^2?

(b) Using a two-sided alternative hypothesis, test whether there is really no relationship between cholesterol and age.

We are not considering the larger cholesterol dataset of Exercise 2 here.

4.3 Some brief intervals

The results of Subsection 4.1 also allow us to provide interval estimators of a number of quantities associated with the linear regression model. Since these formulas are rather similar to the t-intervals of Section 4 of Unit 8, to avoid too much repetition and tedium, we do little more than list the formulas here, along with some brief comments and a single activity.

A more enjoyable sort of interval, at a cultural event

Associated with the test of $H_0 : \beta = 0$ that we have just considered in Subsection 4.2 is a confidence interval for the value of the slope parameter, β. This too arises from manipulation of Distributional Result (11). Throughout this subsection, write s for the estimated standard deviation, which is the square root of the estimated variance

$$s^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2}.$$

A $100(1-\alpha)\%$ confidence interval for the slope parameter β

A $100(1-\alpha)\%$ **confidence interval for the slope parameter** β of the regression line is given by

$$\left(\widehat{\beta} - t\,\frac{s}{\sqrt{S_{xx}}}, \widehat{\beta} + t\,\frac{s}{\sqrt{S_{xx}}}\right),$$

where t is the $(1-(\alpha/2))$-quantile of $t(n-2)$.

In a linear regression model, the mean of the response Y_i is $\alpha + \beta x_i$, that is, it depends on the value of the explanatory variable x_i. So a confidence interval for the *mean response* will also depend on the value of the explanatory variable, and therefore varies for different values of x. Suppose we are interested in the mean response *for a given value* x_0 *of* x, that is, $\alpha + \beta x_0$. The natural point estimator of $\alpha + \beta x_0$ is

$$\widehat{\alpha} + \widehat{\beta} x_0,$$

which turns out to be an unbiased estimator of $\alpha + \beta x_0$. The corresponding interval estimator of $\alpha + \beta x_0$ is given next.

A $100(1-\alpha)\%$ confidence interval for the mean response

A $100(1-\alpha)\%$ **confidence interval for the mean response** of Y_0, $\alpha + \beta x_0$, is given by

$$\left(\widehat{\alpha} + \widehat{\beta} x_0 - t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}} + \frac{1}{n}},\ \widehat{\alpha} + \widehat{\beta} x_0 + t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}} + \frac{1}{n}}\right), \tag{12}$$

where t is the $(1-(\alpha/2))$-quantile of $t(n-2)$.

$t\,s$ is not a new symbol, just the product of the quantile, t, and the estimated standard deviation, s.

Suppose, finally, that there is interest in predicting the value of the response Y_0 for a given value x_0 of the explanatory variable. Then the obvious *predictor* of Y_0 is

$$\widehat{y}_0 = \widehat{\alpha} + \widehat{\beta} x_0.$$

Of course, you saw and used this quantity in Subsection 2.3.

This is precisely the same as the point estimator of the mean response at x_0 given above, that is, the point predictor and the point estimator of the mean response are the same. There is, however, a difference between the confidence interval for the mean response (given above) and the confidence

interval for the predicted response – the interval predictor, or *prediction interval* (given below). This is because there are *two* sources of variation in connection with prediction.

First, there is the variability associated with the least squares line that estimates the mean of the response, this variability being used in forming the confidence interval for the mean response above. In addition, though, for a given value x_0 of the explanatory variable, the response is a random variable:

$$Y_0 = \alpha + \beta x_0 + W_0.$$

So, as well as the variability associated with estimating the line at x_0, $\alpha + \beta x_0$, there is the added variation due to the random term, W_0. (Because of W_0, even if the true values of α and β were known, it would still not be possible to predict Y_0 exactly!) In a prediction interval, we have to allow for variation coming from the random term W_0 in addition to variation coming from the estimation of the predictor.

A $100(1-\alpha)\%$ prediction interval for the response

A $100(1-\alpha)\%$ **prediction interval** for the response Y_0 when $x = x_0$ is given by

$$\left(\widehat{\alpha} + \widehat{\beta} x_0 - t\, s\sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n} + 1},\ \widehat{\alpha} + \widehat{\beta} x_0 + t\, s\sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n} + 1}\right), \tag{13}$$

where t is the $(1 - (\alpha/2))$-quantile of $t(n-2)$.

As you can see, prediction intervals are calculated in a similar way to confidence intervals. A prediction is made; and lower and upper limits are calculated, allowing for error in the prediction. However, a prediction interval has to allow for more variation than a confidence interval does. So prediction intervals are wider than confidence intervals. (There is an extra term of '1' added beneath the square root signs in Interval (13), compared with Interval (12).)

Note that the quantiles of the t-distribution required in the intervals in the boxes above are the same in all three cases.

Activity 18 *Intervals from the finger-tapping data*

In Activity 12, you fitted the following model to the finger-tapping data:

$$y = 244.75 + 0.0175\, x.$$

Suppose that the researchers were interested in the frequency of finger-tapping when an $x_0 = 40$ mg dose of caffeine was administered.

The estimated mean tapping frequency in response to a dose of $x_0 = 40$ mg of caffeine is

$$244.75 + 0.0175 \times 40 = 245.45$$

taps per minute.

The following summary statistics for these data will be needed:

$$n = 30, \quad \overline{x} = 100, \quad S_{xx} = 200\,000.$$

Also, the estimate $s^2 = 4.7946$ was found in Example 15.

(a) You will be concerned with 95% intervals, for the mean response and for the predicted response, in this question. You will require the value of t to be an appropriate quantile of an appropriate t-distribution. Find the value of t.

(b) Obtain a 95% confidence interval for the mean tapping frequency of individuals receiving a dose of $x_0 = 40$ mg of caffeine.

(c) Obtain a 95% prediction interval for the tapping frequency of a particular individual receiving a dose of $x_0 = 40$ mg of caffeine.

Hypothesis tests, confidence intervals, and so on, in relation to linear regression models can be calculated using Minitab. However, we will not spend time on this at this juncture.

Exercise on Section 4

Exercise 3 *A little inference on the full cholesterol dataset*

Consider again the cholesterol data for the full set of 24 hyperlipoproteinaemia patients, with ages from 20 years upwards, that was considered in Exercise 2. The equation of the least squares line for these data, which was found in Exercise 2, is

$$y = 1.28 + 0.05\,x,$$

where y represents the total cholesterol in mg/ml, and x represents the patient's age in years. The summary statistics needed to answer this question are as follows:

$$n = 24, \quad \overline{x} = 39.42, \quad S_{xx} = 4139.77, \quad \sum (y_i - \widehat{y}_i)^2 \simeq 2.455.$$

(a) Using a two-sided alternative hypothesis, test whether there is actually no regression relationship between age and cholesterol over the wide range of ages in the dataset.

(b) Calculate a 90% prediction interval for the value of total cholesterol for a hyperlipoproteinaemia patient of age 35 years.

(c) The prediction interval that you calculated in part (b) is actually rather wide. By reference to the data in Table 8, explain why this interval is still useful, despite its width.

5 Multiple regression

In the final section of this unit, we consider the situation in which there is more than one explanatory variable. You have already seen an example of such a scenario in Example 4 in which the response variable was the strength of timber beams and there were two possible explanatory variables: specific gravity and moisture content. In such situations, the linear regression model can be extended to incorporate more than one explanatory variable into the model; this is called **multiple regression**. Multiple regression is an important statistical method which is widely used by practising statisticians; this section provides just a brief introduction to the topic.

There continues to be a single response variable.

The section begins in Subsection 5.1 by extending the linear regression model to incorporate more than one explanatory variable into the model. The interpretation of the model parameters is not the same in multiple regression as it is in linear regression with one explanatory variable; this is discussed in Subsection 5.2. Checking the model assumptions in multiple regression is the subject of Subsection 5.3. Finally, multiple regression in Minitab is the subject of Subsection 5.4 (and its associated chapter in Computer Book C).

5.1 Extending the linear regression model

We start this subsection with an example of a problem in which there are two potential explanatory variables.

Example 16 *Student satisfaction*

Official statistics concerning UK universities are collected annually. This example considers three of the variables on which data were collected for 2015. The example focuses on data for the 24 UK universities known collectively as Russell Group universities. (This group represents some of the leading UK universities.)

The National Student Survey (NSS) surveys final-year UK undergraduate students. Surveyed students score how satisfied they are with the quality of various aspects of the teaching that they received, using a scale from 0 to 5 (where 5 represents the highest level of satisfaction). The first variable in this example is an overall student satisfaction score for each university: this is an average of the individual student satisfaction scores within that university for 2015. Although individual scores are discrete, with range $\{0, 1, 2, 3, 4, 5\}$, scores for whole universities have quite a narrow range of essentially continuous values, and for these data range from 3.89 to 4.18. Student satisfaction is our response variable Y.

As before, an explanatory variable is regarded as non-random for the purposes of regression modelling, and so is denoted by a lower-case letter.

The second variable that we'll consider here is the student–staff ratio. For each university, this is the total number of undergraduate and postgraduate students for 2015 divided by the number of academic staff for that year. The data for this variable were collected by the Higher Education Statistics Agency (HESA). Denote this explanatory variable by x_1.

The final variable that we'll consider is 'academic services spend'. These data were also collected by HESA and use the average expenditure over three academic financial years (2012/13, 2013/14 and 2014/15) to allow for uneven patterns of expenditure. Academic services spend was calculated as being the expenditure, in pounds, on library and computing facilities (staff, books, journals, computer hardware and software, but not buildings), museums, galleries and observatories, divided by the number of full-time equivalent students in the latest academic year. Denote this second explanatory variable by x_2.

The data are given in Table 9.

Table 9 Student satisfaction in Russell Group universities, 2015

Student satisfaction	Student–staff ratio	Academic services spend (£)
4.08	14.0	1883
3.96	13.8	1453
4.17	11.0	2628
4.09	13.7	1245
4.14	14.7	1542
3.93	12.0	1436
4.18	15.8	1689
4.12	14.5	1702
4.15	11.1	2309
3.91	11.7	1499
4.16	13.5	1970
4.01	11.8	1685
3.89	11.6	2105
4.05	13.4	1513
4.14	15.5	1352
4.06	13.2	1594
4.17	10.5	2700
4.12	11.9	1548
4.13	15.1	1266
4.14	14.6	1540
4.08	12.0	1694
3.95	10.2	2212
4.09	12.5	1826
4.14	14.5	1441

Figure 32 shows a scatterplot of student satisfaction (y) against the first explanatory variable, student–staff ratio (x_1). Included on the plot is the least squares line, calculated using the method given in Subsection 2.3 as

$$y = 3.797 + 0.0215\,x_1.$$

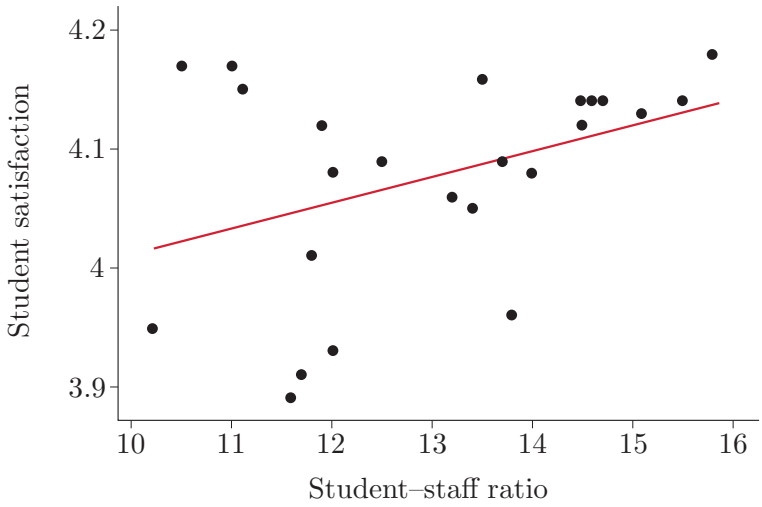


Figure 32 Student satisfaction against student–staff ratio, and the least squares line

From Figure 32, student satisfaction generally increases as the student–staff ratio increases. This is reflected in the positive slope parameter in the least squares line. You might have expected the student satisfaction to decrease as the student–staff ratio increases; and indeed this is the case when all UK universities are considered. The observed increase when considering only Russell Group universities therefore seems to be specific to these universities. (For instance, it's possible that the student–staff ratio could be a reflection of the popularity and quality of some Russell Group universities, which can attract large numbers of applicants.)

Now consider the second explanatory variable, academic services spend (x_2). Figure 33 shows a scatterplot of student satisfaction (y) against x_2, together with the least squares line

$$y = 4.019 + 0.000034\,x_2.$$

Although the relationship between this second explanatory variable and the response variable appears weak, what relationship there is appears to be positive, indicating that student satisfaction increases (slightly) as academic services spend increases.

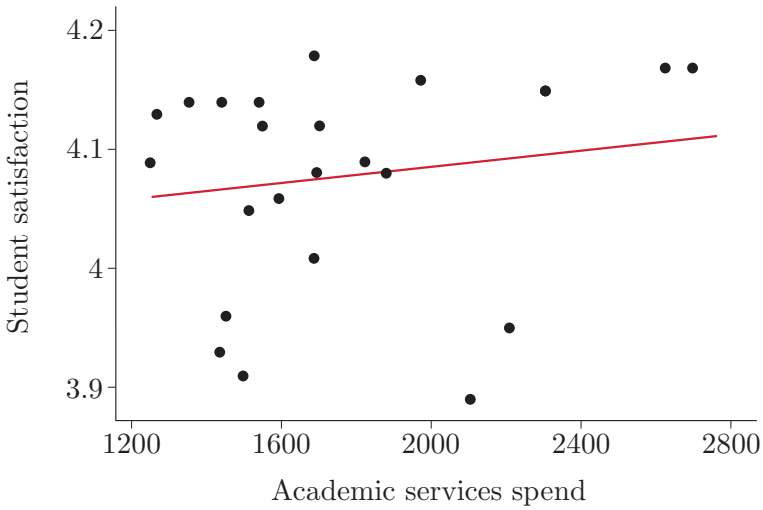


Figure 33 Student satisfaction against academic services spend, and the least squares line

When carrying out a two-sided test of the null hypothesis $H_0 : \beta = 0$ in the regression model using x_1, the associated p-value for the slope is 0.057, and when carrying out the same test in the regression model using x_2, the p-value for the slope is 0.486. So, in actual fact, there is only weak evidence that student–staff ratio on its own affects student satisfaction, and there is little or no evidence that academic services spend on its own affects student satisfaction. Is it possible, however, that student–staff ratio and academic services spend can act *together* to affect student satisfaction in Russell Group universities in a rather more substantial way? You will see that, by using a regression model which uses *both* explanatory variables at the same time, this is indeed the case!

These p-values were obtained from Minitab.

In Example 16, we had a response variable Y with two explanatory variables x_1 and x_2. Denote the ith observations of x_1 and x_2 (associated with y_i) by x_{i1} and x_{i2}, respectively. Now, the linear regression model with one explanatory variable can be written as

$$Y_i = \alpha + \beta\,x_i + W_i,$$

where the W_is are independent random variables with zero mean and constant variance. This model can be extended to incorporate two

explanatory variables thus:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + W_i.$$

Once again the W_is are independent random variables with zero mean and constant variance. In fact, here we will consider only the case in which the W_is are additionally assumed to be normally distributed.

This model can be naturally extended to the situation in which there are p explanatory variables $x_1, x_2, \ldots, x_p$, with the ith observation of the jth explanatory variable being denoted by x_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$. The *multiple linear regression model*, or the *multiple regression model* for short, is then defined as follows.

The multiple linear regression model

If data $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$, $i = 1, 2, \ldots, n$, comprise observations on p explanatory variables $x_1, x_2, \ldots, x_p$ and a response variable Y, then the **multiple linear regression model** can be written

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + W_i, \qquad (14)$$

$i = 1, 2, \ldots, n$. The terms W_i are independent normal random variables with zero mean and constant variance.

Note that we are considering only the situation in which the relationship between Y and $x_1, x_2, \ldots, x_p$ is linear and the random terms W_i, $i = 1, 2, \ldots, n$, come from independent normal distributions.

Activity 19 *Formulating a model*

A zoologist would like to use a multiple linear regression model to model the heights of young giraffes using their weight and age (in days). Write down the form of the zoologist's multiple regression model.

5.2 Interpreting regression coefficients

So, how are the parameters of the multiple linear regression given in Equation (14) to be interpreted?

Well, first, the parameter α can still be considered as an intercept parameter because it is the value of the linear trend $\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ when all of $x_{i1}, x_{i2}, \ldots, x_{ip}$ are zero.

The parameters $\beta_1, \beta_2, \ldots, \beta_p$, however, are now **partial regression coefficients**. They are usually just called **regression coefficients** for short, but the word 'partial' is important in reminding us of their meaning. In the multiple regression model, the parameter β_1 measures the effect of increasing x_1 by one unit when $x_2, x_3, \ldots, x_p$ *are all kept fixed*; β_2 measures the effect of increasing x_2 by one unit when $x_1, x_3, \ldots, x_p$ *are all kept fixed*;

and so on. As such, the regression coefficients are not the same as the slope parameter in the simple linear regression model with one explanatory variable, and they do not have the same interpretation: a regression coefficient represents the 'partial' effect of the associated explanatory variable given the values of the other explanatory variables, while the slope parameter represents the effect of a single explanatory variable on its own.

In Section 2, the method of least squares was used to estimate the parameters in the linear regression model with a single explanatory variable. You also saw, in Subsection 2.4, that when the random terms W_i are normally distributed, maximum likelihood estimates of the parameters of the linear regression model are the same as those obtained via the method of least squares. Parameter estimation when there is more than one explanatory variable follows the same ideas, but is a bit more complicated due to the increased number of parameters. Because of this, in M248 we will simply use Minitab for estimating the intercept parameter and regression coefficients. (Details of estimation are left to modules at a higher level.)

The fitted multiple regression model

If $\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p$ are estimates of the intercept and regression coefficients in a multiple regression model, then the fitted multiple regression model is

$$y = \widehat{\alpha} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_p x_p.$$

Example 17 *Interpreting the fitted model*

Consider once again Example 16 in which we had the response variable student satisfaction (Y) and two explanatory variables: student–staff ratio (x_1) and academic services spend (x_2). The fitted multiple regression model obtained by using Minitab with the data of Table 9 is

$$y = 3.157 + 0.0484\, x_1 + 0.000166\, x_2.$$

The interpretation of the regression coefficients is as follows.

- If the value of the student–staff ratio (x_1) increases by one unit (that is, by one more student per staff member), and the value of the academic services spend (x_2) remains fixed, then the student satisfaction score (y) would be expected to increase by 0.0484.
- If the value of the academic services spend (x_2) increases by one unit (that is, by one pound per student), and the value of the student–staff ratio (x_1) remains fixed, then the student satisfaction score (y) would be expected to increase by 0.000166.

Notice that the regression coefficients are not the same values as the corresponding slope parameters for x_1 and x_2 in the separate least squares lines in Example 16. In each least squares line, the slope parameter represents the effect of the individual explanatory variable on the response

variable. However, when both explanatory variables are in the model, as they are here, each regression coefficient represents the partial effect that the individual explanatory variable has on the response variable, given the other explanatory variable.

In Example 16, you saw that there wasn't very much evidence to suggest that either of the slope parameters in the separate linear regression models for modelling student satisfaction was non-zero. So, treated individually, it looked likely that neither the first explanatory variable, student–staff ratio, nor the second explanatory variable, academic services spend, was going to be very useful for modelling student satisfaction. How do we know that the regression model with both explanatory variables given in Example 17 is any better? The answer lies in carrying out a two-sided test of the null hypothesis

$$H_0 : \beta_1 = 0,$$

and a second two-sided test of the null hypothesis

$$H_0 : \beta_2 = 0,$$

within the context of the multiple linear regression model with two explanatory variables. These tests are similar in construction to the two-sided test in the linear regression model with one explanatory variable of the null hypothesis $H_0 : \beta = 0$. But the pair of multiple regression tests yields different results from the pair of simple linear regression tests, for reasons again associated with the partial nature of the regression coefficients in the multiple regression context. You will be spared the details of these tests here, and instead we will just consider the resulting p-values which are routinely provided by Minitab when fitting a multiple regression model.

These p-values are used to assess the evidence against the null hypothesis in the usual way.

Activity 20 *Are the student satisfaction regression coefficients zero?*

The fitted multiple regression model for response variable student satisfaction (Y) and two explanatory variables, student–staff ratio (x_1) and academic services spend (x_2), obtained from Minitab is

$$y = 3.157 + 0.0484\,x_1 + 0.000166\,x_2.$$

The p-value for the two-sided test of the null hypothesis $H_0 : \beta_1 = 0$ is calculated in Minitab to be 0.000, and the p-value for the two-sided test of the null hypothesis $H_0 : \beta_2 = 0$ is calculated in Minitab to be 0.002. What do you conclude about the regression coefficients for x_1 and x_2? Hence, what do you conclude about how student–staff ratio and academic services spend affect student satisfaction in Russell Group universities?

As was the case with a single explanatory variable, the fitted multiple linear regression line can be used for prediction. Point prediction is particularly straightforward and is illustrated in the context of student satisfaction scores in Example 18.

Example 18 *Predicting student satisfaction*

Suppose that another university felt that the Russell Group fitted line applied equally well to it. In 2015, this university had a student–staff ratio of 14.5 students per staff member and an academic services spend of £1441 per student. The fitted line predicts a student satisfaction score of

$$3.157 + 0.0484 \times 14.5 + 0.000166 \times 1441 \simeq 4.10.$$

(Perhaps this university was right: its actual 2015 student satisfaction score turns out to have been 4.14.)

Activity 21 *Strength of timber beams*

Example 4 introduced a dataset involving timber beams. The response variable Y is the strength of a timber beam, and there are two explanatory variables, specific gravity (x_1) and moisture content (x_2). Scatterplots of y against x_1 and of y against x_2 were given in Figure 5. These suggested that there may be an increasing linear relationship between y and x_1, but a weaker, decreasing, relationship between y and x_2. We can use multiple regression to investigate whether specific gravity and moisture content together affect the strength of timber beams.

The fitted multiple regression model for this dataset, obtained from Minitab, is

$$y = 10.29 + 8.50\,x_1 - 0.265\,x_2.$$

The p-value for the two-sided test of the null hypothesis $H_0 : \beta_1 = 0$ is 0.002, while that for the two-sided test of the null hypothesis $H_0 : \beta_2 = 0$ is 0.069.

(a) Interpret the regression coefficients.

(b) Do the data suggest that both x_1 and x_2 together influence the strength of timber beams?

(c) Using the fitted multiple regression model, predict the strength of a timber beam with specific gravity 0.5 and moisture content 10.

In the next activity you will consider a dataset in which there are more than two explanatory variables.

Activity 22 *Gross domestic product*

The average level of income per person varies widely across countries and changes over time as some countries decline and others grow. Economists are interested in the question: 'Why do some countries grow faster than others?' In this activity, a multiple regression model is used to investigate this question. Economic data for 128 countries are available. The response

variable, Y, is the rate of growth, specifically the rate of change between 2000 and 2010 of the gross domestic product (GDP) per head, where the GDP is the total output produced in the country in one year per person. In the dataset, the growth is given as a decimal rather than as a percentage.

There are three explanatory variables, x_1, x_2 and x_3.

There will be consideration of transformations of variables within regression models in Unit 12; just take this logarithmic transformation as providing a helpful scale for this variable in this case.

- x_1 is a measure of the output (GDP) per head in 2000, the initial year of the period; more specifically, it happens to be the logarithm of the GDP per head, where GDP has been translated to the value in US dollars from 2005. Differences in GDP per head are related to differences in the level of technology used. Countries that are more technologically advanced tend to have high levels of GDP per head, while countries that tend to use older and less efficient technology tend to have low levels of GDP per head. Since it is much more difficult and expensive to generate technological innovation than to copy existing technology, it should be easier for poorer countries to grow faster than richer countries by copying better existing technology and therefore improving their efficiency. In turn, this means that countries with low initial GDP per head in 2000 have greater scope for growing and therefore catching up with richer countries.

- x_2 is the share of gross fixed capital formation in GDP in the ten-year period. This is a percentage. Gross fixed capital formation is the investment in new plants, machinery and equipment that is necessary to produce more output (goods and services) and is considered by economists to be a key engine of growth. Intuitively, the argument runs as follows. The output produced can consist of either consumer goods which are used up, such as a loaf of bread, or capital goods which are used as inputs to produce new output in the future, such as a new milling machine. Countries that invest more by producing a greater share of capital goods increase their stock of capital available for production of future output, so they should grow faster than countries that focus more on consumption. So a high share of gross fixed capital formation in GDP should be associated with higher growth.
- x_3 is the total enrolment in secondary schools. This too is a percentage, in this case of the population aged 15 or over. The total enrolment in secondary schools is a measure of human capital, the level of education of the workforce, which is associated with higher productivity and therefore faster economic growth.

The fitted multiple regression model for this dataset, obtained from Minitab, turns out to be

$$y = 0.312 - 0.0923\,x_1 + 0.02425\,x_2 + 0.00493\,x_3.$$

The p-value for each individual two-sided test of the null hypothesis $H_0 : \beta_j = 0$, for $j = 1, 2, 3$, is reported by Minitab to be 0.000.

(a) Explain why this analysis suggests that all three explanatory variables together influence the rate of growth of GDP.

(b) Interpret each of the regression coefficients.

(c) Predict what the rate of growth between 2000 and 2010 would have been for a fictional South Asian country whose 'logged' output per head in 2000 (in the appropriate units) was 6, whose gross fixed capital formation share was 25%, and whose total enrolment in secondary schools was 40%.

5.3 Checking the assumptions

An essential part of any regression analysis is to check the model assumptions. For the multiple linear regression model we have the following assumptions.

1 The random terms W_i are independent.

2 The W_is have zero mean and constant variance.

3 The W_is are normally distributed.

You might be thinking that these three assumptions look very familiar, and you would be right! We have exactly the same assumptions for the multiple linear regression model that we had for the linear regression model with one explanatory variable. As such, these assumptions can be checked in exactly the same way. (As for simple linear regression, although Assumption 1, the independence of the W_is, can be checked, we will not do so in M248.)

In much the same way as for linear regression with one explanatory variable, the fitted multiple regression model can be used to calculate the

fitted value of the response, $\widehat{y}_i$, given values of the explanatory variables $x_{i1}, x_{i2}, \ldots, x_{ip}$. The formula is

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip}.$$

The fitted values can then be used to calculate the residuals

$$w_i = y_i - \widehat{y}_i.$$

This is illustrated in the next example.

Example 19 *Calculating a fitted value and residual*

From Example 17, the fitted multiple regression model for response variable student satisfaction (Y) and two explanatory variables, student–staff ratio (x_1) and academic services spend (x_2), obtained from Minitab is

$$y = 3.157 + 0.0484\, x_1 + 0.000166\, x_2.$$

A very predictable university?!

Out of the 24 Russell Group universities in the sample, the University of Liverpool achieved a student satisfaction score of 4.01. For this university, the student–staff ratio was 11.8, while the academic services spend was £1685 per student. The University of Liverpool is the 12th university listed in the sample. The fitted value of student satisfaction for Liverpool is then

$$\widehat{y}_{12} = 3.157 + 0.0484 \times 11.8 + 0.000166 \times 1685 = 4.0078 \simeq 4.01.$$

The associated residual is therefore

$$w_{12} = 4.01 - 4.01 = 0.$$

So, for this university, the actual student satisfaction score is the same value as the fitted student satisfaction score estimated from the multiple regression model. The values of student–staff ratio and academic services spend allow us to predict the student satisfaction score very well.

Activity 23 *More fitted values and residuals*

For the University of Exeter, the student satisfaction score was 4.18, the student–staff ratio was 15.8, and the academic services spend was £1689 per student. The University of Exeter is the 7th university listed in the sample. For Queen Mary University of London, the student satisfaction score was 4.12, the student–staff ratio was 11.9, while the academic services spend was £1548 per student. Queen Mary is the 18th university listed in the sample.

Calculate the residuals w_7 and w_{18}. Comment on the values of the residuals you obtain.

With fitted values and residuals in place, you should be able to do Activity 24.

Activity 24 *How to check the model assumptions*

(a) Explain how you would check that Assumption 2, that the W_is have zero mean and constant variance, is reasonable.

(b) Explain how you would check that Assumption 3, that the W_is are normally distributed, is reasonable.

You will check the model assumptions for the university student satisfaction data in the following activity.

Activity 25 *Checking the assumptions for the student satisfaction model*

Figure 34 shows the residual plot and the normal probability plot of the residuals for the fitted multiple linear regression model given in Example 17 for the university student satisfaction dataset first considered in Example 16.

Do these plots suggest that the model assumptions are reasonable?

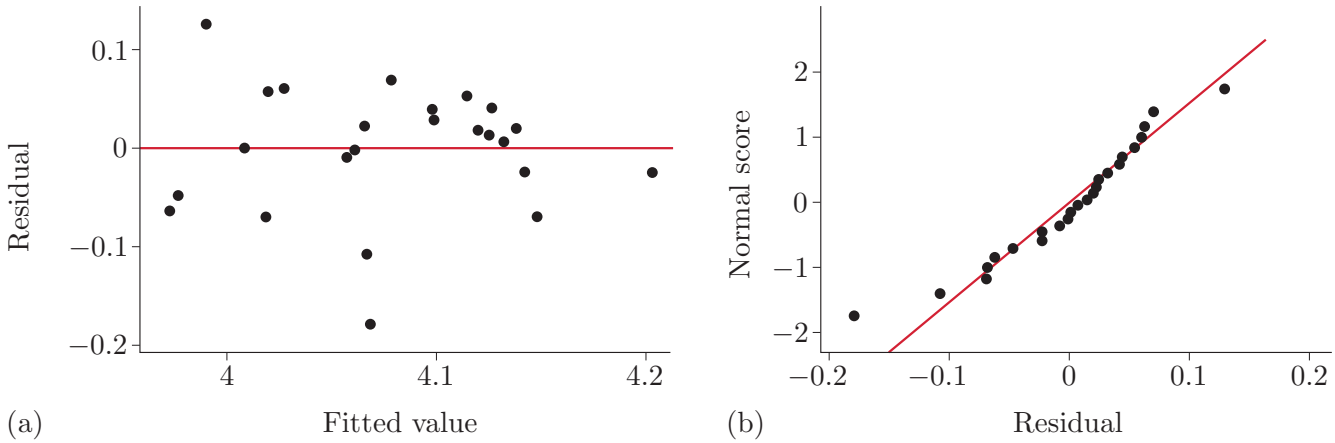


Figure 34 Checking the assumptions for the student satisfaction data: (a) residual plot; (b) normal probability plot

5.4 Multiple regression in Minitab

The final part of this section involves using multiple linear regression in Minitab.

Refer to Chapter 3 of Computer Book C for the rest of the work in this section.

Exercise on Section 5

Exercise 4 *Another multiple regression model for growth of GDP*

In Activity 22, a multiple regression model was fitted to data from 128 countries with response variable the rate of growth of gross domestic product (Y) over 2000–2010, and three explanatory variables: log of output per head in 2000 (x_1), share of gross fixed capital formation in the ten-year period (x_2), and total enrolment in secondary school (x_3). There is also available a fourth explanatory variable, x_4, the prevalence of HIV as a proportion of population for ages 15–49. The prevalence of HIV is a factor that might affect growth in some poorer countries because a high prevalence of HIV can reduce the contribution of productive workers. Data for this explanatory variable were available for only 78 of the 128 countries considered in Activity 22. A multiple regression model with all four explanatory variables was fitted for the 78 countries with complete data. The fitted model is

'Prevalence' is a word often used for a proportion or percentage when talking about medical conditions.

$$Y = 0.357 - 0.0895\,x_1 + 0.02118\,x_2 + 0.00519\,x_3 - 0.00892\,x_4.$$

The p-values for individual two-sided tests of the hypotheses $H_0 : \beta_j = 0$, are 0.000 for $j = 1, 2, 3$ and 0.032 for $j = 4$. The residual plot and normal probability plot of residuals are given in Figure 35.

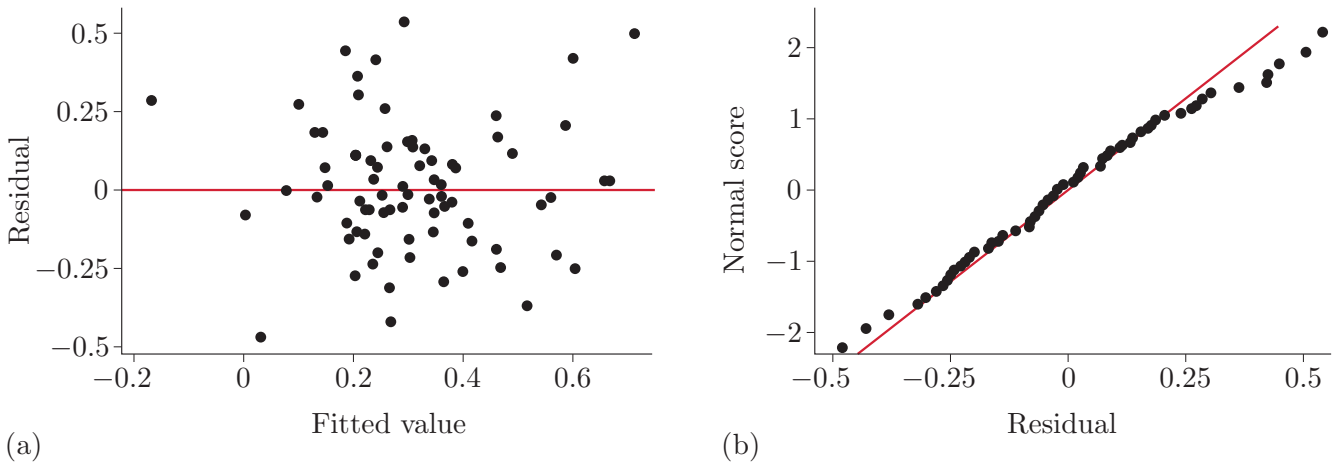


Figure 35 Checking the assumptions for the GDP dataset: (a) residual plot; (b) normal probability plot

(a) Explain why this analysis suggests that all four explanatory variables together affect the rate of growth of GDP.

(b) Do the model assumptions seem reasonable?

(c) Interpret the regression coefficients in the fitted model.

(d) For the fictional South Asian country in Activity 22(c) whose log of output per head in 2000 was 6, whose gross fixed capital formation share was 25%, whose total enrolment in secondary schools was 40%, and whose HIV prevalence is 0.1, use the fitted multiple regression model to predict its growth in GDP. How does this prediction compare

with the prediction you made on the basis of the multiple regression model with just three explanatory variables in Activity 22(c)?

Summary

In this unit, you have learned about regression models. The general regression model has been defined and a particular simple case, the linear regression model with one explanatory variable, has been treated in some depth. You have learned how to fit linear regression models to data, and to check the assumptions of a fitted model. Also, you have learned how to test whether there really is any linear regression relationship at all, and have briefly explored confidence intervals for the slope and for the mean response for a given value of the explanatory variable, and prediction intervals for the response for a given value of the explanatory variable. Finally, you saw how linear regression can be extended to incorporate more than one explanatory variable through multiple regression.

You have used Minitab to fit linear regression models, both simple and multiple, and to produce appropriate plots in order to check the modelling assumptions.

Learning outcomes

After you have worked through this unit, you should be able to:

- appreciate that an explanatory variable might be thought of as 'explaining' the value of another (response) variable, and that a response variable 'responds' to the value of one or more other (explanatory) variables
- understand that a general regression model contains a function describing how the response variable is related to the explanatory variable, and a random term which models the variation in the response
- appreciate that a linear regression model is a special case of the general regression model in which the relationship between the variables is linear
- understand that the random terms in the linear regression model are assumed to be independent with constant, zero mean and constant variance
- use a scatterplot to decide if a regression model (or a linear regression model) might be an appropriate model for the data

- fit a straight-line model to data using the method of least squares, both by hand given summary statistics for the data, and using Minitab
- calculate fitted values, residuals and predicted values
- use Minitab to produce residual plots and normal probability plots of the residuals in order to check the assumptions of a linear regression model
- appreciate that if a residual plot shows a pattern, then the assumption of constant, zero mean and constant variance of the random terms might not be justified
- appreciate that if the residuals in a normal probability plot do not fall close to a straight line, then the random terms of a linear regression model might not be normally distributed
- given summary statistics for the data, test if the response variable is related to the explanatory variable in a simple linear regression model
- given summary statistics for the data, obtain a confidence interval for the mean response in a linear regression model
- given summary statistics for the data, calculate a prediction interval for the response in a linear regression model
- appreciate how the linear regression variable with one explanatory variable is extended to the multiple regression model with several explanatory variables
- interpret the (partial) regression coefficients of the multiple linear regression model
- appreciate that the assumptions of the multiple regression model are the same as those of the simple linear regression model, and use residual plots and normal probability plots of residuals in the same way to check the assumptions
- use Minitab to fit a multiple regression model.

Solutions to activities

Solution to Activity 1

It would be natural to regard height as the response variable and age as the explanatory variable. This is because age 'explains' height and it wouldn't make sense to think of height 'changing' age.

Solution to Activity 2

(a) The natural background for this example would be a paper manufacturer wishing to estimate the optimal amount of hardwood to use in production to ensure the strongest possible paper. To do this, he must know how tensile strength depends on the percentage of hardwood in the pulp. That is, tensile strength is the response variable and hardwood content is the explanatory variable.

(b) In the scatterplot in Figure 7, there is a very evident relationship between the two variables. However, the relationship is not linear. It appears (from this experiment) that kraft paper is at its strongest for some intermediate level of pulp hardwood content (about 10%). A curve (quadratic or cubic) might be useful to model the relationship.

Solution to Activity 3

An appropriate regression model for these data might also be of the form

$$Y_i = \alpha + \beta x_i + W_i.$$

As in Example 6, α and β are the intercept and slope, respectively, of the straight line relating the variables, and the W_is are random terms accounting for the scatter around the straight line. In this case, the random terms W_i might have normal distributions with zero mean and some constant variance σ^2. Moreover, the W_is are independent because the height of one schoolboy has no affect on the height of another schoolboy.

Solution to Activity 4

Since $\alpha + \beta x_i$ is a constant, use the results from Unit 4 that, for any random variable X, $E(a + bX) = a + b\,E(X)$, $V(a + bX) = b^2\,V(X)$, with $a = \alpha + \beta x_i$, $b = 1$ and $X = W_i$, to find that

$$\begin{aligned} E(Y_i) &= E(\alpha + \beta x_i + W_i) = \alpha + \beta x_i + E(W_i) \\ &= \alpha + \beta x_i + 0 = \alpha + \beta x_i, \end{aligned}$$

$$V(Y_i) = V(\alpha + \beta x_i + W_i) = V(W_i) = \sigma^2.$$

Solution to Activity 5

(a) The problem with using the sum of residuals is that positive and negative residuals (which might be quite large in absolute value) cancel each other out. By summing the squared residuals, residuals that are large in absolute value add substantially to the sum, whether they be positive or negative.

(b) Instead of summing squared residuals, you could sum the absolute values of the residuals, also forcing large residuals to contribute substantially to the sum whether they are positive or negative. Other possibilities include taking the residuals to the fourth power prior to summing.

As an aside, minimising the sum of absolute values of residuals is also quite a popular method in statistics. An advantage it has over least squares is that it is less readily influenced by outliers; a disadvantage is that it does not afford explicit formulas for parameter estimates (the sum of absolute residuals has to be minimised numerically using a computer). This method will not be considered further in this module.

Solution to Activity 6

Equation (2) gives

$$R(\gamma) = \sum_{i=1}^{n}(y_i - \gamma x_i)^2 = \sum_{i=1}^{n}(y_i^2 - 2\gamma y_i x_i + \gamma^2 x_i^2)$$
$$= \sum_{i=1}^{n} y_i^2 - 2\gamma \sum_{i=1}^{n} y_i x_i + \gamma^2 \sum_{i=1}^{n} x_i^2.$$

This is of the form $a\gamma^2 + b\gamma + c$ with

$$a = \sum_{i=1}^{n} x_i^2, \quad b = -2\sum_{i=1}^{n} x_i y_i, \quad c = \sum_{i=1}^{n} y_i^2.$$

(Here, the standard convention of writing $\sum_{i=1}^{n} x_i y_i$ rather than the equivalent $\sum_{i=1}^{n} y_i x_i$ has been followed.)

Solution to Activity 7

(a) (i) Expanding the square in the right-hand side of Equation (3) and manipulating further, we find that

$$a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a} + c = a\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2}\right) - \frac{b^2}{4a} + c$$
$$= ax^2 + bx + \frac{b^2}{4a} - \frac{b^2}{4a} + c$$
$$= ax^2 + bx + c,$$

as required.

(ii) When $a > 0$, the expression on the right-hand side of Equation (3) comprises a positive constant times a squared term depending on x, plus constants. It is therefore minimised if we can choose x to make the squared term zero. This happens if

$$x + \frac{b}{2a} = 0,$$

that is,

$$x = -\frac{b}{2a},$$

as required.

(b) With $x = \gamma$ and, from the solution to Activity 6, $b = -2\sum_{i=1}^{n} x_i y_i$ and $a = \sum_{i=1}^{n} x_i^2$, the minimiser of $R(\gamma)$ is given by

$$\gamma = -\frac{-2\sum_{i=1}^{n} x_i y_i}{2\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Solution to Activity 8

The least squares estimate of the slope γ is

$$\widehat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{219\,817}{796\,253} \simeq 0.276.$$

The least squares line through the scattered data points has equation

$$y = 0.276\,x.$$

That is, the regression relationship between the explanatory variable and the response variable can be written

$$\text{beetle count} = 0.276 \times \text{bracket weight}.$$

(In practice, you should always obtain a scatterplot before fitting a regression model. In fact, in this case, a scatterplot suggests that an unconstrained line would be more appropriate than a line through the origin.)

Solution to Activity 9

(a) Starting from Equation (4),

$$\begin{aligned} S_{xx} &= \sum (x_i - \overline{x})^2 = \sum (x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\ &= \sum x_i^2 - 2\overline{x}\sum x_i + n\overline{x}^2 = \sum x_i^2 - 2n\overline{x}^2 + n\overline{x}^2 \\ &= \sum x_i^2 - n\overline{x}^2, \end{aligned}$$

which is the second version of Equation (7).

(b) Mathematically, the only difference is a notational change, from xs in part (a) to ys here.

(c) Starting from Equation (6),

$$\begin{aligned} S_{xy} &= \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum (x_i y_i - x_i\overline{y} - \overline{x}y_i + \overline{x}\,\overline{y}) \\ &= \sum x_i y_i - \overline{y}\sum x_i - \overline{x}\sum y_i + n\overline{x}\,\overline{y} \\ &= \sum x_i y_i - n\overline{y}\,\overline{x} - n\overline{x}\,\overline{y} + n\overline{x}\,\overline{y} \\ &= \sum x_i y_i - n\overline{x}\,\overline{y}, \end{aligned}$$

which is the second version of Equation (9).

Solution to Activity 10

Using Equations (7) to (9), S_{xx}, S_{yy} and S_{xy} can be calculated from the summary statistics as

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 30\,409 - \frac{575^2}{11} \simeq 352.182,$$

$$S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 179.14 - \frac{44.2^2}{11} \simeq 1.536,$$

$$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 2324.8 - \frac{575 \times 44.2}{11} \simeq 14.345.$$

Solution to Activity 11

When $x = \overline{x}$ is inserted in the equation for the least squares line $y = \overline{y} + \widehat{\beta}(x - \overline{x})$, we find that

$$y = \overline{y} + \widehat{\beta}(\overline{x} - \overline{x}) = \overline{y},$$

as required.

Solution to Activity 12

(a) For these data,

$$S_{xx} = 500\,000 - \frac{3000^2}{30} = 200\,000,$$

$$S_{xy} = 743\,000 - \frac{3000 \times 7395}{30} = 3500.$$

(b) The least squares estimate of the slope is

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{3500}{200\,000} = 0.0175.$$

The estimate of the intercept term is

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = \frac{7395}{30} - 0.0175 \times \frac{3000}{30} = 244.75.$$

The equation of the least squares line is

$$y = 244.75 + 0.0175\,x,$$

or, equivalently,

$$\text{taps} = 244.75 + 0.0175 \times \text{caffeine dose},$$

where taps are counted per minute, and the caffeine dose is measured in mg.

(c) The value $\widehat{\alpha} = 244.75$ is the estimated value of the intercept, that is, the value of the regression line when $x = 0$. It is meaningful in this case as the estimated value of the average number of taps per minute (or the predicted number of taps per minute) for a student in receipt of no caffeine. (As an aside, this value is not the same as the average response of the 10 no-caffeine students who happened to be measured in the experiment; it is, however, extremely close since that average happens to be 244.8.)

The value $\widehat{\beta} = 0.0175$ is the estimated value of the slope. It estimates that for each additional milligram of caffeine, a student might on average be able to increase his average number of taps per minute by 0.0175.

(d) If the caffeine dose is 50 mg, the predicted number of taps per minute is

$$244.75 + 0.0175 \times 50 = 245.625.$$

Solution to Activity 13

There is a definite pattern in the residual plot in Figure 23. The residuals are increasing at first, then there is a single large negative residual, and finally the residuals return to a high positive level before decreasing. That is, Assumption 2, that the residuals come from distributions with constant, zero mean and constant variance, appears to be violated. It seems that a linear regression model is not a good model for these data after all.

(The residual plot suggests a systematic discrepancy from linearity throughout the range of the data as well as an outlier. This is despite the claim in Example 3, based on Figure 4, that 'there may well be a straight-line relationship' and the fitting of a linear regression model in Exercise 1. As well as the outlier, which can be seen in Figure 4, perhaps the data deserve to be modelled by lines of different slope either side of the outlier.)

Solution to Activity 14

The pattern of points in the residual plot gives no reason to doubt the assumption of constant, zero mean but gives plenty of reason to doubt the assumption of constant variance. Instead, the variability of the residuals appears to increase as the sizes of the fitted values increase. (The 'band' of points widens towards the right.) The linear regression model appears not to be appropriate for these data in the sense that constant variance cannot be assumed.

(Actually, in Example 26 of Unit 1 it was commented that 'an extra feature that you might perceive in [the scatterplot of these data] is that the amount of spread of the points about any central line appears to increase as the values of the measurements increase'.)

Solution to Activity 15

$$\begin{aligned}\sum_{i=1}^{n} w_i &= \sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}\,x_i) = \sum_{i=1}^{n}\{y_i - (\overline{y} - \widehat{\beta}\,\overline{x}) - \widehat{\beta}\,x_i\} \\ &= \sum_{i=1}^{n}\{y_i - \overline{y} - \widehat{\beta}\,(x_i - \overline{x})\} = \sum_{i=1}^{n}(y_i - \overline{y}) - \widehat{\beta}\sum_{i=1}^{n}(x_i - \overline{x}) \\ &= n\overline{y} - n\overline{y} - \widehat{\beta}(n\overline{x} - n\overline{x}) = 0.\end{aligned}$$

Solution to Activity 16

There is no particular pattern in the residual plot in Figure 27(a) (other than that due to the very discrete nature of the values of the explanatory variable). It seems that Assumption 2, that the W_is come from distributions with constant, zero mean and constant variance, is a reasonable one.

Also, the normal probability plot of the residuals in Figure 27(b) appears to accommodate Assumption 3, that the W_is are normally distributed. This is because the points in the plot roughly follow a straight line; the main departures from this, if any, are due to the 'stacking up' (that is,

jittering) of points with the same value of their explanatory variables and their response variables, and hence their residuals.

Overall, the linear regression model with normally distributed random terms appears to be a reasonable one to explain the dependence of the number of taps per minute on caffeine dose.

Solution to Activity 17

(a) $s^2 = \dfrac{\sum(y_i - \widehat{y}_i)^2}{n-2} = \dfrac{0.952}{9} \simeq 0.1058.$

(b) The null hypothesis is $H_0 : \beta = 0$. From Distributional Result (11), the null distribution of the test statistic is $t(n-2) = t(9)$. The observed value of the test statistic is

$$\frac{\widehat{\beta} - 0}{s/\sqrt{S_{xx}}} = \frac{0.04}{\sqrt{0.1058}/\sqrt{352.18}} \simeq 2.308.$$

The 0.975-quantile of $t(9)$ is 2.262 and the 0.99-quantile of $t(9)$ is 2.821, so the p-value for a two-sided test is slightly less than 0.05. (A computer gave 0.047 for the p-value.) There is therefore moderate evidence against H_0, that there is no relationship between cholesterol and age, but with a p-value close to 0.05, the evidence is somewhat marginally moderate to weak in this case.

Solution to Activity 18

(a) To calculate 95% intervals for the finger-tapping data, the 0.975-quantile of $t(28)$ is required: from the table in the Handbook, this is $t = 2.048$.

(b) Using Interval (12), a 95% confidence interval for the mean $\alpha + 40\beta$ is given by

$$\left(\widehat{\alpha} + \widehat{\beta}\, x_0 \pm t\, s \sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n}}\right)$$

$$= \left(245.45 \pm 2.048\sqrt{4.7946}\sqrt{\frac{(40-100)^2}{200\,000} + \frac{1}{30}}\right)$$

$$\simeq (245.45 \pm 1.016) \simeq (244.43, 246.47).$$

(c) Using Interval (13), a 95% prediction interval for the finger-tapping frequency attained by an individual after a 40 mg dose of caffeine is given by

$$\left(\widehat{\alpha} + \widehat{\beta}\, x_0 \pm t\, s \sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n} + 1}\right)$$

$$= \left(245.45 \pm 2.048\sqrt{4.7946}\sqrt{\frac{(40-100)^2}{200\,000} + \frac{1}{30} + 1}\right)$$

$$\simeq (245.45 \pm 4.598) \simeq (240.85, 250.05).$$

Notice that this prediction interval is wider than the confidence interval calculated in part (b).

Solution to Activity 19

The zoologist is interested in modelling height using the weight and age of the giraffes. So height is the response variable Y, while weight and age are the explanatory variables; let weight be denoted x_1 and age be denoted x_2. Then the multiple regression model for data on weight, age and height is

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + W_i,$$

where W_i is a normally distributed random variable with zero mean and constant variance.

Solution to Activity 20

Since the p-value for the two-sided test of the null hypothesis $H_0 : \beta_1 = 0$ is 0.000, this means that $p < 0.01$. Therefore there is strong evidence to suggest that β_1 is not 0, that is, the regression coefficient for x_1 is not equal to 0.

The p-value for the two-sided test of the null hypothesis $H_0 : \beta_2 = 0$ is 0.002. So once again $p < 0.01$ and there is strong evidence to suggest that β_2 is not 0, that is, the regression coefficient for x_2 is also not equal to 0.

Notice that when x_1 and x_2 were considered individually in separate linear regression models in Example 16, the p-values for the slope parameters suggested that there wasn't enough evidence to suggest that either of them was non-zero. (This was especially true for x_2.) However, when we have both x_1 and x_2 in the model, the p-values suggest that there is strong evidence that the regression coefficients for both explanatory variables are non-zero. So it looks like student–staff ratio and academic services spend work *together* to affect student satisfaction.

Solution to Activity 21

(a) The interpretation of the regression coefficients is as follows.

- If the value of specific gravity (x_1) increases by one unit, and the value of moisture content (x_2) remains fixed, then the strength of timber beams (y) would be expected to increase by 8.50.
- If the value of moisture content (x_2) increases by one unit, and the value of specific gravity (x_1) remains fixed, then the strength of timber beams (y) would be expected to decrease by 0.265. The decrease is because of the negative regression coefficient.

(b) For the two-sided test of the null hypothesis $H_0 : \beta_1 = 0$, since $p = 0.002 < 0.01$, there is strong evidence to suggest that β_1 is not zero. However, for the two-sided test of the null hypothesis $H_0 : \beta_2 = 0$, since $p = 0.069$ satisfies $0.05 < p < 0.1$, there is only weak evidence to suggest that β_2 is not equal to zero. Therefore, overall there is only weak evidence to suggest that both x_1 and x_2 *together* influence the strength of timber beams.

(c) Using the fitted multiple regression line, a beam with $x_1 = 0.5$ and $x_2 = 10$ is predicted to have strength

$$10.29 + 8.50 \times 0.5 - 0.265 \times 10 = 11.89.$$

Solution to Activity 22

(a) The p-values for each individual two-sided test of the null hypothesis $H_0 : \beta_j = 0$, for $j = 1, 2, 3$, are 0.000, which means that for each regression coefficient $p < 0.01$. There is therefore strong evidence that each regression coefficient is non-zero, which in turn implies that together the three explanatory variables influence Y, the rate of growth of GDP.

(b) The regression coefficients can be interpreted as follows.

- Regression coefficient for x_1: If the value of x_1 increases by one unit, and the values of x_2 and x_3 remain fixed, then the rate of growth of GDP (y) would be expected to decrease by 0.0923 (or a little over 9% over the ten-year period). The decrease is because of the negative regression coefficient. From the information in the question, this makes sense because it means that poorer countries tend to catch up with richer countries by copying existing technology available on global markets, and countries who are initially richer, with higher values of x_1, will grow more slowly.
- Regression coefficient for x_2: If the value of x_2 increases by one unit, and the values of x_1 and x_3 remain fixed, then the rate of growth of GDP (y) would be expected to increase by 0.02425 (or about 2.4% over the ten-year period). The increase is because of the positive regression coefficient. From the information in the question, this makes sense because it suggests that countries that invest a greater share of their resources in capital goods, such as industrial plants, machinery and equipment, than consumption (and so have a higher value of x_2), grow faster than countries that focus more on consumption (and so have a lower value of x_2).
- Regression coefficient for x_3: If the value of x_3 increases by one unit, and the values of x_1 and x_2 remain fixed, then the rate of growth of GDP (y) would be expected to increase by 0.00493 (or about 0.5% over the ten-year period). The increase is because of the positive regression coefficient. From the information in the question, this makes sense because an increase in enrolment in secondary school (x_3) increases the education of the workforce, which would be associated with faster economic growth and increased change in GDP.

(c) Using the fitted multiple regression line, a country with $x_1 = 6$, $x_2 = 25$ and $x_3 = 40$ is predicted to have had a growth rate over the ten-year period of

$$0.312 - 0.0923 \times 6 + 0.02425 \times 25 + 0.00493 \times 40 \simeq 0.56$$

(or about 56%).

Solution to Activity 23

The fitted student satisfaction score for Exeter is

$$\widehat{y}_7 = 3.157 + 0.0484 \times 15.8 + 0.000166 \times 1689 \simeq 4.2021 \simeq 4.20.$$

The associated residual is therefore

$$w_7 = y_7 - \widehat{y}_7 = 4.18 - 4.20 = -0.02.$$

For the University of Exeter, the student satisfaction score seems to be fairly close to the fitted student satisfaction score estimated from the multiple regression model. The values of student–staff ratio and academic services spend allow us to predict the student satisfaction score well.

The fitted student satisfaction score for Queen Mary University of London is

$$\widehat{y}_{18} = 3.157 + 0.0484 \times 11.9 + 0.000166 \times 1548 \simeq 3.9900 \simeq 3.99.$$

The associated residual is therefore

$$w_{18} = y_{18} - \widehat{y}_{18} = 4.12 - 3.99 = 0.13.$$

For this university, the student satisfaction score is quite a bit higher than the fitted student satisfaction score estimated from the multiple regression model. (In fact, the student satisfaction score for this university has the largest positive residual in the sample.) The values of student–staff ratio and academic services spend do not allow us to predict the student satisfaction score so well in this case.

Solution to Activity 24

(a) Assumption 2, that the W_is have zero mean and constant variance, can be checked by using a residual plot which plots the observed residuals w_i against the fitted values $\widehat{y}_i$. The residuals should be scattered randomly about zero if the assumption is true.

(b) Assumption 3, that the W_is are normally distributed, can be checked using a normal probability plot for the observed residuals w_i. If the assumption is plausible, then the residuals should lie reasonably close to a straight line.

Solution to Activity 25

With the possible exception of one large positive and one large negative residual, the points in the residual plot appear to be scattered randomly about zero, suggesting that the assumption that the W_is have constant, zero mean and constant variance seems plausible.

The residuals lie reasonably close to a straight line in the normal probability plot, so the assumption that the W_is are normally distributed seems plausible. There is perhaps a hint of curvature, but with only 24 data points it doesn't seem to be sufficient to rule out the assumption of normality.

Solutions to exercises

Solution to Exercise 1

(a) For Forbes's data, S_{xx} and S_{xy} are given by

$$S_{xx} = 10\,820.9966 - \frac{426^2}{17} \simeq 145.938,$$

$$S_{xy} = 86\,735.495 - \frac{426 \times 3450.2}{17} \simeq 277.542.$$

The least squares estimates of β and α are

$$\widehat{\beta} = \frac{277.542}{145.938} \simeq 1.90$$

and

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = \frac{3450.2}{17} - \widehat{\beta} \times \frac{426}{17} \simeq 155.30.$$

The equation of the least squares line is therefore

$$y = 155.30 + 1.90\,x.$$

That is, the fitted model is

$$\text{boiling point} = 155.30 + 1.90 \times \text{atmospheric pressure},$$

where temperature is measured in °F and pressure in inches Hg.

(b) The estimated value of the intercept, $\widehat{\alpha}$, is of little interest in this context because it refers to zero atmospheric pressure, which is of no interest on a mountain and is way beyond the range of the data to which the linear regression model was fitted.

The value $\widehat{\beta} = 1.90$ is the estimated value of the slope. It estimates that for each increase in atmospheric pressure of one inch of mercury, the boiling point of water will, on average, increase by about 1.9 °F.

(c) If the pressure is 25 inches Hg, the predicted boiling point of water is

$$155.30 + 1.90 \times 25 = 202.8\ ^\circ\text{F}.$$

Solution to Exercise 2

(a) On the basis of Figure 28, yes, a linear regression model appears to continue to provide a good model for the full dataset.

(b) This plot shows no particular pattern; the points seem to be randomly scattered around zero. That is, Assumption 2 seems to be satisfied. (Or do you think you perceive a curve to the plot, in which case the linearity of the model would appear to be in doubt?)

(c) The points on the probability plot fall in a pretty good straight line. The assumption of normality does not appear to be in doubt.

(d) Numerically, the slopes of the lines are very similar, but the intercepts are rather different. The lines are plotted in Figure 36.

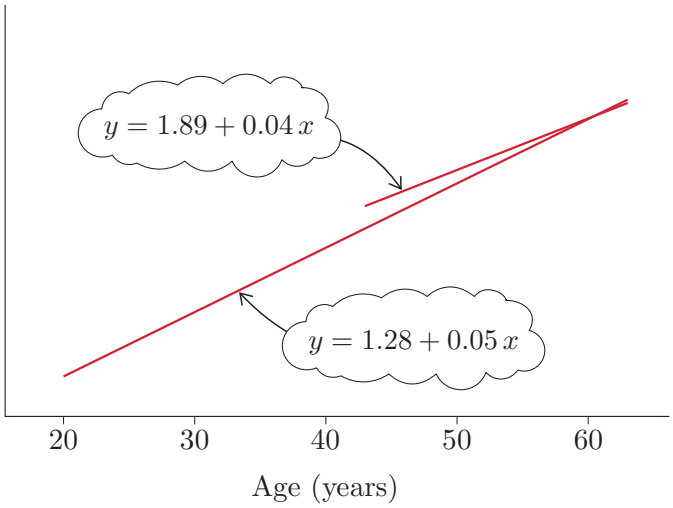


Figure 36 The lines $y = 1.89 + 0.04\,x$ plotted for $43 \leq x \leq 63$ and $y = 1.28 + 0.05\,x$ plotted for $20 \leq x \leq 63$

The fitted lines are similar for the older age range. However, they will clearly differ more for lower ages. So, yes, the line has changed appreciably with the inclusion of younger patients. (In particular, the intercept has changed substantially.)

(Or is there indeed a curve in the residual plot of part (b), suggesting a slightly different, non-linear, relationship over the wider age range? Statistics is full of such ambiguities, especially when arguments are being made, as here, on the basis of small datasets.)

Solution to Exercise 3

(a) First,

$$s^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2} = \frac{2.455}{22} \simeq 0.1116.$$

The null hypothesis is $H_0 : \beta = 0$. From Distributional Result (11), the null distribution of the test statistic is $t(n-2) = t(22)$. The observed value of the test statistic is

$$\frac{\widehat{\beta} - 0}{s/\sqrt{S_{xx}}} = \frac{0.05}{\sqrt{0.1116}/\sqrt{4139.77}} \simeq 9.63.$$

From the table in the Handbook, the 0.999-quantile of $t(22)$ is 3.505 so the p-value for a two-sided test is considerably less than 0.002. (In fact, the p-value is very small indeed.) There is therefore very strong evidence against H_0; there does seem to be a relationship between age and cholesterol over the wide range of ages in the dataset.

(b) The point prediction for the value of total cholesterol for a patient with hyperlipoproteinaemia aged 35 years is

$$1.28 + 0.05 \times 35 = 3.03 \text{ mg/ml}.$$

For a 90% prediction interval, we need the 0.95-quantile of the $t(22)$ distribution, which is 1.717. Using Interval (13), a 90% prediction

interval for the value of total cholesterol for a patient with hyperlipoproteinaemia aged 35 years is given by

$$\left(\widehat{\alpha}+\widehat{\beta}\,x_0 \pm t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}}+\frac{1}{n}+1}\right)$$

$$=\left(3.03 \pm 1.717\sqrt{0.1116}\sqrt{\frac{(35-39.42)^2}{4139.77}+\frac{1}{24}+1}\right)$$

$$\simeq (3.03 \pm 0.587) \simeq (2.44, 3.62).$$

(c) The prediction interval in part (b) suggests that it is plausible that a 35-year-old individual with hyperlipoproteinaemia would have a total cholesterol level of somewhere between about 2.4 mg/ml and 3.6 mg/ml. Although this is quite a wide range of values, this is useful information since the prediction interval contains only values higher than the observed values associated with some of the younger individuals in the dataset and lower than the observed values associated with many of the older individuals in the dataset.

Solution to Exercise 4

(a) The p-values for each individual two-sided test of the null hypothesis $H_0 : \beta_j = 0$, for $j = 1, 2, 3$, are 0.000, which means that for each of the first three regression coefficients $p < 0.01$. There is therefore strong evidence that β_1, β_2 and β_3 are all non-zero. Also, the p-value for the two-sided test of the null hypothesis $H_0 : \beta_4 = 0$ is $0.032 < 0.05$. There is therefore moderate evidence that β_4 is also non-zero. Therefore there is evidence that the four explanatory variables together influence Y, the rate of growth of GDP.

(b) The points in the residual plot appear to be scattered randomly about zero, suggesting that the assumption that the W_is have constant, zero mean and constant variance seems plausible. Most of the residuals in the normal probability plot lie roughly along a straight line, so the assumption of normality of residuals also seems plausible. Having said that, a number of the larger residuals deviate from the line, so the assumption of normality might be called into question.

(c) The regression coefficients can be interpreted as follows.

- Regression coefficient for x_1: If the value of x_1 increases by one unit, and the values of x_2, x_3 and x_4 remain fixed, then the rate of growth of GDP (y) would be expected to decrease by 0.0895. (The decrease is because of the negative regression coefficient.)
- Regression coefficient for x_2: If the value of x_2 increases by one unit, and the values of x_1, x_3 and x_4 remain fixed, then the rate of growth of GDP (y) would be expected to increase by 0.02118. (The increase is because of the positive regression coefficient.)
- Regression coefficient for x_3: If the value of x_3 increases by one unit, and the values of x_1, x_2 and x_4 remain fixed, then the rate of growth of GDP (y) would be expected to increase by 0.00519. (The increase is because of the positive regression coefficient.)

- Regression coefficient for x_4: If the value of x_4 increases by one unit, and the values of x_1, x_2 and x_3 remain fixed, then the rate of growth of GDP (y) would be expected to decrease by 0.00892. (The decrease is because of the negative regression coefficient.)

Reasons why the regression coefficients for x_1, x_2 and x_3 make sense were given in the solution to Activity 22. The negative regression coefficient for x_4 makes sense because having a high prevalence of HIV can reduce productivity and therefore decrease growth.

(d) Using the fitted multiple regression line, a country with $x_1 = 6$, $x_2 = 25$, $x_3 = 40$ and $x_4 = 0.1$ is predicted to have a growth rate over the ten-year period of

$$0.357 - 0.0895 \times 6 + 0.02118 \times 25 + 0.00519 \times 40 - 0.00892 \times 0.1 \simeq 0.56.$$

Addition of HIV prevalence into the model has not changed the prediction of the growth of GDP of this country (at least, not to second-decimal-place precision).

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 3: © http://ushistoryscene.com/article/rise-of-public-education/

Page 5: © Thanavut Chao-ragam / www.123rf.com

Page 7: © eltpics This file is licensed under the Creative Commons Attribution-Non-commercial Licence http://creativecommons.org/licenses/by-nc/3.0/

Page 8: © Ina van Hateren / www.123rf.com

Page 17: © 2000–2017 vBulletin Solutions Inc

Page 19: © BruceBlaus / https://commons.wikimedia.org/wiki/File:Blausen_0052_Artery_NormalvPartially-BlockedVessel.png This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 21: © Geography Photos / Universal Images Group

Page 25: © 2008 Joyce Gross, University of California, Berkeley

Page 29: © kzenon / www.123rf.com

Page 30: © Sara Riggare

Page 33: © odessa4 / www.123rf.com

Page 44: © pifate / www.123rf.com

Page 46: © Paul Sableman This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 47: © Alan Light This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 51: © edella / iStock Editorial / Getty Images Plus

Page 54: © Zoo New England

Page 56: © kasto / www.123rf.com

Page 58: © kzenon / www.123rf.com

Page 60: © Ilbusca / iStock Unreleased / Getty Images

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.

Unit 12

Transformations and the modelling process

Introduction

You have now met some of the most important ideas of statistics:

- you can summarise the key features of a dataset and can represent it graphically in different ways
- you have seen that variability in a population can be represented by a probability distribution
- with a few assumptions, you are able to use a variety of distributions to model the behaviour of both discrete and continuous random variables
- you can use data and models to perform inference and hence to answer practical questions.

Each unit so far has dealt with particular topics or methods in statistics, which have been illustrated by examples. You might think of the techniques you have encountered in the module as statistical tools. In the first part of this unit (Sections 1 and 2), you will meet the final tools to be added to your statistical toolbox in this module: the use of *transformations* of the data, first in a one-sample context, in Section 1, and then in a regression context, in Section 2.

In contrast, the second part of this unit focuses on the overall statistical *modelling process*. Having assembled your toolbox, the aim now is to work out how to use it when confronted with a statistical problem. An introduction to the modelling process is provided in Section 3, along with some reminders of some of the basic tools at your disposal. In Section 4, you will practise undertaking a complete analysis using a variety of tools; the section consists of a chapter of Computer Book C. When you have finished your analysis, you must be able to summarise what you have done and tell interested parties about it: writing a statistical report is discussed in Section 5.

1 Transforming the data: the one-sample case

The continuous distributions available to you that are most commonly used for modelling purposes are the continuous uniform, exponential and normal distributions. These were introduced in Units 3, 5 and 6, respectively. On the next page is a summary of the ranges and shapes of these three distributions followed, in Figure 1, by graphs of examples of their p.d.f.s.

The chi-squared and t-distributions are continuous also, but are more rarely used directly for modelling purposes.

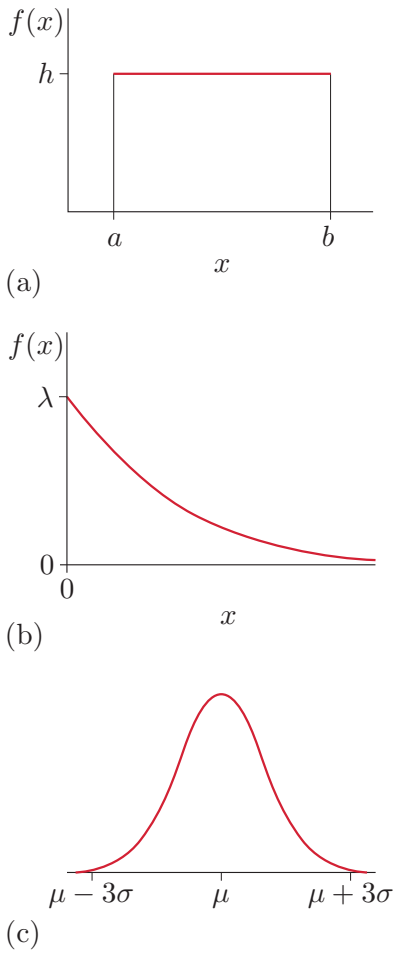
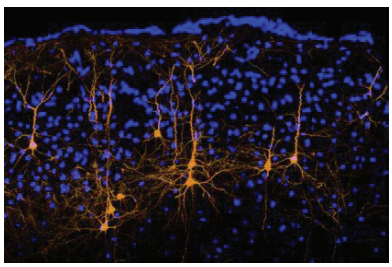


Figure 1 P.d.f.s of the (a) continuous uniform, (b) exponential and (c) normal distributions



Motor cortex neurons (the orange structures looking a little like root systems) in a mouse

Continuous models: range and shape

- The continuous uniform distribution has finite range $a < x < b$ and its p.d.f. is flat.
- The exponential distribution has range $0 < x < \infty$, unbounded to the right, and its p.d.f. is a decreasing function of x.
- The normal distribution has unbounded range $-\infty < x < \infty$ and is symmetric about a single mode that coincides with the mean. Values far from the mean have low probability.

At first sight, this presents a serious problem, since the choice of shapes covered by these models is very restricted.

Example 1 *Interspike intervals*

The motor cortex is the part of the brain concerned with movement. Neurons are cells that experience momentary electric potential changes, or 'spikes', the occurrence of which can be tracked over time. In the experiment of interest here, $n = 100$ 'interspike' intervals of motor cortex neurons of a monkey were measured (in milliseconds). The aims of the study were to describe the distribution of waiting times between spikes, and to estimate the firing rate of neurons.

Since the data in this case are waiting times between spikes, it seems a reasonable first assumption to assume that these spikes arise at random. Hence, a reasonable first model for the waiting times between spikes is the exponential distribution. An exponential p.d.f. was shown in Figure 1(b).

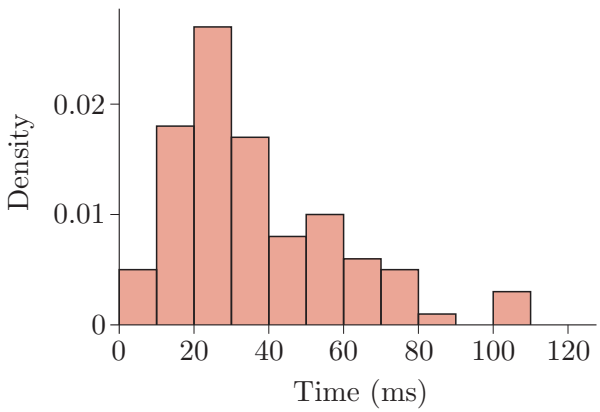


Figure 2 Histogram of interspike intervals

(Source: Zeger, S.L. and Qaqish, B. (1988) 'Markov regression models for time series: a quasi-likelihood approach', *Biometrics*, vol. 44, no. 4, pp. 1019–31)

Figure 2, on the other hand, shows a unit-area histogram of the data. The data are certainly skew and display a long right tail, as would be expected of an exponential distribution. However, it is noticeable that the mode is

not at zero, but somewhere in the range 20 to 30. This could be due to random variation, but it is more likely that it reflects a failure of the exponential model.

Activity 1 *Other inappropriate models for interspike intervals*

Why does neither the continuous uniform distribution nor the normal distribution provide a good model for the data of Example 1?

1.1 Transformations: some general considerations

When dealing with continuous data, one method by which the available collection of modelling distributions can be extended enormously is by using transformations. The idea is simple: if the data do not have the shape required, you can try to transform them, that is, take a function of them, so that the transformed data do have the shape required.

Transformations can be used for certain purposes with discrete data too, but we will not consider that here.

You met *linear* transformations of data in Units 4 and 6. Unfortunately for our current purposes, linear transformations do not change the shape of a distribution. For example, you saw in Subsection 3.1 of Unit 6 that a linear transformation of normally distributed data results in another set of normally distributed data. (It is just the mean and variance of the normal distribution that are changed.) We therefore need to consider *non-linear* transformations.

You briefly met the notion of non-linear transformations of *parameters*, not of data values, in the context of confidence intervals in Subsection 3.1 of Unit 8.

The idea behind the use of (non-linear) transformations is illustrated in Example 2 using simulated data.

Example 2 *Transforming data*

Four histograms of datasets, each based on 300 data points, are shown in Figure 3 (overleaf).

The histogram in Figure 3(a) looks as though the data might be normally distributed, but those in Figures 3(b), 3(c) and 3(d) are progressively more skew and the distributions appear to be far from normal. It may surprise you to learn that the same sample of data was used for all four histograms. First, a computer was used to generate a sample of size $n = 300$ from a normal distribution; these data are represented by the histogram in Figure 3(a). Suppose that a typical value in this dataset is denoted x_i. Then the data points used for Figure 3(b) are the values x_i^2; the data points used for Figure 3(c) are the values $e^{2(x_i-1)}$; and the data points used for Figure 3(d) are the values $1/(251x_i)$.

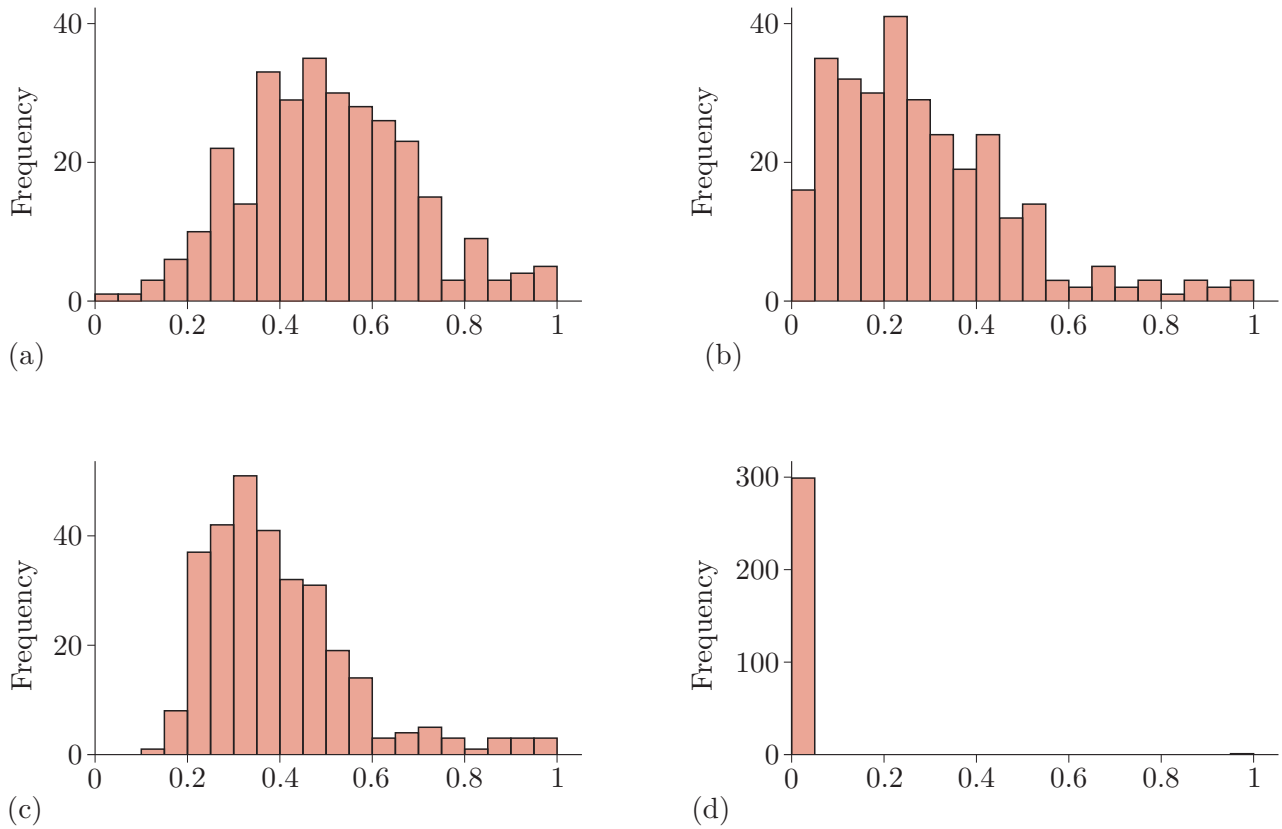


Figure 3 Four histograms

Since the data represented in Figure 3(a) are normally distributed, they could be used to carry out a t-test, for example. However, it would not be legitimate to carry out a t-test using the data in any of Figures 3(b), 3(c) or 3(d) because the variation is far from normal. Suppose now that data resembling those in Figures 3(b), 3(c) or 3(d) were to arise in practice. It would clearly be worth considering transforming them. For example, if the data looked like those in Figure 3(b), then an appropriate procedure to follow would be to take the inverse transformation to the transformation that led to Figure 3(b) from Figure 3(a) in the first place. Since the initial transformation, in this case, was to square the normally distributed data values, then the transformation of the data in Figure 3(b) that leads back to Figure 3(a) must be to take the square root of each value. It would then be appropriate to carry out t-tests on the square-root-transformed data, based on the assumption of normality.

A popular alternative target of transforming data is just to make them more symmetric rather than specifically normal.

One aim of transforming a set of data values to a different set of values by means of a mathematical transformation is, as in the second half of Example 2, to render the transformed data more plausibly normal. If we are able to do this, then we can perform statistical modelling and inference by applying the techniques we already have available for normally distributed data to the transformed dataset. Information that we obtain via the transformed version of the dataset can then be reinterpreted in terms of questions concerning the original, untransformed, data. So this is

the aim of transforming data on which we will concentrate in this section. In this case, assessment of the success or otherwise of proposed transformations can be made using normal probability plots, as introduced in Section 5 of Unit 6 and used for residuals in a regression context in Unit 11.

Wedding day in the rain in Minneapolis–St Paul

Example 3 *March precipitation in Minneapolis–St Paul*

The total precipitation (in inches) in the month of March was recorded in 30 successive years in Minneapolis–St Paul, Minnesota, USA. Figure 4(a) shows a normal probability plot for these data. The normal probability plots in Figures 4(b) and 4(c) were produced after transforming the data using the log transformation $\log x$ and the cube root transformation $x^{1/3}$, respectively. That is, if the original data are denoted by x_i, then the normal probability plot in Figure 4(b) is based on the values $\log x_i$, and the normal probability plot in Figure 4(c) is based on the values $x_i^{1/3}$.

As always in this module, 'log' (without a subscript) denotes the natural logarithm

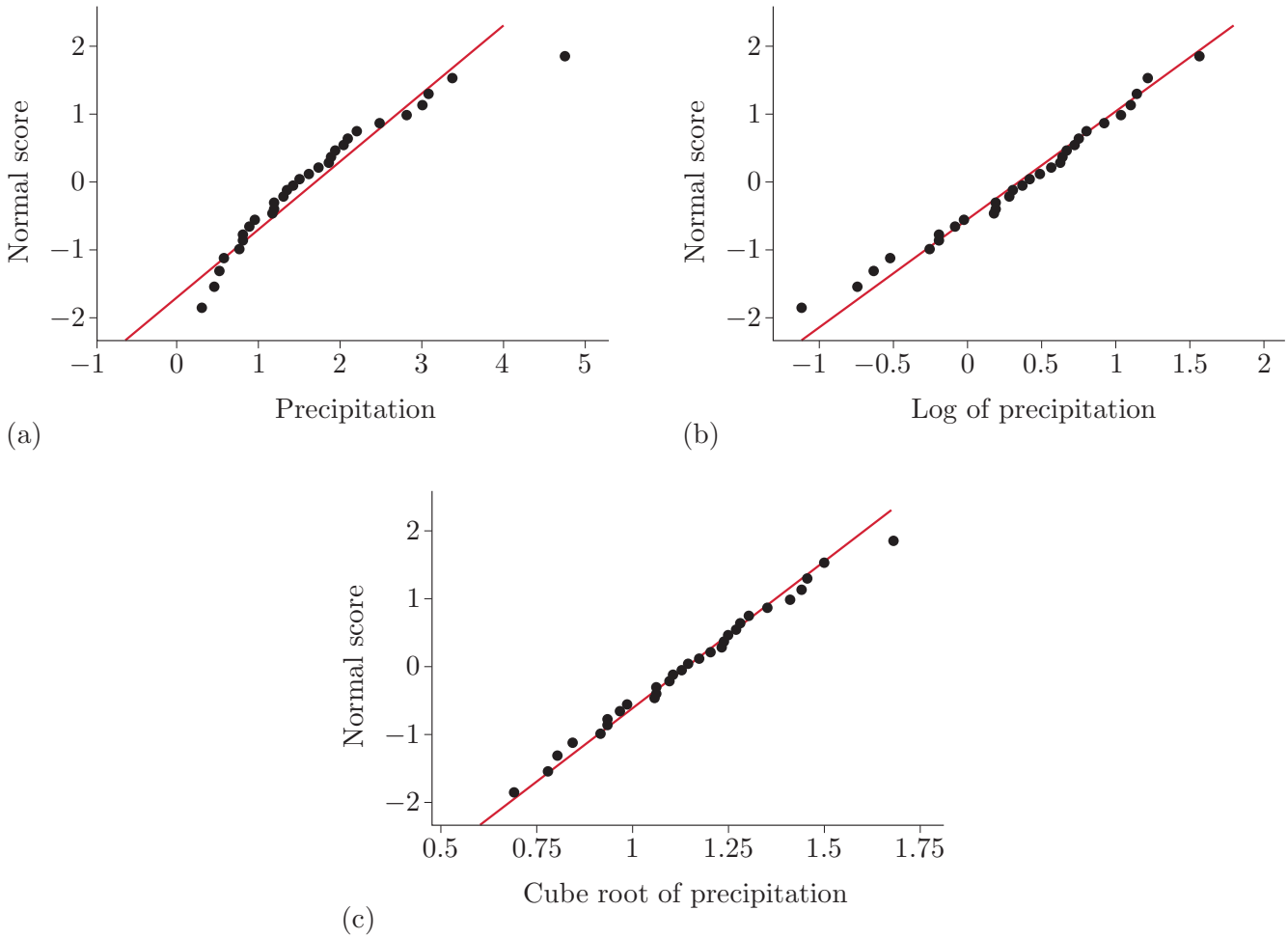


Figure 4 Normal probability plots for precipitation data: (a) untransformed, (b) log transformed, (c) cube root transformed

(Source: Hinkley, D. (1977) 'On quick choice of power transformation', *Applied Statistics*, vol. 26, no. 1, pp. 67–9)

The probability plot for the untransformed data displays a systematic pattern giving an indication of non-normality. The log transformation results in a straighter plot, suggesting that while the x_i values might not be modelled (directly) by a normal distribution, the $\log x_i$ values might be modelled by a normal distribution. Arguably, an even straighter probability plot is obtained using the cube root transformation; it seems reasonable also to model the values of $x_i^{1/3}$ using a normal distribution.

Suppose now that a typical value in a dataset is denoted by x and its transformed value by $y = h(x)$, where h is some transformation function. How do we go about choosing h? We have seen that if we can think of a possible function to act as h, then we can check whether using it results in normality of the transformed data by using a normal probability plot. But possibilities for h include those you have already seen, like $y = \log x$ and $y = x^{1/3}$, and others like $y = x^2$, $y = e^x$ and $y = 1/x$; in fact, the list is endless.

Well, first, we can observe that some transformations are simply not available for use with some data. For example, the transformation $y = \log x$ is not defined for negative (or zero) values of x, so it can be used only when all the data values are positive. The same goes for the transformation $y = \sqrt{x}$. That is, the transformation $y = h(x)$ needs to be defined for all values in the range of the distribution of the data.

Second, it turns out that the approach makes sense only if a transformation is either increasing or decreasing over the range of the distribution of x. Recall that a transformation of x is increasing if its graph rises as you move to the right through the range of x; it is decreasing if its graph falls. Alternatively, as you were reminded in Subsection 3.1 of Unit 8, a transformation $h(x)$ is increasing if its derivative is positive, that is, $h'(x) > 0$, and is decreasing if its derivative is negative, that is, $h'(x) < 0$. So we need to choose h so that either $h'(x) > 0$ or $h'(x) < 0$ over the range of values of the distribution of x.

Example 4 *The transformation $y = x^2$*

Suppose that x can take both positive and negative values. Then the transformation $y = x^2$ is neither increasing nor decreasing over the range of the distribution of x. This is illustrated in Figure 5, where the graph of $y = x^2$ is shown for $-2 < x < 2$: $y = x^2$ is decreasing between -2 and 0, then increasing between 0 and 2. Mathematically, if $h(x) = x^2$, then $h'(x) = 2x$, so $h'(x) < 0$ for $-2 < x < 0$ (in fact, for all $x < 0$) and $h'(x) > 0$ for $0 < x < 2$ (in fact, for all $x > 0$). However, if the range of the distribution of x included only values $x > 0$, then the transformation $y = x^2$ would be appropriate as it is increasing over this range.

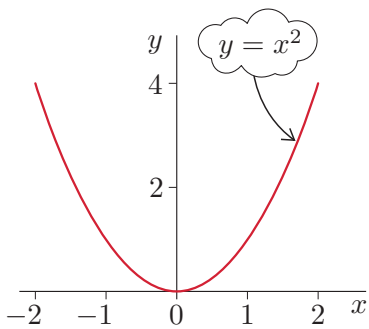


Figure 5 The graph of $y = x^2$ for x between -2 and 2

Activity 2 *Which transformation is possible?*

Suppose that a dataset consists of values x such that $-1 < x < 1$. Only one of the four transformations listed below is both defined and either

increasing or decreasing over the range of the distribution of x. Identify which one, and explain why each transformation is or is not available for such data.

$$y = \sqrt{x} = x^{1/2}, \quad y = x^4, \quad y = (2 + x)^2, \quad y = -\log x.$$

A summary of the basic requirements of transformations is given in the following box.

Basic requirements of transformations

To be suitable for consideration as a transformation of data x, a transformation $y = h(x)$ has to be defined and either increasing or decreasing over the range of the distribution of x.

The case for transforming data is strengthened if some natural physical interpretation of the transformation is available. This might, for example, have something to do with units of measurement: suppose that x represents a volume, in units of m^3, so that $x^{1/3}$ is in units of metres; then working with the latter *might* be preferable for some purposes. On the other hand, if x actually arises from the product of two terms, $x = tw$ say, though you can't individually observe t and w, then the basic rules of logarithms mean that

$$y = \log x = \log(tw) = \log t + \log w. \tag{1}$$

Since y is now a sum, it is *possible* that its distribution is 'more normal' than that of x, by a kind of Central Limit Theorem effect, albeit for a sum of only two random variables.

Most often, however, as in Example 3, there is no convenient physical interpretation, and the data are transformed simply to satisfy the requirements of the statistical procedure that you wish to use. Moreover, in such cases, again as you saw in Example 3, two (or more) transformations might prove to be broadly equally justifiable, and it doesn't then really matter which of them you choose to use.

Some general indications for choosing a transformation may, however, be given, especially for data x that are positive; it is this scenario on which we concentrate in the remainder of this section. When positive data are highly right-skew, with many relatively small values and fewer higher values, and possibly some very high values, the following transformations – which are all shown in Figure 6 (overleaf) – tend to reduce the spread of higher values more than that of lower values:

$$y = \sqrt{x} = x^{1/2}, \quad y = x^{1/3}, \quad y = \log x, \quad y = \log(1 + x).$$

This is because they all, while being increasing transformations, slow down their rate of increase as x increases.

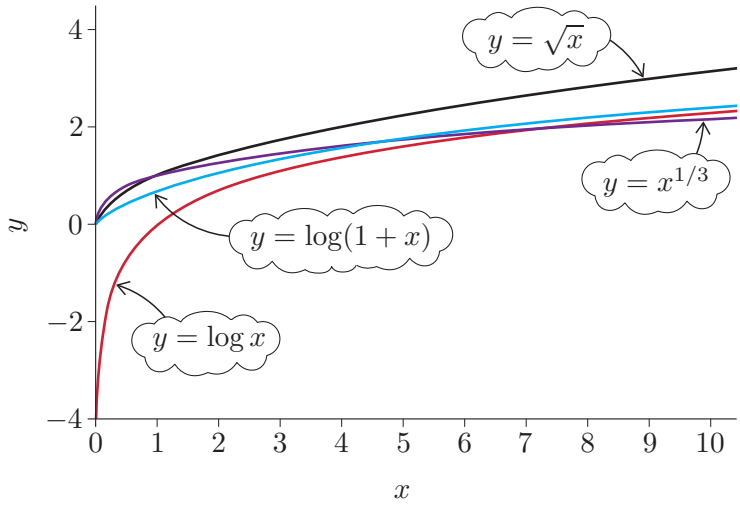


Figure 6 The graphs of $y = \sqrt{x}$, $y = x^{1/3}$, $y = \log x$ and $y = \log(1 + x)$ for $x > 0$

Each of the other transformations in Figure 6 works in the same way.

This is clarified in Figure 7 for the particular case of $y = \sqrt{x}$. You can see from the figure how the right-skew set of data values x_i is transformed to a much less skew set of transformed data values y_i. The overall effect of any one of these transformations is therefore to reduce the right-skew in the data, and potentially to make them symmetric (and even normally distributed!).

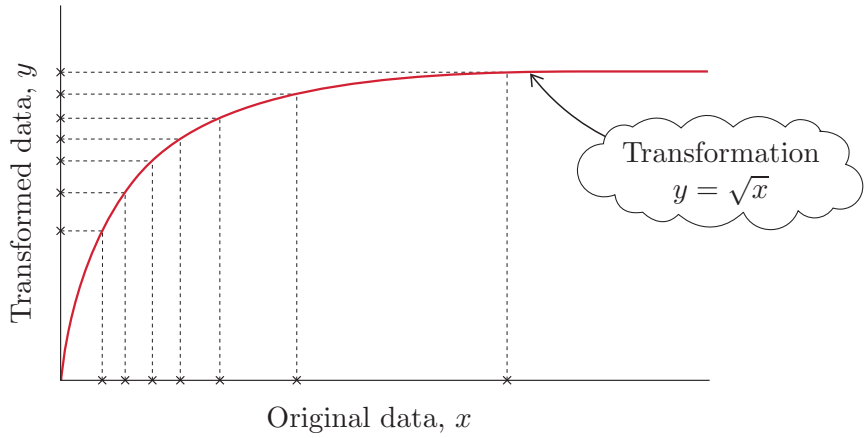


Figure 7 The way the transformation $y = \sqrt{x}$ works

1.2 The ladder of powers

The notion at the end of Subsection 1.1 can be taken a bit further. It is popular to consider a *ladder of powers*, which lists transformations of the form

$$\ldots,\ x^{-2},\ x^{-1},\ x^{-1/2},\ \log x,\ x^{1/2},\ \boxed{x^1},\ x^2,\ x^3,\ x^4,\ \ldots.$$

The transformation $y = x^1 = x$ leaves data values unchanged. Notice the position of $\log x$ in the ladder of powers: although not, in fact, a power transformation, it fills the position of x^0, which is not a valid transformation because it collapses all values to 1.

Transformations corresponding to powers below 1 on the ladder (that is, transformations to the left of x^1 in the list above) all contract high data values relative to low data values. The first two of these transformations, $y = \sqrt{x}$ and $y = \log x$, were shown in Figure 6.

The next three of these transformations,

$$y = x^{-1/2} = \frac{1}{\sqrt{x}}, \quad y = x^{-1} = \frac{1}{x}, \quad y = x^{-2} = \frac{1}{x^2}$$

are shown in Figure 8.

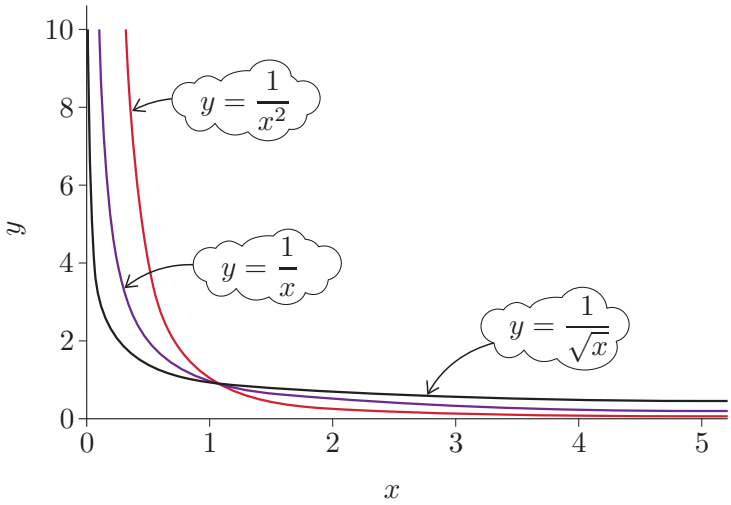


Figure 8 The graphs of $y = 1/\sqrt{x}$, $y = 1/x$ and $y = 1/x^2$ for $x > 0$

Figure 9 (overleaf) shows (via the particular example of $y = 1/x$) how these three transformations work: they too expand low values while contracting high values. However, because these are decreasing transformations, they also flip the values around: high values of x become low values of y, while low values of x become high values of y. This is not a problem in the context of using such a transformation to transform data to normality: all the transformations on the ladder of powers with powers below 1 can reduce any right-skew in the data.

On the other hand, transformations with powers above 1 on the ladder of powers (that is, transformations to the right of x^1 in the list) all expand high data values relative to low data values. These transformations are shown in Figure 10 (overleaf), and the way they work (in the particular case of $y = x^2$) is shown in Figure 11 (overleaf). These transformations are therefore most useful if we need to reduce left-skew in the distribution of the data.

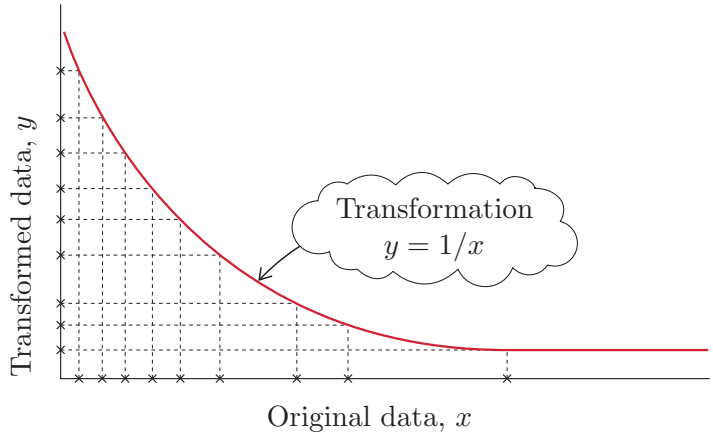


Figure 9 The way the transformation $y = 1/x$ works

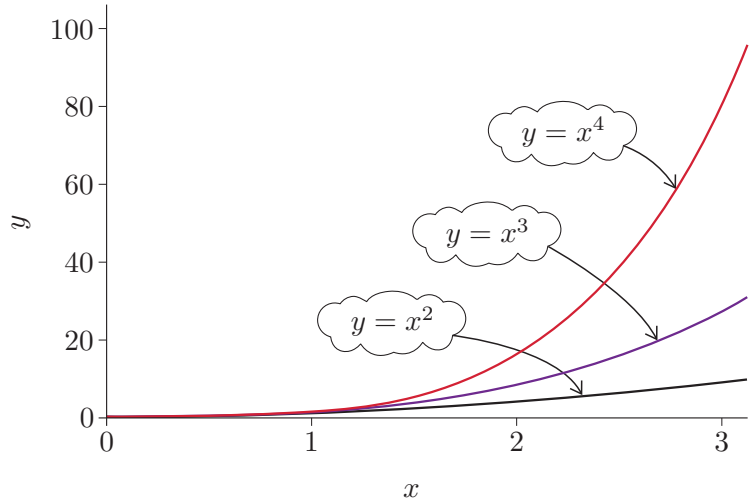


Figure 10 The graphs of $y = x^2$, $y = x^3$ and $y = x^4$ for $x > 0$

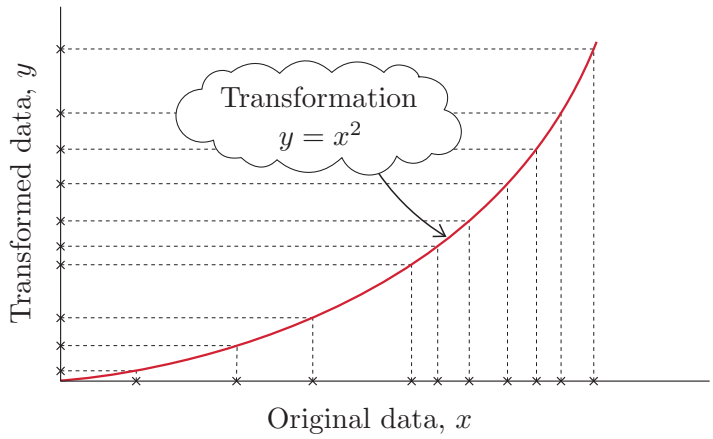


Figure 11 The way the transformation $y = x^2$ works

Screencast 12.1 further explores the effects of transformations on the ladder of powers.

Screencast 12.1 Transformations on the ladder of powers

Transformations: ladder of powers

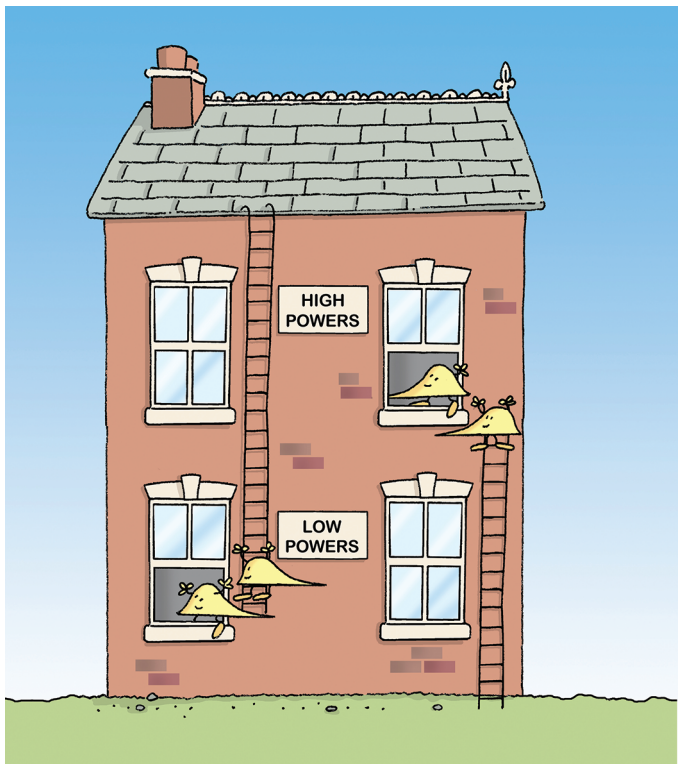
The **ladder of powers** lists transformations of the form

$$\ldots,\ x^{-2},\ x^{-1},\ x^{-1/2},\ \log x,\ x^{1/2},\ x^{1},\ x^{2},\ x^{3},\ x^{4},\ \ldots.$$

The transformation x^1 leaves the data unchanged.

When transforming skew, positive, data to make them more symmetric, and hence more amenable to modelling with a normal distribution:

- for right-skew data, go **down** the ladder of powers
- for left-skew data, go **up** the ladder of powers.



By the way, you might have noticed that, with the exception of $\log x$, which transforms positive data to data which can take any value, the transformations in the ladder of powers transform positive data to a different set of positive data. As mentioned earlier in the module, the normal distribution can still be used as a reasonable model for such data if the probability it assigns to negative values is suitably small.

Activity 3 *Which transformation to use?*

Figure 12 shows histograms for three datasets (which each happen to have been simulated using a computer). For each dataset, suggest a suitable possible transformation that would make the transformed data more symmetric and hence more nearly normally distributed.

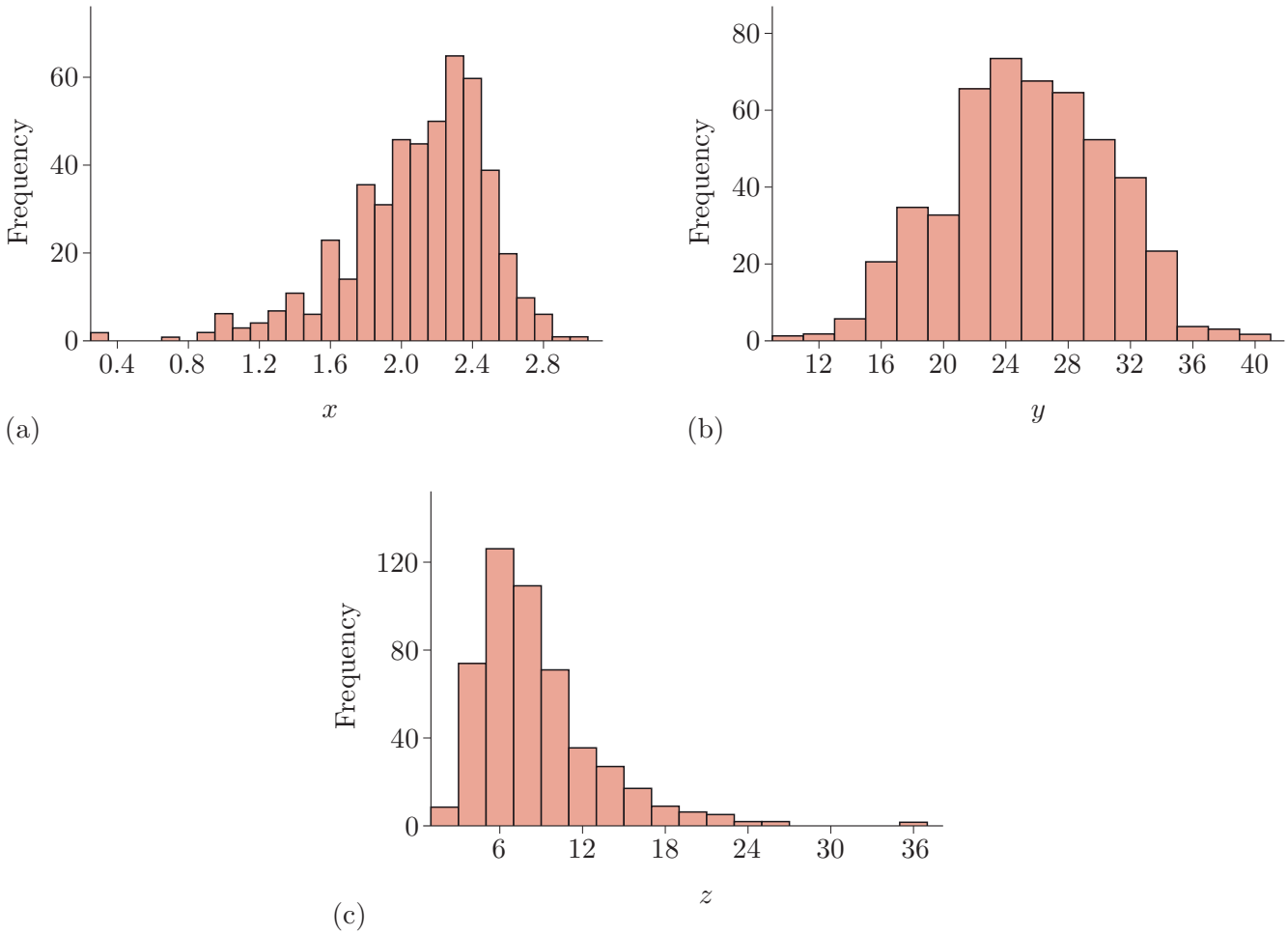


Figure 12 Histograms of three simulated datasets

Activity 4 *Interspike intervals*

In Example 1, it was found that an exponential model is probably not appropriate for the data on interspike intervals. (See the histogram in Figure 2.) In Activity 1, it was argued that a normal model is probably not appropriate for the data on interspike intervals either. This is now confirmed in the normal probability plot of the interspike intervals given in Figure 13: there is a distinct bend in the normal probability plot. In fact, a bend of the type observed in Figure 13 – the points increase first faster than the overall straight line, then slower – is typical when the data reflect a unimodal, right-skew, distribution.

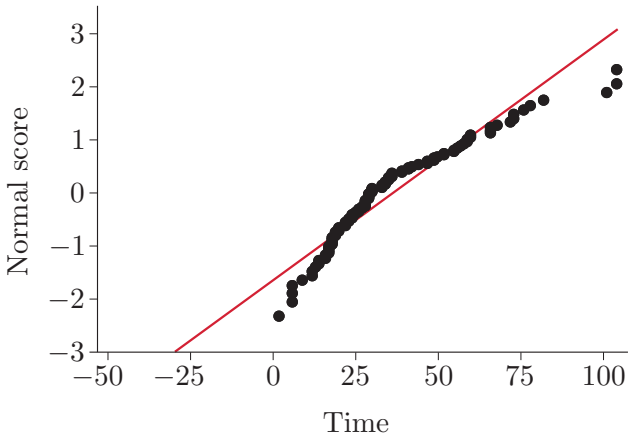


Figure 13 Normal probability plot of interspike intervals

(a) Which transformations from the ladder of powers would you expect to be the most appropriate when attempting to transform the interspike interval data to normality, and why?

(b) Figure 14(a) shows a normal probability plot of the data after a log transformation, and Figure 14(b) shows a normal probability plot of the data after a square root transformation. Comment on the effects of the two transformations. In your view, which transformation has produced the more normally distributed result?

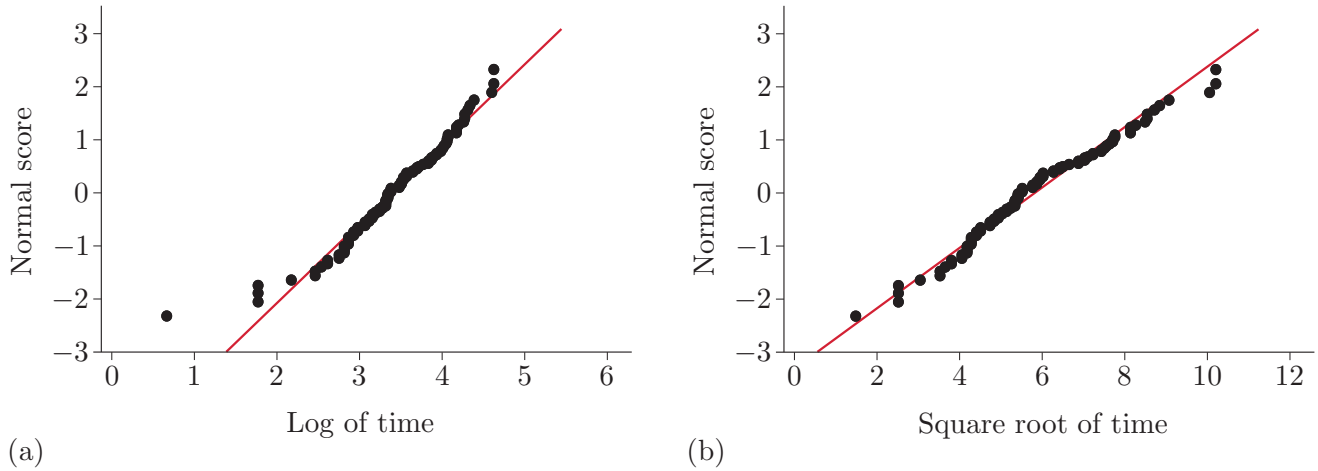


Figure 14 Normal probability plots of interspike intervals: (a) log transformed, (b) square root transformed

It is worth pointing out at this stage that it is not always necessary to obtain a more symmetric or normal distribution! For example, suppose that the aim of the analysis of the interspike intervals data were to calculate the mean interspike interval, together with a 95% confidence interval for the mean. Then, since the sample size is sufficiently large – there are 100 observations – by the Central Limit Theorem, the

The sample mean of the interspike intervals turns out to be 36.49 ms with 95% large-sample confidence interval for the mean of (32.20, 40.78) ms.

distribution of the sample mean is approximately normal, even though the underlying distribution is skew. So large-sample methods can be used to find an approximate 95% confidence interval for the mean. Therefore it is not necessary to transform the data to find a confidence interval for the mean. There may, of course, be other reasons to transform the data.

Exercises on Section 1

Exercise 1 *Validity of powers on the ladder*

Implicit in Subsection 1.2 is the claim that all transformations on the ladder of powers,

$$\ldots,\ x^{-2},\ x^{-1},\ x^{-1/2},\ \log x,\ x^{1/2},\ x^{1},\ x^{2},\ x^{3},\ x^{4},\ \ldots,$$

are valid transformations for positive data in the sense that they are defined and either increasing or decreasing functions of positive x. The former claim is obvious – you can take any power, or a log, of a positive value; the latter claim appears to be true from Figures 6, 8 and 10. By writing any member of the ladder of powers other than log in a unified mathematical form, verify mathematically that every member of the ladder of powers (other than log) is either increasing or decreasing, and identify which are which.

Of course, log is also an increasing function of positive x.

Exercise 2 *Strength of glass fibres*

In Example 2 of Unit 8, a dataset comprising $n = 63$ strengths (in unspecified units) of glass fibres, each of length 1.5 cm, was described. In Unit 8, this dataset was used to introduce the ideas underlying the provision of a large-sample, approximate, confidence interval for the mean strength of such glass fibres. For that purpose, the actual distribution of the glass fibre strengths was not needed (because of the Central Limit Theorem). Suppose now, however, that for some other reason there is interest in the distribution of glass fibre strengths.

(a) A histogram of the data is given in Figure 15. Does a normal distribution appear to be a possible model for these data? If not, why not?

(b) Consider the possibility of transforming the glass fibre data to improve their symmetry and hence potential normality. Which transformations from the ladder of powers would you expect to be the most appropriate when attempting to transform the glass fibre data to normality, and why?

(c) Figure 16(a) shows a normal probability plot of the data. The rest of Figure 16 shows normal probability plots after transforming the data: Figure 16(b) after a square transformation, Figure 16(c) after a cube transformation, and Figure 16(d) after a fourth power transformation. Comment on the effects of the three transformations. In your view, which transformation has produced the most normally distributed result?

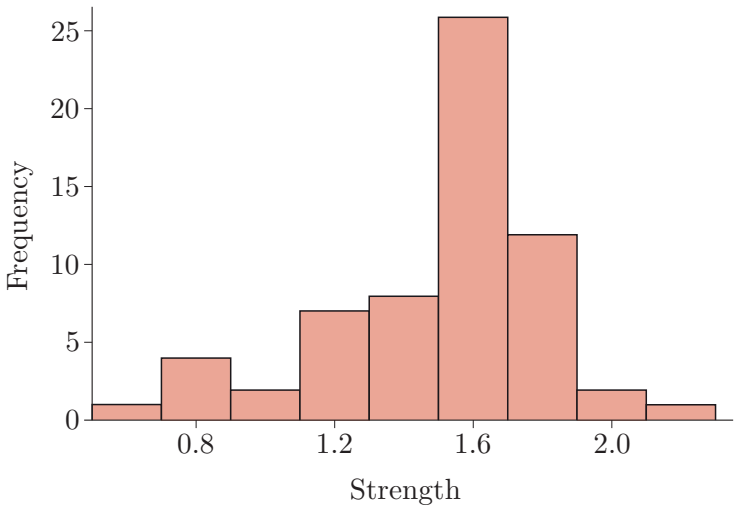


Figure 15 Histogram of glass fibre strengths

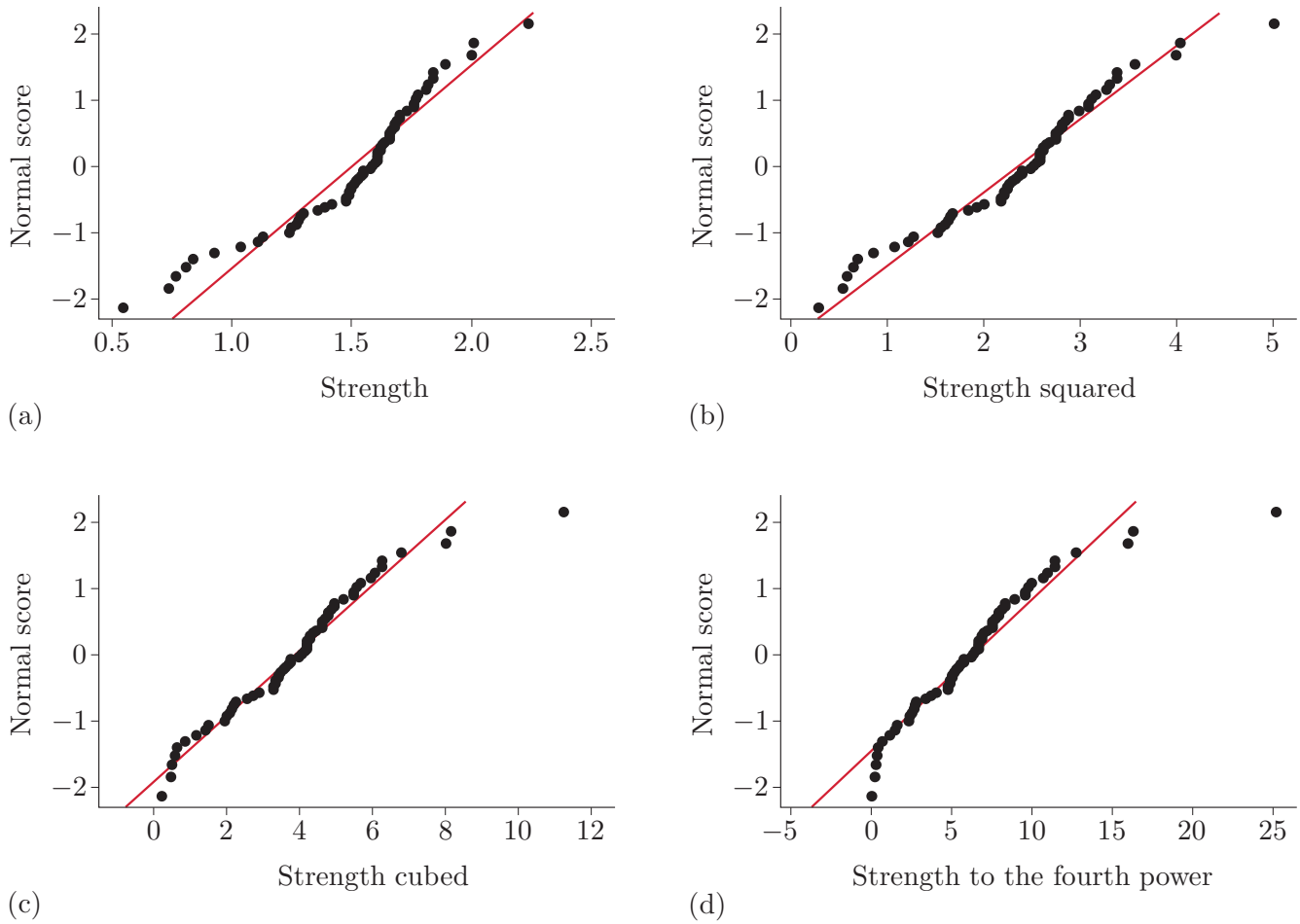


Figure 16 Normal probability plots of glass fibre strengths: (a) untransformed, (b) squared, (c) cubed, (d) taken to the fourth power

2 Transformations in regression

In this section, let us turn our attention back to regression modelling. In Section 1 of Unit 11, you saw that a scatterplot of the data often gives some indication of the relationship between two variables. In particular, when the relationship appears to be linear, the data might be fitted by a simple linear regression model. You then learned, in Sections 2 to 4 of Unit 11, how to fit such a linear model to the data using least squares, how to check the modelling assumptions, and how to perform statistical inference for linear regression models. And you were introduced to multiple regression in Section 5 of Unit 11.

It turns out that transformations can also prove to be very useful – in more ways than one – in regression modelling.

Non-linear to linear? As if by magic ...

First, in some situations, apparently non-linear relationships can be reformulated as linear relationships, and the techniques of Unit 11 applied to them. This is done by transforming the *explanatory variable* and then modelling the dependence of the response variable on the transformed explanatory variable. This is the topic of Subsection 2.1. The core of the work in that subsection is contained in Computer Book C, however.

Second, the other important reason for transforming regression data has to do with the variation of the random terms. For data where the random terms W_i appear to be non-normal, or where the assumption of constant variance of the random terms does not appear to be reasonable, it is sometimes possible to make an appropriate transformation of the *response variable* such that the assumptions seem to be satisfied for the transformed data. This is the topic of Subsection 2.2.

These two uses of transformations in linear regression are summarised in the following box.

In linear regression, it is sometimes possible to:

- straighten out the regression function by **transforming the explanatory variable**
- make the assumptions associated with the random terms conform to those of the linear regression model by **transforming the response variable**.

It is also possible to apply transformations to both explanatory and response variables at the same time, but this will not be investigated in this module. Transformations are also relevant to multiple regression; this is discussed briefly in Subsection 2.3, only for the main focus of that subsection to become the use of multiple regression to solve a specific type of transformation problem in linear regression with one explanatory variable!

2.1 Linear regression on a function of the explanatory variable

We start this subsection with an example of linear regression with one explanatory variable where, by considering the nature of the explanatory variable, the possibility of working with a transformation is raised.

Example 5 *Removing arsenic from drinking water*

This example concerns the results of an experiment investigating how the effectiveness of a method to remove traces of arsenic from drinking water depends on the water's pH, that is, its degree of acidity or alkalinity. Figure 17 shows a scatterplot of the response variable, the percentage of arsenic removed, against the explanatory variable, the pH level, along with the line fitted by least squares. The fitted line has equation

$$y = 190.3 - 18.03\,z, \tag{2}$$

where y is the percentage of arsenic removed and z is the pH level. (You will see why we have used z instead of the usual x in a moment.) Normal probability and residual plots (not shown) suggest that the linear regression model is reasonable for these data.

Filtering through sand to remove arsenic from drinking water in the Red River delta, Vietnam

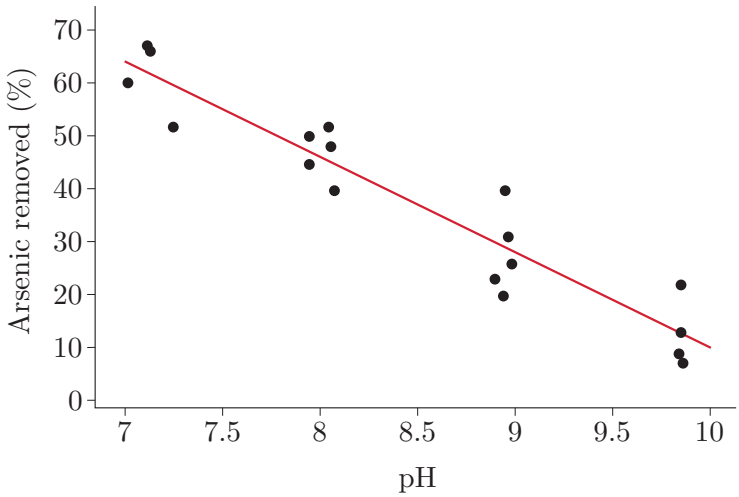


Figure 17 Percentage of arsenic removed, y, against pH, z

(Source: data taken from Devore, J.L. (2014) *Probability and Statistics for Engineering and the Sciences*, 9th edn, Boston, Cengage Learning, p. 490; Devore approximated it from Lytle, D.A., Sorg, T.J. and Snoeyink, V.L. (2005) 'Optimizing arsenic removal during iron removal: theoretical and practical considerations', *Journal of Water Supply Research and Technology – AQUA*, vol. 54, no. 8, pp. 545–60)

Now, values of pH are on a scale running from 1 to 14; the value 7 corresponds to neutrality, and to pure water. Values greater than 7 (as in the experiment reported above) correspond to an alkaline solution; values lower than 7 indicate acidity. This is a standard, internationally agreed, scale for pH, but it is derived from a more basic measured quantity: the pH level is the negative of the logarithm to base 10 of the activity of the

hydrogen ion. It is equally meaningful, therefore, to ask how the percentage of arsenic removed, y, depends (on average) on the activity of the hydrogen ion; call the latter x and note that $z = -\log_{10} x$.

We therefore have a relationship between y and x by putting $z = -\log_{10} x$ in Equation (2):

$$y = 190.3 - 18.03 \times (-\log_{10} x) = 190.3 + 18.03 \log_{10} x. \quad (3)$$

Figure 18 shows a scatterplot of the response variable, the percentage of arsenic removed, against this alternative explanatory variable, the activity of the hydrogen ion, along with the fitted curve given by Equation (3).

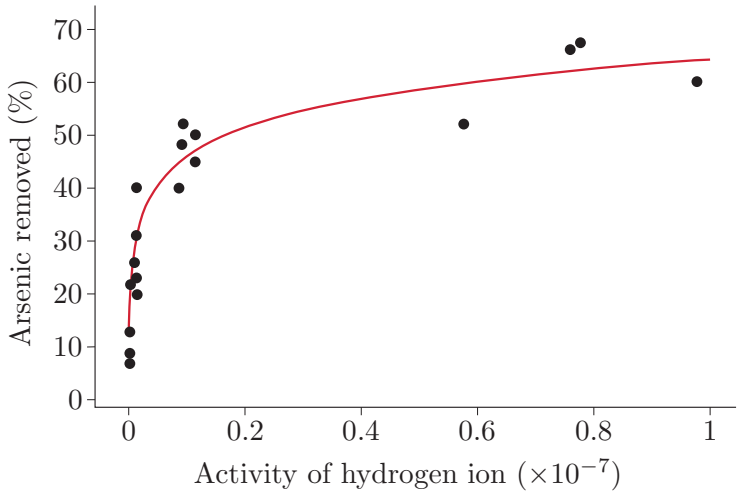


Figure 18 Percentage of arsenic removed, y, against activity of the hydrogen ion, x

Now, Formula (3) is non-linear in x. It therefore seems that we have managed to fit a suitable (very) non-linear relationship between y and x via least squares fitting of a straight-line relationship between y and z! The main point here, therefore, is that there appears to be a straight-line relationship between y and some function, or transformation, of x, namely $-\log_{10} x$.

We could equally well have written $\log_{10} x$ in place of $-\log_{10} x$ here; a straight-line relationship applies to either transformation of x because they are (in this case, very simple) linear functions of one another.

What we have seen in Example 5 is that it is sometimes possible to 'straighten out' or 'linearise' the data by a suitable *transformation of the explanatory variable* so that a linear regression model can be fitted to the transformed data. Then the results from Unit 11 can be used on the transformed data. Explicitly, if h denotes the transformation of x so that $x' = h(x)$, say, then the model for the data becomes

$$Y_i = \alpha + \beta h(x_i) + W_i$$

or equivalently

$$Y_i = \alpha + \beta x'_i + W_i.$$

This model is non-linear in x but linear in the new explanatory variable x'.

Note that the variation in the random terms W_i is not affected by this transformation.

Naturally, results from an analysis made on the transformed data will refer to the transformed data and *not* directly to the original data. You should always take care to present results with reference to the appropriate set of data in order to avoid misunderstandings or misinterpretations.

Finding an appropriate transformation in this context can be very difficult and it often involves a certain amount of trial and error. Statisticians always use a computer to do this, so almost all the activities for this subsection are in Computer Book C. Unfortunately, apart from including many of the transformations that you might wish to try, the ladder of powers of Section 1 is not really directly useful here.

This non-linear river has linearised itself with the creation of an oxbow lake

Refer to Chapter 4 of Computer Book C for the next part of the work in this section.

Two of the datasets introduced in Section 1 of Unit 11 look as though they might be modelled in such a way that they could be treated by linear regression on a suitable function of x, as above. But can they? The first of these, the duckweed data introduced in Example 5 of Unit 11, will be considered in the following example; the second of these, the paper strength data introduced in Activity 2 of Unit 11, will be considered in Subsection 2.3.

Example 6 *More on the model for the duckweed data*

Figure 19 is a repeat of Figure 6 of Unit 11, showing a scatterplot of the data in this case.

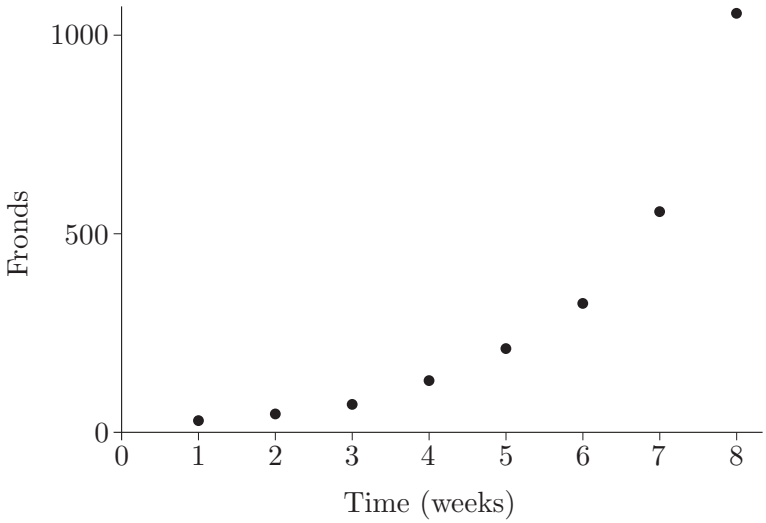


Figure 19 Number of duckweed fronds against time

In Example 7 of Unit 11, it was argued that a possible regression model for the duckweed data might be

$$Y_i = 20e^{\lambda x_i} + W_i,$$

where Y represents the number of duckweed fronds, and x represents the time after the start of the experiment, in weeks. This model is certainly

non-linear in x. However, it appears that if we define $x' = h(x) = 20e^{\lambda x}$, then we would have a linear regression model of the form $Y_i = x'_i + W_i$ (which is, in fact, a linear regression model going through the origin). There is a snag, however. Unlike the transformations of x that you have been considering above and in Computer Book C, this transformation involves the unknown parameter λ. If λ were known to be 1, say, then we could indeed proceed by linear regression on the transformed explanatory variable $x' = 20e^{x}$. But λ is not known and it too needs to be estimated. In such cases, the transformation approach is not available and the model is said to be inherently or intrinsically non-linear. It then needs to be treated by the methods of *non-linear regression*, but these are beyond the scope of this module.

Similarly for any other fixed value of λ.

An important aspect, then, of being able to linearise the data by a suitable transformation of the explanatory variable is that such a transformation should be fully specified and not itself depend on an unknown parameter.

Activity 5 *Which of these regression functions can be linearised?*

Which of the following regression functions can be linearised by employing a suitable transformation so that a linear regression model can be fitted to the transformed data?

$$\alpha + \beta x^3, \quad \alpha + \beta \log(x + \lambda), \quad \alpha + \beta \log\left(\frac{x}{1-x}\right), \quad \beta \exp\left(\mu x + e^{-\gamma x^2}\right).$$

After our brief dalliance with logs to the base 10 in Example 5, we revert to natural logarithms from here on.

2.2 Transforming the response variable

Straightening out a non-linear relationship in order to be able to fit a linear relationship to the data, as in Subsection 2.1, is one of the two main reasons for transforming regression data. The second main reason for transforming regression data is to try to make the assumptions associated with the random terms, W_i, conform to those of the linear regression model – constant, zero mean, constant variance, normality – in situations where they don't. While the method in Subsection 2.1 for linearising the regression function was transformation of the explanatory variable, the method in this subsection for 'normalising' the random terms in the model is *transformation of the response variable*. (In this case, we leave the explanatory variable as it is.)

Example 7 *Fuel efficiency and weight of cars*

For the purposes of a competition associated with the 1983 Annual Meeting of the American Statistical Association, a dataset was compiled on attributes of a number of models of car then in use in the USA. This example concerns two of the variables from that dataset: the response variable is fuel efficiency measured as the number of miles per gallon of

petrol (mpg) achieved by that model of car, and the explanatory variable is its weight in pounds (lb). There are $n = 398$ data points in the dataset. How does fuel efficiency depend on weight?

A 1982 Ford Mustang GL

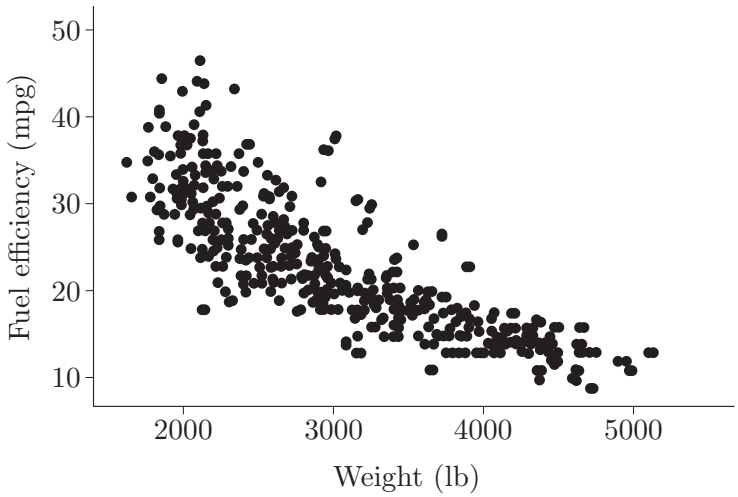


Figure 20 Fuel efficiency against weight

(Source: data of E. Ramos and D. Donoho, extracted from the *Statlib* data archive at Carnegie Mellon University, Pittsburgh, USA)

It is clear from Figure 20 that fuel efficiency decreases as weight increases. The decrease looks a bit non-linear, so maybe a transformation of the explanatory variable, weight, might be considered. However, another, striking feature of Figure 20 is that the amount of variability in the response variable appears not to be constant but also to decrease with increasing weight. A transformation of the response variable, fuel efficiency, might be worth trying, therefore. Figure 21 shows the transformed response variable log(fuel efficiency) plotted against weight. The effect of the transformation is as we might have hoped: the variability in the data appears to be constant (and, as a bonus, the dependence of log(fuel efficiency) on weight appears to be at least approximately linear!).

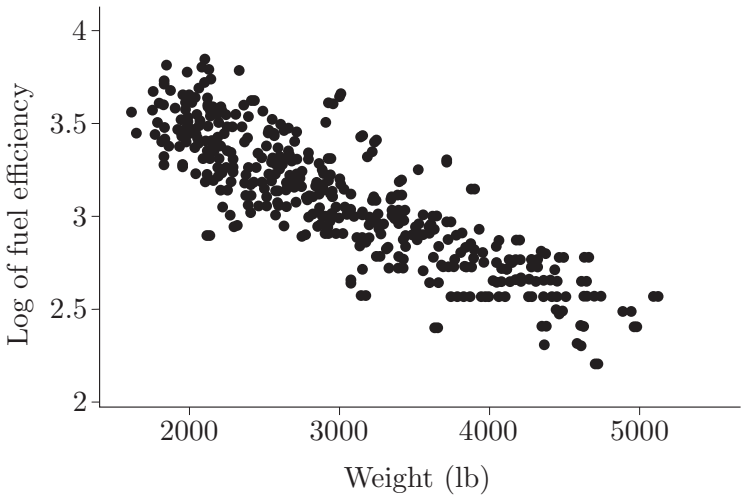


Figure 21 Log(fuel efficiency) against weight

Normal probability and residual plots, shown in Figure 22, suggest that the linear regression model is reasonable for these transformed data. (The only discrepancy from the linear regression model that is evident from these plots is a possible non-normality in the tails of the distribution of residuals in Figure 22(b).)

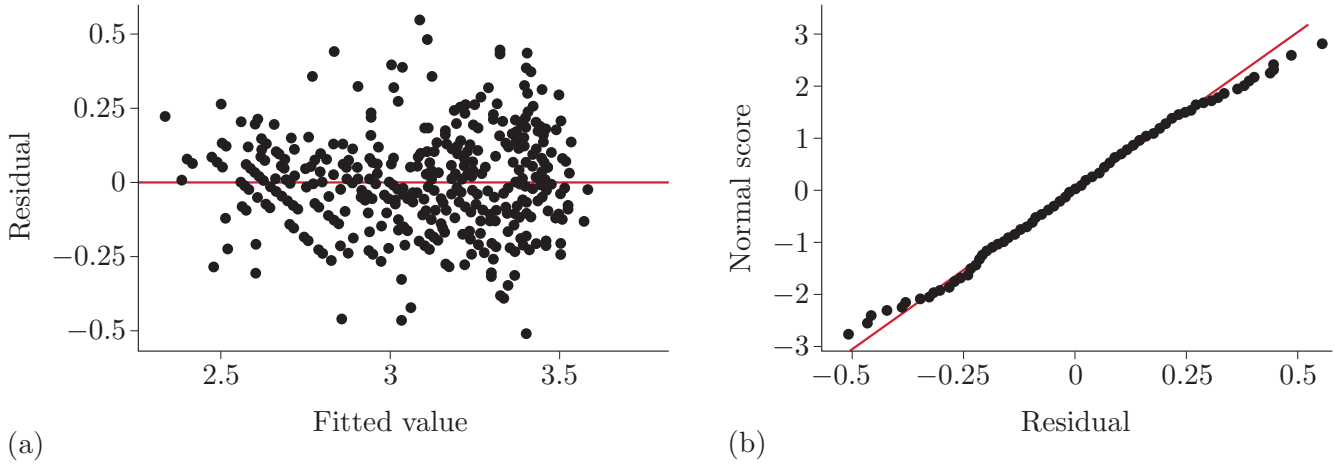


Figure 22 Regression of log(fuel efficiency) on weight: (a) residual plot; (b) normal probability plot of residuals

The line fitted by least squares has been overlaid on the transformed data in Figure 23. The equation of the fitted line is

$$y = 4.1445 - 0.000351\,x,$$

where y is the log of fuel efficiency and x is the weight (in lb). That is,

$$\log(\text{fuel efficiency}) = 4.1445 - 0.000351 \times \text{weight}.$$

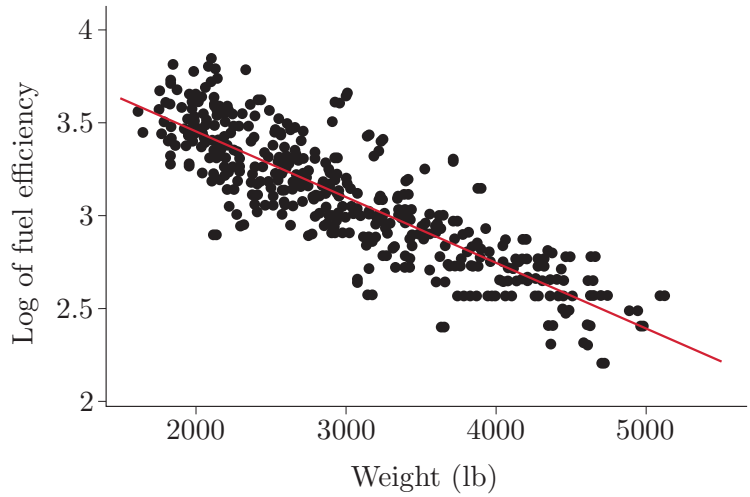


Figure 23 Log(fuel efficiency) against weight, and the least squares line

As in Subsection 2.1 where the explanatory variable was transformed, results from an analysis made on the data with transformed response variable will refer to the transformed data and not the original data. For

example, in this case a unit (pound) increase in weight corresponds (on average) to a reduction in log(fuel efficiency) of 0.000351. Moreover, the following calculation allows us to interpret the effect of weight on the original scale. We have that

$$\begin{aligned}\log(\text{fuel efficiency}) - 0.000351 &= \log(\text{fuel efficiency}) + \log(0.9996)\\ &= \log(0.9996 \times \text{fuel efficiency}),\end{aligned}$$

$e^{-0.000351} = 0.9996.$

the last equality corresponding to use of Equation (1). Therefore a unit (pound) increase in weight corresponds (on average) to a reduction in log(fuel efficiency) of 0.000351, which corresponds in turn to multiplying the actual fuel efficiency by 0.9996; this is a reduction in fuel efficiency of 0.04% for each pound increase in weight.

Other transformations of the response variable are possible too, as in the next example.

Example 8 *Male elephants and their matings*

An eight-year study of the mating behaviour of African elephants was reported in 1989. As part of this study, data were provided on the number of matings (the response variable, y) of each of $n = 41$ male African elephants; this is plotted against the elephant's age (x, in years) in Figure 24. Note that because some of the elephants are of the same age and had the same number of matings, some data points in Figure 24 are jittered vertically, that is, displaced slightly from their true position in order to avoid plotting points on top of one another.

A male African elephant

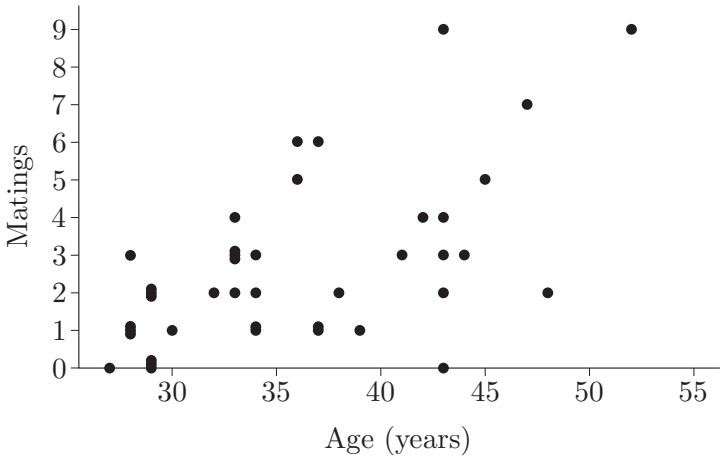


Figure 24 Number of matings against age

(Source: Poole, J.H. (1989) 'Mate guarding, reproductive success and female choice in African elephants', *Animal Behaviour*, vol. 37, no. 5, pp. 842–9)

The number of matings clearly increases as age increases. But as was the case in Figure 20, a feature of Figure 24 is that the amount of variability in the response variable appears not to be constant, in this case increasing with increasing values of the explanatory variable. A transformation of the response variable again seems worth trying, therefore. You could contemplate taking logs, as in Example 7, but there's an immediate

You *could* remedy this by taking $\log(y + a)$ where $a > 0$, but how do you choose a?

problem: some elephants had no matings, and the log of zero is undefined. An alternative transformation that is defined at zero (and increasing for positive x) is the square root transformation, so let's try that. Figure 25 shows the transformed response variable $\sqrt{y}$ plotted against age (again with a little jittering). The effect of the transformation on the scatterplot seems at least to be an improvement: the amount of variability is certainly more constant across ages than it was in Figure 24 (especially if you think of the elephant aged 43 with no matings as an outlier and ignore him), and the increase is plausibly linear.

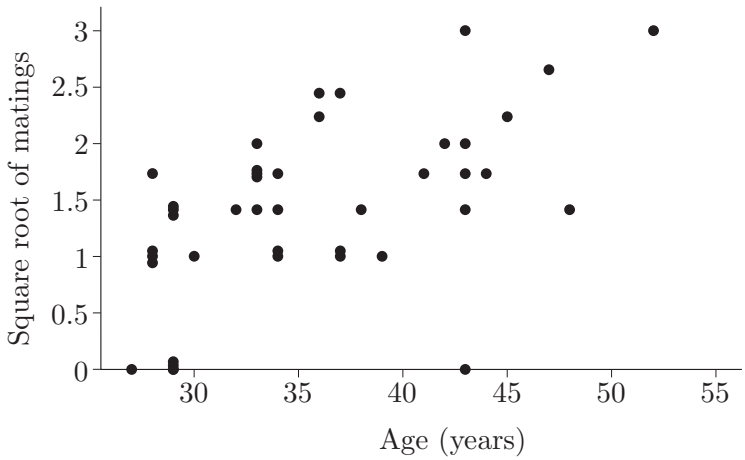


Figure 25 Square root of number of matings against age

A line is fitted by least squares to the data – in original form, no jittering! – and shown on the scatterplot (with jittered points) in Figure 26. In fact, normal probability and residual plots (shown in Figure 27) confirm, in perhaps a slightly surprisingly unequivocal manner, that the linear regression model is reasonable for these transformed data. The fitted line has the formula

$$\sqrt{y} = -0.812 + 0.0632\,x,$$

where y is the number of matings and x is the elephant's age.

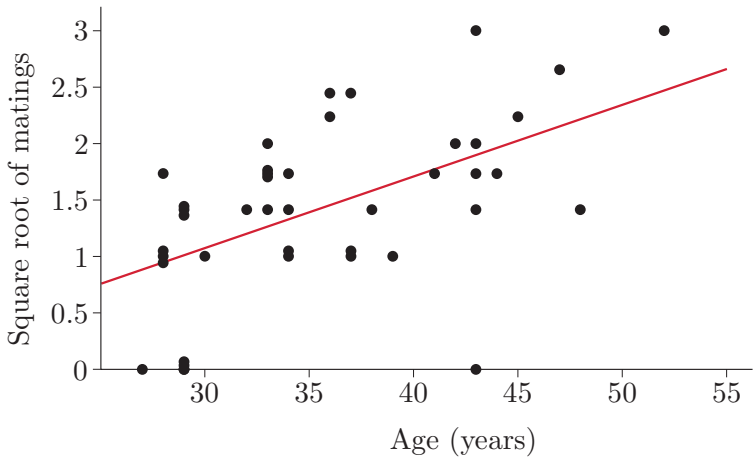


Figure 26 Square root of number of matings against age, and least squares line

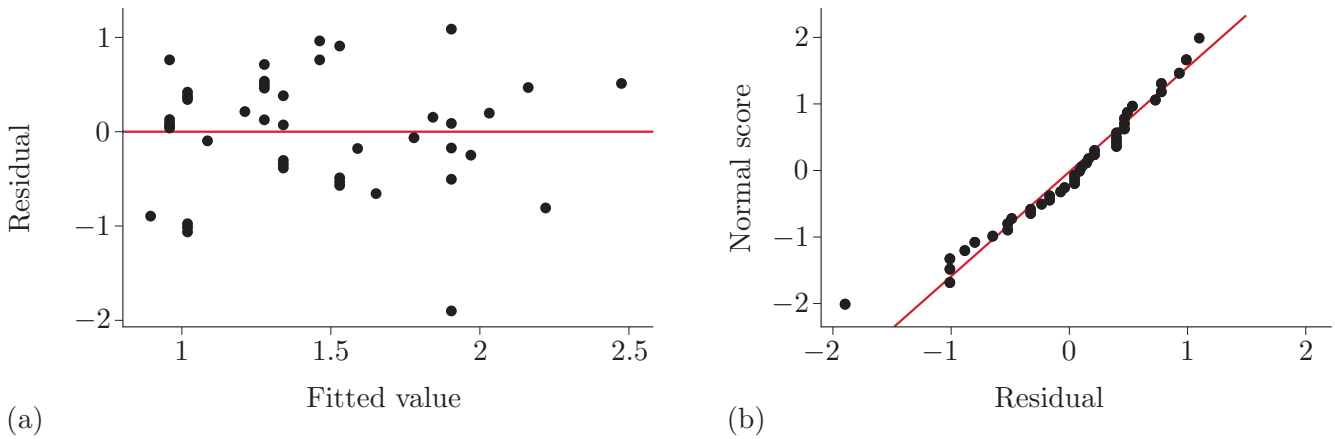


Figure 27 Regression of $\sqrt{\text{matings}}$ on age: (a) residual plot; (b) normal probability plot of residuals

Activity 6 *Predicted number of matings*

Using the model fitted to the data in Example 8, what is the point prediction of the number of matings that might be expected for a male elephant aged 40 years?

It is clear that, as was the case for transformation of the explanatory variable, experimentation with, and comparisons of, different transformations of the response variable might be carried out in practice, with the help of a computer. This can be done in Minitab, but you are not asked to make any such investigations in this subsection.

You might also have an objection to our treatment of Example 8, and you may well be right! The number of matings associated with each elephant is a discrete random variable, not a continuous one. The linear regression model as considered in this unit is designed for use with a continuous response variable. Discrete response variables can be accommodated in a generalisation of the linear regression model that is a topic for a more advanced module. However, in practice, approximating a discrete regression situation by transforming the response and pretending the result is continuous, as we have done in Example 8, is an alternative that remains usefully contained in a statistician's toolbox, at least for occasional use.

'Poisson regression' is one relevant generalisation.

2.3 Multiple regression with transformed variables

So far in this section, we have discussed how and why either the explanatory variable or the response variable can be transformed in linear regression with one explanatory variable. The same principles can be used in multiple regression where any number of the explanatory variables can be transformed (in order to improve the linear dependence on explanatory variables of the regression function) or the response variable can be

Something of this sort might have underlain the decision to take logs of one of the explanatory variables in the economic growth example of Section 5 of Unit 11.

transformed (in order to make the random terms have the properties required in a multiple linear regression model). The multiplicity of explanatory variables makes their transformation – the choice of both which explanatory variable(s) to transform and how to transform each of them – a more difficult problem than in the one explanatory variable case, so we will not investigate it here. On the other hand, transformation of the response variable adds little or no complication over that in the single explanatory variable case and so, for the opposite reason, also won't be investigated here!

There is, however, a problem involving just one explanatory variable which linear regression with one explanatory variable can't cope with, but which *can* be tackled using multiple regression. An example of such a problem is provided by the data on the tensile strength of kraft paper, and its relation to the percentage of hardwood pulp used in its manufacture. These data were first considered in Activity 2 of Unit 11.

Example 9 *A model for the paper strength data*

Rolls of kraft paper in a workshop

Figure 7 of Unit 11 is repeated here as Figure 28. It shows values of the tensile strength of kraft paper (the response variable, y, in units of pounds per square inch, or p.s.i.) plotted against the hardwood content of the pulp from which the paper was made (the explanatory variable, x, in %).

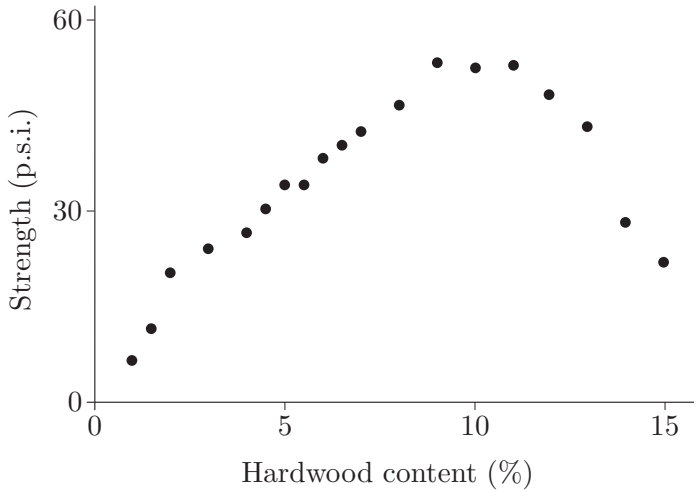


Figure 28 Tensile strength against hardwood content

As was briefly mentioned in the solution to Activity 2 of Unit 11, the up-and-down pattern to these data suggests that a possible model for the data might involve a quadratic function of x or perhaps a cubic function of x. The simpler of these is the quadratic, so let us consider a model like

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + W_i. \qquad (4)$$

A less obvious way of writing the same model is as

$$Y_i = \alpha + \beta(x_i + \delta x_i^2) + W_i,$$

where $\delta = \gamma/\beta$. This makes it clear that if we try to transform the function of x in this model by $x' = x + \delta x^2$ we hit the same problem as in Example 6: the transformation involves an unknown parameter, δ. It appears that non-linear regression is needed.

In this case, however, there is an alternative approach: we can use multiple regression rather than non-linear regression! Look back to Equation (4). We can consider x^2 as a *second* explanatory variable in that regression model, alongside x. Let's set $x_1 = x$ and $x_2 = x^2$. Then the model in Equation (4) is a special case of the multiple regression model

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + W_i,$$

where β and γ in Equation (4) have been renamed β_1 and β_2, respectively. In this case, the two explanatory variables happen to be strongly related ... but there was no requirement in Section 5 of Unit 11 that they should be anything different!

Activity 7 *Models for the paper strength data*

(a) A multiple regression model for the paper strength data was fitted, with the tensile strength of kraft paper as the response variable y, the hardwood content of the pulp from which the paper was made as the first explanatory variable, x_1, and the square of the first explanatory variable as the second explanatory variable, x_2. The fitted model is

$$y = -6.67 + 11.76\, x_1 - 0.6345\, x_2.$$

The residual plot for this model is given in Figure 29.

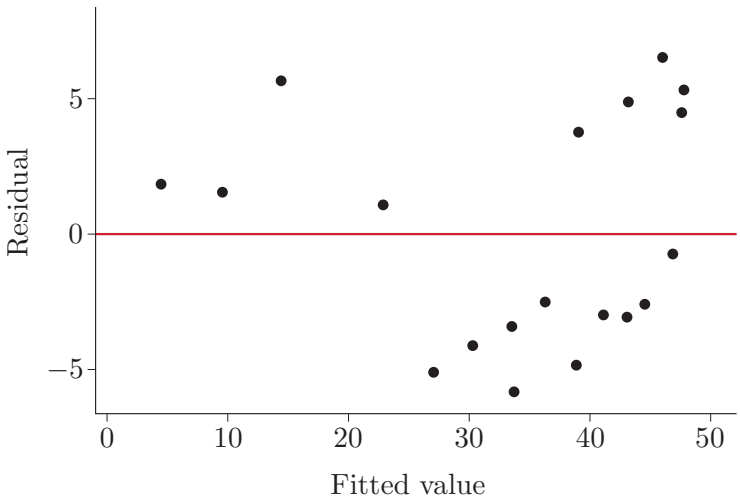


Figure 29 Quadratic model: residual plot

On the basis of the residual plot, do the model assumptions seem reasonable? If not, why not?

(b) Having considered and discarded a quadratic model for the paper strength data in part (a), how about a cubic model instead? In multiple regression terms, the model is of the form

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + W_i,$$

where, in addition to all the ingredients of the multiple regression model using a quadratic curve, $x_{i3} = x_i^3$ represents the cubes of the values of x_i. Such a model was fitted to the data, the fitted model being

$$y = 5.65 + 3.58\,x_1 + 0.654\,x_2 - 0.0552\,x_3.$$

The residual plot and the normal probability plot of the residuals are given for this model in Figure 30.

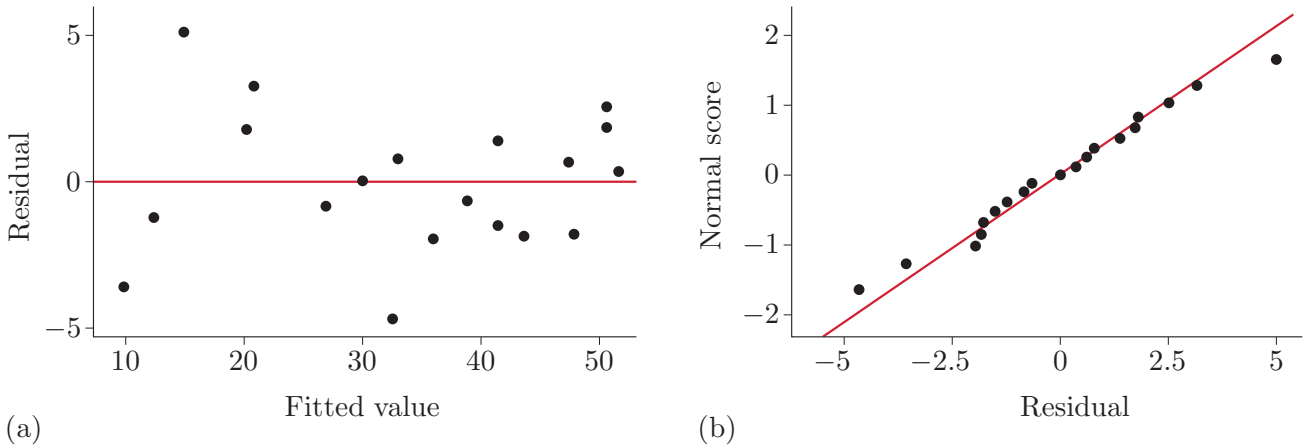


Figure 30 Cubic model: (a) residual plot; (b) normal probability plot of residuals

On the basis of these plots, do the model assumptions seem reasonable? If not, why not?

(c) Suppose that a new batch of kraft paper was to be produced using pulp with a hardwood content of 10%. Using the fitted cubic model, what is your prediction of the tensile strength of that paper?

It is useful to be reminded what has just been achieved in Activity 7. By using multiple *linear* regression, we have managed to fit the *cubic* regression model

$$y = 5.65 + 3.58\,x + 0.654\,x^2 - 0.0552\,x^3,$$

where y is the tensile strength of kraft paper and x is its hardwood content. The non-linearity of the fitted curve is accentuated by plotting it on top of a scatterplot of the data, shown in Figure 31.

Unfortunately, only non-linear functions with a certain specific structure – in fact, those that can be written as linear combinations of functions of x – can be accommodated by multiple linear regression in this way.

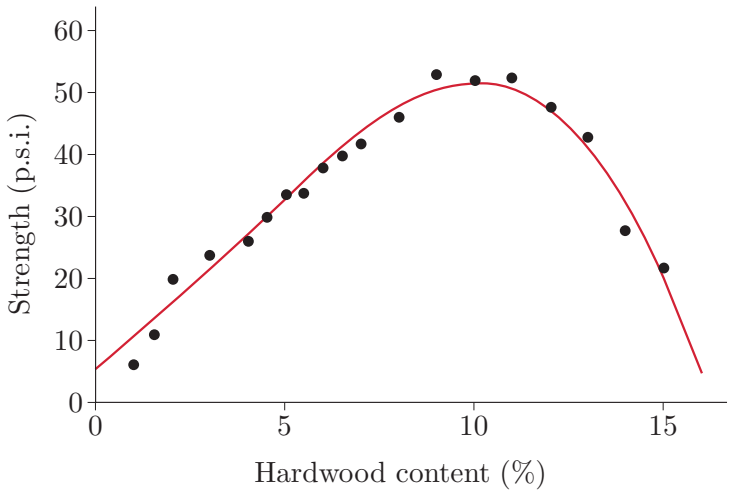


Figure 31 Tensile strength against hardwood content, and fitted cubic regression function

Exercise on Section 2

Exercise 3 *Residual plots and transformations*

Suppose a dataset is made up of measurements of two variables on each of a number of individuals, and we are in the usual regression situation of wishing to model the behaviour of one of the variables, the response variable, as a function of the other, the explanatory variable. Figure 32 is a repeat of Figure 20 of Unit 11. It shows four typical shapes of residual plots produced after fitting a line to such data by least squares.

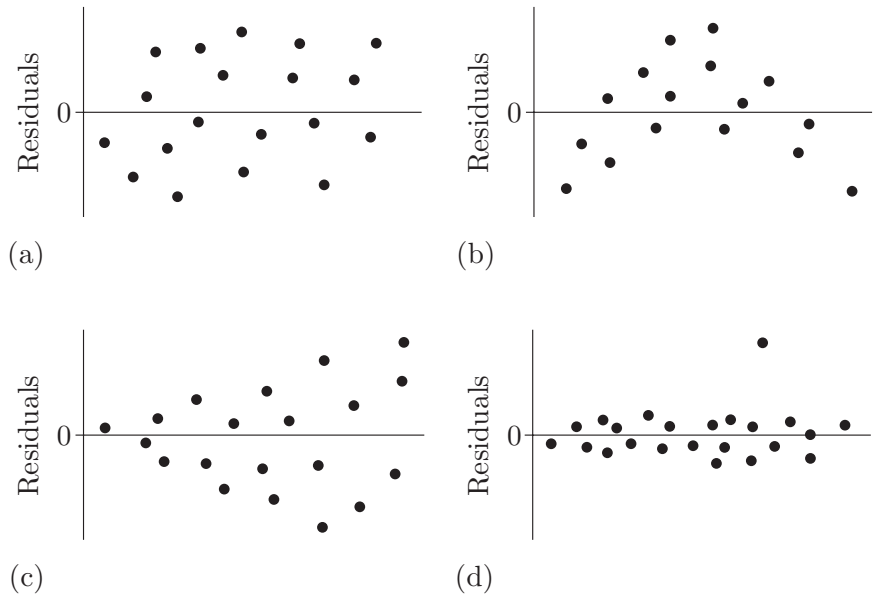


Figure 32 Four residual plots

For each panel of Figure 32 in turn, state whether or not you would transform the data in order to obtain a set-up better aligned with a standard linear regression model, and if you would, which variable you would choose to transform. Give reasons for your answers.

3 The modelling process

The beginning of most statistical investigations is usually a practical problem. For example:

- a medical researcher might want to know whether or not a treatment for cancer works
- an engineer might wish to estimate the tensile strength of a particular material
- a social scientist might seek to understand what factors influence school performance
- an economist might wish to predict future inflation rates.

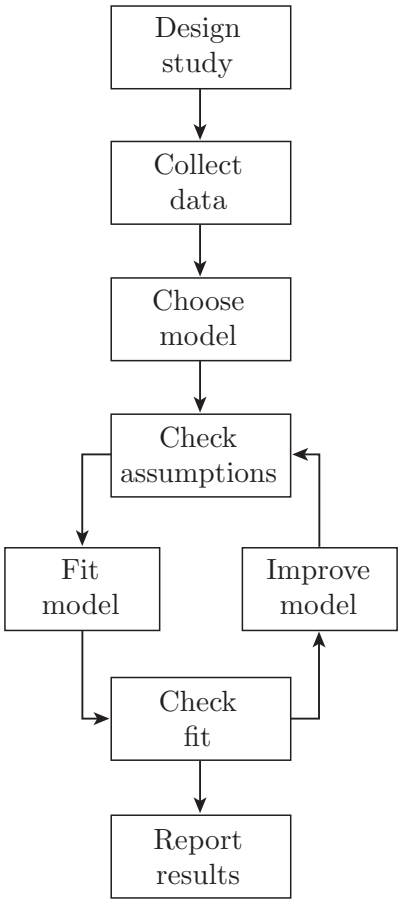


Figure 33 The statistical modelling process

In a statistical investigation, the problem is formulated in statistical terms, appropriate data are collected and analysed, and the conclusions are summarised in a statistical report. The journey from practical problem to statistical report is best thought of as a research process, which can be represented by the flow chart in Figure 33. Note that, in practice, the various stages of the statistical modelling process might arise in a slightly different order from that in Figure 33. For example, it is sometimes more convenient to check assumptions after the model has been fitted.

Formulating the right questions, and designing studies to answer them, are important statistical issues, though they are not dealt with in this module. Typically, they involve collaborations with specialists in other disciplines – medical doctors, engineers, social scientists, economists, and so on – and usually require some knowledge of the particular application area. However, this unit focuses on the issues of statistical modelling and reporting, which are similar in all application areas. The starting point is therefore the problem or question under consideration, together with the data that have been collected to throw light on it. Thus, in terms of the flow chart in Figure 33, you will begin at the box marked 'Choose model'.

What is a model? Well, as you have seen, in general terms, it is a simplified representation of the process generating the data. A key component of a statistical model is the underlying distribution from which the data are sampled; but the model might also include other components – for example, transformations or regression relationships between variables. However, the terms 'distribution' and 'model' are used interchangeably in much of the remainder of this unit, as they have been throughout the module.

A suitable model to start with is one that reflects the most important attributes of the data. For example, if the data consist of measurements on a continuous variable, then it makes sense to choose a continuous distribution to represent the underlying variation. Also, the question to be answered will often suggest how the model will be used – for example, to calculate a confidence interval or to carry out a hypothesis test. Having chosen a model, you will need to check that it fits the data, and that any assumptions required are satisfied. If either is not the case, then you will need to alter the model in some way, or perhaps even try a completely different one. You will then need to repeat the process, improving the model at each stage until it is good enough for its purpose. Having chosen a model, the final stage of the modelling process is to report your results.

The modelling process

Statistical modelling is, to some extent, an art as much as a science, requiring common sense and judgement and, just occasionally, a little inspiration. It is as well to remember that statistical models are at best idealisations of reality: you should not expect to find a 'perfect' model. The real skill is in finding a model that is good enough for your purposes, and from which you can draw valid conclusions.

The remainder of this unit pans out like this. First, in the remainder of this section, we will briefly collect and review *some* of the tools that were put in your statistical toolbox earlier in the module (and hence can be considered as revision, to some extent). However, the focus here is not so much on the individual methods or techniques but on using them to choose the most appropriate method(s) of analysis for the data and questions of interest. The opportunity to put into practice the statistical skills you have acquired will then be provided in a single substantial example in Section 4, while appropriate ways to go about statistical report writing will be described and practised in Section 5.

More practical applications await you in Unit 13.

3.1 Choosing a model: getting started

In this subsection, some guidance is provided about how to approach the task of selecting a model. It is important to remember that there are no fixed rules for this, only general principles – and even they, on occasion, might reasonably be set aside.

Decisions, decisions ...

It is recommended that, as part of the process of choosing a model, you explore the data using graphical methods. The graphical displays first discussed way back in Unit 1 are, therefore, very relevant at this stage of the statistical modelling process. Histograms and, latterly, scatterplots are the types of graph that have been used extensively 'to get a feel for the data' in the module to date; bar charts and boxplots have similar roles, but have been used a little less. Normal probability plots have also been used extensively in the module but not so much at this initial stage of the process; rather, they have been particularly used to contribute strongly to the 'Check fit' stage of the modelling process.

In conjunction with looking at the data using graphical methods, we also wish to emphasise the importance of the setting or context in which the

data are obtained, and the type of data collected, in structuring your ideas. For the purposes of statistical modelling, the first major distinction to be drawn is whether the data should be modelled as being discrete or continuous, as discussed in Units 1 and 2. In many examples, the choice is reasonably clear. Thus even before seeing any data you may be able to narrow down your choice of model to one suited to discrete data or one suited to continuous data. These can then be refined further – or perhaps set aside – after looking at histograms, bar charts or other graphical displays. (The distinction between bar charts and histograms is also, you should recall, that one is appropriate to discrete data, the other to continuous data.)

The distinction between discrete and continuous data is fundamental but it is worth being reminded that the distinction is not always clear-cut. For example, on hearing that one's data consist of counts, a discrete model would be expected. However, arguably, this is premised on the assumption that the counts are low integers $0, 1, 2, 3$, and so on. But what if the counts are typically in the hundreds? The underlying distribution is still discrete, but the data might reasonably be modelled as continuous, as an error of one in hundreds could be considered to be negligible. Conversely, measurements on continuous variables are often rounded to some fixed number of decimal places, and so may be regarded as discrete. If the rounding can be ignored, it is reasonable to treat the rounded measurements as continuous. On the other hand, in some cases, the measurement might be very crude and the data may then best be regarded as discrete. These points are illustrated in Example 10.

Example 10 *Shipwrecks*

In the nineteenth century, there were no fewer than 177 shipwrecks recorded along the stretch of coast from Pevensey to Rye in East Sussex. (Source: Renno, D. (2002) *East Sussex Shipwrecks of the 19th Century (Pevensey-Hastings-Rye)*, Sussex, Book Guild.) Consider the problem of modelling the times between successive shipwrecks, the dates on which they occurred having been recorded. The time between successive shipwrecks is measured in days, so this could arguably be regarded as discrete. However, it could equally be argued that the rounding to the nearest day is immaterial, that many of the intervals between shipwrecks comprise large numbers of days and that, for all practical purposes, the data should be treated as continuous. In situations such as this, the data can be treated either as continuous or as discrete, and both approaches should lead to similar results.

Other important aspects of shipwrecks might be their size, the wrecked vessels ranging from rowing boats to large cargo ships, and the severity of the shipwreck in terms of lives lost and/or goods destroyed. Number of lives lost is clearly a discrete count. Goods destroyed is, in principle, a continuous variable, whether measured in tonnage or financial value. However, there are many reasons why, for such historical events, precise amounts of goods lost (and sometimes even of lives lost) are not known. It might therefore be more reasonable to 'discretise' the continuous variable

'amount of goods destroyed' to, say, one of a small number of approximate round numbers of pounds (or even to discretise lives lost, for some purposes, to a binary variable representing 'some' or 'none').

Example 10 illustrates the point that clear rules even about the apparently simple matter of deciding on a discrete or continuous model can be difficult to specify. When the error involved in treating a continuous variable as discrete, or vice versa, is negligible, then it may not matter which choice is made. In this case, the choice might reasonably be made on the grounds of convenience, and how well the proposed model fits the data.

We, perhaps, claimed an element of this in our treatment of the elephant mating data in Example 8.

3.2 The models at your disposal

In this module, you have met ten families of probability models. They are listed in alphabetical order in Table 1, along with the number of the unit and section or subsection thereof in which they were introduced and primarily discussed.

Table 1 Distributions and where to find them in M248

Distribution	Unit	S(ubs)ection
Bernoulli	3	1.1
binomial	3	2
chi-squared	10	2.2
continuous uniform	3	5.2
discrete uniform	3	5.1
exponential	5	2.2
geometric	3	3.1
normal	6	all
Poisson	3	4
t	8	4.2

Activity 8 *Which are discrete and which are continuous?*

For the ten distributions listed in Table 1, identify which are discrete distributions and which are continuous distributions.

Points made in Subsection 3.1 and the first part of this subsection are summarised below.

Choosing a model: discrete or continuous?

- If the random variable X is discrete, then you might choose a discrete distribution – for example, Bernoulli, binomial, discrete uniform, geometric or Poisson.
- If the random variable X is continuous, then it usually makes sense to choose a continuous distribution – for example, chi-squared, continuous uniform, exponential, normal or t.
- In some circumstances, it is appropriate to model a continuous variable as discrete or a discrete variable as continuous. This is typically the case when the error involved in doing so is negligible.

Having narrowed the field to either discrete models or continuous models, the next step is to choose which of the models in each of these categories is most likely to be suitable. Let us consider discrete distributions first.

3.2.1 Discrete models

The Bernoulli model can be regarded as a special case of the binomial model with $n = 1$. As mentioned when it was introduced in Unit 3, the Bernoulli distribution is the only possible distribution for data taking just

There are many other discrete distributions; these are just the ones you have met in this module.

two values. Thus the choice of discrete models available to you in other discrete situations is really between the binomial, discrete uniform, geometric and Poisson distributions. Choosing between these can be helped by prior understanding, knowledge or intuition about the process generating the random variable X which you wish to model. In particular, the four discrete distributions were introduced – all in Unit 3 – in specific settings; and each setting may be regarded as the standard one for this model. If your data were collected in such a setting, then it makes sense to try the corresponding model.

Choosing a model: standard settings for discrete distributions

- If X may be regarded as the number of successes in some known number n of independent Bernoulli trials with constant probability p of success at each trial, then choose the binomial distribution $B(n, p)$.
- If X has a finite range and every outcome has the same probability, or at least is believed to be equally likely, then try the discrete uniform distribution.
- If X may be regarded as the number of trials up to and including the first success in a sequence of independent Bernoulli trials with constant success probability p, then choose the geometric distribution $G(p)$.
- If X may be regarded as a count of a number of events, then try the Poisson distribution. This applies in particular if events are believed to occur at random and X is the number of events that occur during intervals of fixed length.

These derivations of the models are not guaranteed to apply in practice – in particular, you will need to check the assumptions required in each case. On the other hand, the 'mechanism' underlying the data at hand might be unclear and, for example, a geometric distribution might still prove to be a better model for a certain set of counts than a Poisson distribution, even if the assumptions associated with waiting times in Bernoulli trials are not met. In such circumstances, the distribution must be chosen on *empirical* grounds; that is, your choice of distribution can be guided by the range and shape of the distribution.

Below is a reminder of the ranges and shapes of the binomial, discrete uniform, geometric and Poisson distributions.

Choosing a discrete model: range and shape

- Binomial, $B(n, p)$: finite range $\{0, 1, \ldots, n\}$; one mode, which can take any value within the range (depending on the value of p); symmetric for $p = 1/2$, left-skew for $p > 1/2$, right-skew for $p < 1/2$.
- Discrete uniform: finite range; constant p.m.f.; no mode.

- Geometric, $G(p)$: range $\{1, 2, 3, \ldots\}$, unbounded to the right; decreasing p.m.f. so that its mode is always at 1.
- Poisson(λ): range $\{0, 1, 2, \ldots\}$, unbounded to the right; one mode, which can take any value within the range (depending on the value of λ), including decreasing p.m.f. when $\lambda < 1$.

In some cases, no distribution fits all the requirements. Even then, all may not be lost. First, the models as listed often provide a convenient starting point. Second, it is important to remember that the purpose of statistical modelling is not to find a perfect model, but to find a 'good enough' model. Provided that the model does not fail in some key respect, it might still be useful.

The following quirky example and its associated activity illustrate some of the above considerations.

Example 11 *Reusing reusable envelopes*

In 1990, William Sutherland worked in a large organisation in which internal notes and memoranda were sent in reusable envelopes. This type of envelope is little used these days, since communication via email and other electronic methods has become more convenient; examples of such reusable envelopes are shown in the picture accompanying this example. Here's how they work. Each envelope has a number of spaces (windows) for the names of recipients; the type considered by Dr Sutherland had twelve such windows. New users cross out their own name, and write in the next window the name of the person they wish to contact. Dr Sutherland kept a count of how many names, including his own, were written on some of the envelopes he received. Notice that the first window is always filled. Also, the probability that a window is filled depends on the position of the window: those higher up the envelope are more likely to have been filled than those below (since people usually used up the windows in sequence).

Reusable envelopes of the type considered in Example 11

The purpose of the analysis is to describe the distribution of the number of used windows, and thus to obtain some idea of the age structure of envelopes in circulation.

Activity 9 *Models for numbers of used windows on envelopes*

Consider the setting described in Example 11. The data are counts of the number of used windows on envelopes, ranging from 1 to 12, so clearly they are discrete.

(a) For each of the four discrete distributions in the box above, consider the standard settings underlying each distribution and the range of each distribution to see if one of them might fit this problem.

(b) So far, the discussion of this example has not involved any data, just a rather abstract discussion of the setting. Figure 34 shows a bar chart of the numbers of used windows for a sample of 311 windowed envelopes.

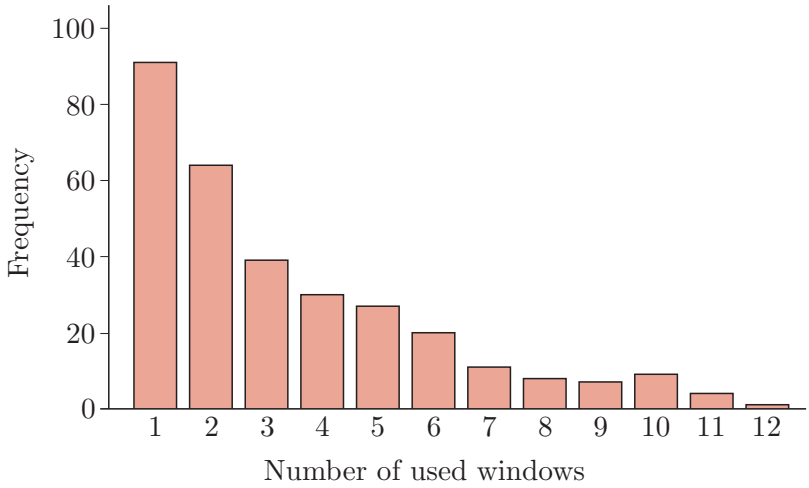


Figure 34 Numbers of used windows

(Source: Sutherland, W. (1990) 'The great pigeonhole in the sky', *New Scientist*, 9 June, vol. 1720, pp. 73–4)

Do you think a discrete uniform model is appropriate for these data? What model (or models) does the shape of the data suggest?

(c) The model preferred in the solution to part (b) can be checked using a chi-squared goodness-of-fit test. This was done after suitable combination of cells with expected frequencies less than 5, yielding a chi-squared test statistic of 13.646 on 9 degrees of freedom, and hence a p-value of 0.135. Does the model fit the data?

It seems that, from Activity 9, a geometric distribution provides a good model for the envelope windows data despite the facts that we failed to identify an underlying Bernoulli process and that the geometric distribution has an unbounded range, whereas the number of used windows can be no greater than 12. This illustrates the point that it is important to keep an open mind when modelling and concentrate on the important aspects of the data, rather than just focus on the limitations of the distribution: in this case, getting the shape right is probably more important than abiding by constraints about the range of the data. What is more, the fitted geometric model happens to give a value for $P(X > 12)$ of 0.016. This is indeed quite a small probability of obtaining values greater than 12. And, in this particular situation, a twist might be that, occasionally, if all the windows on an envelope were full, the envelope would still be re-used, with an extra window or windows drawn on by users. The geometric model might therefore be an even more appropriate model for these data than it appeared to be!

An over-full reused envelope!

3.2.2 Continuous models

Choosing an appropriate continuous distribution follows much the same process as choosing a discrete distribution, with one major exception: if none of the available distributions seems appropriate, then you can try transforming the data (as discussed in Section 1). For the time being, consider the standard continuous distributions you have met so far. These are the chi-squared, continuous uniform, exponential, normal and t-distributions. The chi-squared distribution and t-distribution were introduced specifically in the context of statistical tests. Nothing precludes you from using them for modelling, and they are increasingly used in this way in the modern world, but in this module only the continuous uniform, exponential and normal distributions are used for this purpose.

There are many other continuous distributions; these are just the ones you have met in this module.

As for the discrete distributions above, there are standard settings for these continuous distributions. These can help you to select a candidate distribution.

Choosing a model: standard settings for continuous distributions

- If X takes values between a and b, and each value in the interval $a < x < b$ is equally likely, or believed to be equally likely, then try the continuous uniform distribution $U(a, b)$.
- If events are thought to occur at random in time and X is the waiting time between successive events, then try the exponential distribution.
- The normal distribution is a good first choice when X is clustered around a central value, and is as likely to lie below as above this value. It is also likely to be suitable if you can perceive your data values as being means of other values.

Bringing transformations back into the mix, the box below summarises the ranges and shapes of the continuous distributions available to you. (Apart from the transformation item, the information is the same as in the box at the start of Section 1.)

Choosing a continuous model: range and shape

- Continuous uniform: finite range $a < x < b$; constant p.d.f.; no mode.
- Exponential distribution, $M(\lambda)$: range $0 < x < \infty$, unbounded to the right; decreasing p.d.f.
- Normal distribution, $N(\mu, \sigma^2)$: unbounded range $-\infty < x < \infty$ and is symmetric about a single mode that coincides with the mean; values far from the mean have low probability.
- For data on any range, a suitable transformation of the data *might* be modelled by a normal distribution.

The following activity illustrates some of the above considerations. It also serves to illustrate the important point that models are at best approximate representations of reality. The aim is to formulate a reasonable model, not a perfect one.

Activity 10 *Heredity and head shape*

How similar are the head shapes of these two famous first and second sons, Princes William (right) and Harry (left)?

In a study of 25 families where there were at least two sons, measurements were taken on the head length and head breadth of the first and second sons. Head size can be measured as length + breadth, head shape can be measured by the head shape index calculated as 100 × (breadth/length). One issue of interest is whether there is a difference between the head shapes, as just defined, of first and second sons.

Use some of the above considerations to identify suitable candidate models for the following variables on head size and shape.

(a) Head size, that is, length + breadth.

(b) Head shape, that is, 100 × (breadth/length).

(c) Head shape difference, measured as head shape of first son minus head shape of second son.

Before ending this subsection, it is worth emphasising that there are many more distributions than those covered in detail in this module, although this module does include some of the most important ones in statistics. Minitab provides several other distributions, which you might care to explore (this is entirely optional); and there are many others beside these. However, the approach to selecting an appropriate distribution is much the same whatever the collection of distributions to which you have access.

3.3 Dealing with outliers

You are already aware of the notion of statistical outliers; in this subsection, we seek to give a little more advice about how to deal with them in practice. Outliers are particularly disconcerting if they are found in very small datasets, as in Example 12.

Example 12 *Radiocarbon age determinations*

The data in Table 2 are a set of radiocarbon age determinations, in years, of eight samples from the same stratigraphic layer of a site at Lamoka Lake, New York, USA. The samples should all come from the same period, that of the earliest occupation of the site by ancient people of the Lamoka culture, and so should be, at least approximately, the same age.

Most determinations in Table 2 suggest an age of between 2400 and 2600 years. However, sample C-367 indicates an age of 3433 years, which is quite out of step with the other values: this sample is a clear outlier.

Table 2 Radiocarbon dating

Sample number	Radiocarbon age determination
C-288	2419
M-26	2485
C-367	3433
M-195	2575
M-911	2521
M-912	2451
Y-1279	2550
Y-1280	2540

(Source: Long, A. and Rippeteau, B. (1974) 'Testing contemporaneity and averaging radiocarbon dates', *American Antiquity*, vol. 39, no. 2, pp. 205–15)

If at all possible, when outliers are present, your first step should be to check that they are not the result of recording, coding or data entry errors. Such errors are very common. It is well worth repeating here that if you enter your own data, you should always check your computer data file against the original. Not all data entry errors will necessarily appear as outliers!

The study of outliers and how to treat them can be rather complex, so only a little general guidance will be given in this module. Broadly speaking, the treatment of outliers depends on how many appear in the data, what effect they have on the conclusions, and how far you are prepared to go in believing that you have been unlucky enough to obtain a few 'atypical' values, rather than believing that the distributional assumptions are not viable. This last point is important: the outliers might just reflect the fact that you have chosen the 'wrong' model. The effect of model choice on outliers is illustrated in Example 13.

We offer no specific advice about when a data point is sufficiently unusual to be classified as an outlier; as with much of statistics, this question is not clear-cut, its answer depending on context and purpose.

Activity 11 *Treatment duration*

The duration (in days) of treatment was recorded for a set of 86 long-term hospital patients. Treatment can last several months, so it makes sense to model these data as continuous; they are also positive. But which model is appropriate? Despite the positivity of the data, the normal model might be a good one, if there is a 'typical' treatment length around which the values cluster and, as suggested above, treatment durations tend to be long (corresponding to a normal model with negligible probability of negative durations). Alternatively, you could think of treatment durations as waiting times, which might suggest that an exponential model may be suitable.

A hospital main entrance: the way out as well as the way in!

Let us now consider the actual data. Figure 35 (overleaf) shows a frequency histogram of treatment times; the sample mean of the data is 122.3 days and the sample standard deviation is 146.7 days.

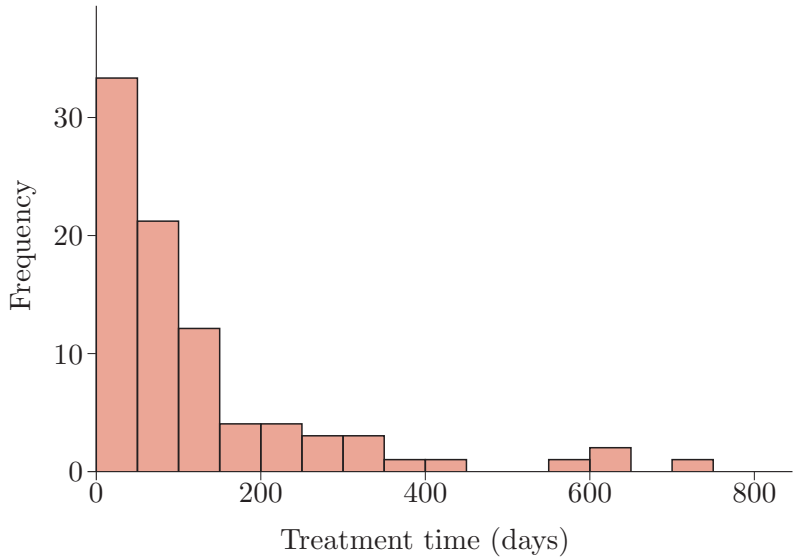


Figure 35 Treatment times for hospital patients

(Source: Copas, J.B. and Fryer, M.J. (1980) 'Density estimation and suicide risks in psychiatric treatment', *Journal of the Royal Statistical Society, Series A*, vol. 143, no. 2, pp. 167–76)

(a) On the basis of the histogram, do you think that an exponential model is likely to be a good model for these data?

(b) In Subsection 2.2 of Unit 5, it was noted that the mean and standard deviation of the exponential distribution are equal, which 'provides a method for checking quickly, given data, whether an exponential model is worth considering'. On the basis of the given sample statistics, do you think that an exponential model is likely to be a good model for these data?

Example 13 *Treatment duration, continued*

In this example, we take consideration of the treatment duration data introduced in Activity 11 somewhat further. From the histogram in Figure 35 and by looking at the mean–standard deviation relationship, you may well have concluded that an exponential model might be a reasonable one for these data. Using the reciprocal of the sample mean (which is the maximum likelihood estimate) as an estimate of the exponential parameter λ leads to the fitted exponential distribution superimposed on top of the unit-area version of the histogram of Figure 35, which is shown in Figure 36.

Arguably, this figure suggests that the exponential distribution fits the main body of the data very well, but there are four data values – those making up the two little bumps to the right in the histogram – that are not in line with this distribution. (Further analysis using tools that you have not met in M248 makes this observation even clearer.) So, relative to

the behaviour expected under an exponential model, it seems that there may be four outliers in this dataset. These outliers could be interpreted as relating to an atypical group of patients with unusually long treatment times.

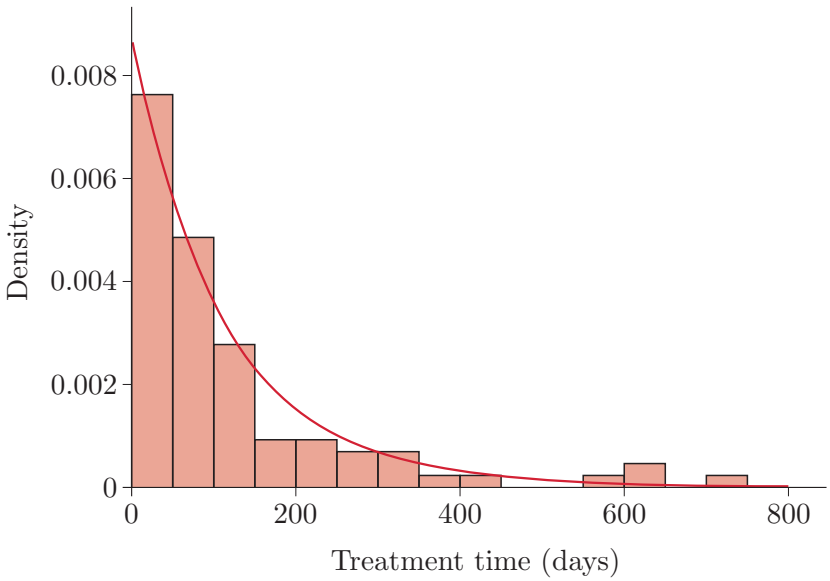


Figure 36 Unit-area histogram of treatment times with fitted exponential distribution

However, if the transformation $x^{1/4}$ is applied to the data, then the corresponding frequency histogram, which is shown in Figure 37(a), is roughly symmetric; and the normal probability plot in Figure 37(b) suggests that a normal model might be appropriate for the transformed data.

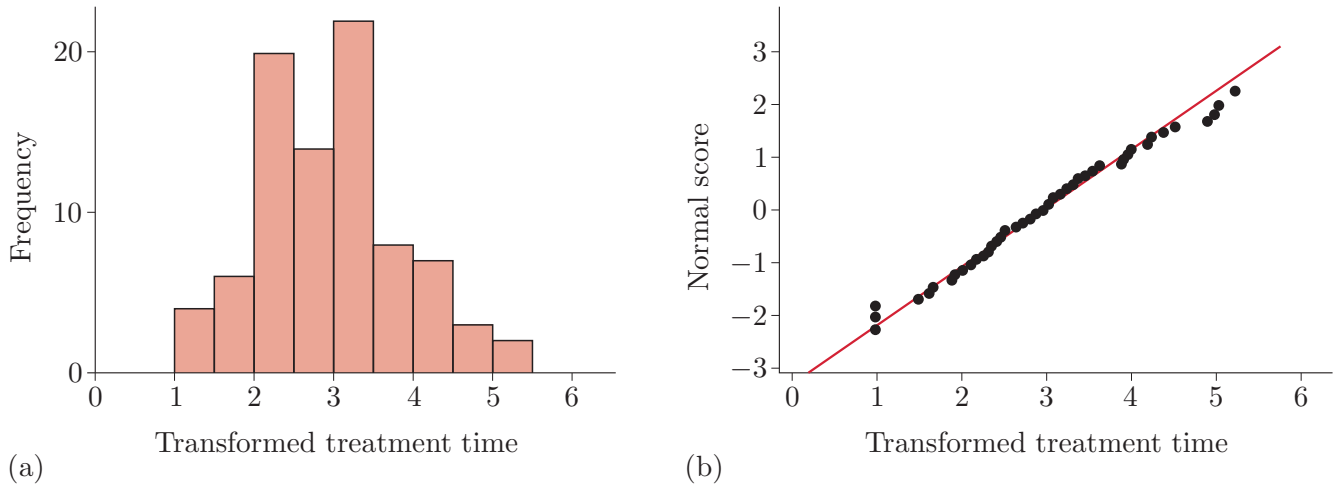


Figure 37 Transformed data: (a) histogram (b) normal probability plot

There is no strong suggestion of a problem at longer treatment times now! There are three ill-fitting points with treatment times of one day. Since times are counted in whole days, it is possible that some of the patients to which these times relate were treated for a slightly longer or slightly shorter time than one day. The poor fit in the lower tail might thus just be the result of crude rounding of the data.

So which model is 'correct'? Are the untransformed data exponentially distributed, with a group of four atypical patients with long treatment times? Or are the transformed data normally distributed? It is not possible to settle the issue without further information, in particular about the four possibly atypical patients, or a lot more data. As it stands, both models for analysis are equally reasonable. In any case, which model to use would depend on the purpose of the analysis.

An important consideration in dealing with outliers is the purpose of the analysis. If there are relatively few outliers, and the model you have selected appears to fit the rest of the data reasonably well, then a sensible procedure is to examine the sensitivity of your conclusions to the outliers. This is easily done by undertaking two analyses, one including the outliers and the other excluding them. If your conclusions do not differ substantially under the two procedures, then the outliers are not influential and should not be a major source of concern. In this case, you should report the presence of the outliers, and state that the conclusion reached is not sensitive to inclusion or exclusion of these outliers. On the other hand, if the outliers do have a big impact on the conclusions, then it can be appropriate to report your findings with the data analysed both ways.

Lamoka Lake

Activity 12 *Age of the Lamoka Lake site*

The radiocarbon data of Example 12 are to be used to provide an estimate of the age of the Lamoka Lake site. Provide such an estimate using all of the data points, and investigate its sensitivity to the outlier. How would you report your results?

There are no hard and fast rules to decide how many data values may be deleted in order to salvage a particular modelling assumption. In practice, it is best to remove only a very few values.

If in doubt, an alternative is to keep all the values and revert to a distribution-free method. Using ranks instead of data values loses information about how far apart the values are but, on the other hand, it removes sensitivity to abnormally large or abnormally small values.

If decisions about which method to use seem unduly vague, you should remember that there is not always a definitely right or wrong way of performing a statistical analysis. All you can do is use your common sense.

4 Modelling with Minitab

In this section, you will have the opportunity to tackle an exercise, or mini-project, in statistical modelling using Minitab. The aim is to give you practice in the skills of statistical modelling, pulling together a number of things that you have learned about in the module and been reminded about in this unit. The mini-project begins with some background, a scientific question and a description of a dataset relevant to the question. You will then progress through the various stages of exploratory analysis, model and method choice, model checking, and performing relevant statistical calculations.

Refer to Chapter 5 of Computer Book C for the work in this section.

5 Writing a statistical report

A statistical investigation usually begins with a practical problem and ends with the results being summarised in a statistical report. The stages involved were discussed briefly in Section 3 and represented in a flow chart in Figure 33. In this section, some general advice is given on how to write – and how not to write – a statistical report.

The statistical report is an account of what you did, why you did it, what you found, and what your results mean in terms of the original scientific question. The main challenge in writing a report is that it is aimed at two very different readerships. First, it should be sufficiently detailed to allow other statisticians to understand clearly what you did, and enable them to assess the validity of your conclusions. But it is equally important that it should provide a non-technical account of your investigation to non-statisticians who are interested primarily in the original question. These distinct aims are usually reconciled by writing a non-technical summary of your investigation to accompany the more technical report.

The report is also important for your own record.

It is also important to stress that the report should be succinct and, if possible, short. A long-winded, rambling document is of little use to anyone!

In Subsection 5.1, the structure of a typical statistical report is described, and in Subsection 5.2 an example is discussed in detail. You will then have the opportunity to assemble some statistical reports yourself.

5.1 The structure of a statistical report

The key to a good report is its structure. This makes for easy reading. For example, a non-specialist might ignore the more technical parts of the report dealing with statistical methods, and read only the summary

and the discussion. Also, structure makes a report easier to write, as it helps to organise the material.

A possible structure for a statistical report is set out in the following box.

The structure of a statistical report

A statistical report comprises the following sections.

- Summary
- Introduction
- Methods
- Results
- Discussion

This structure is reasonably standard, though some authors might use different section headings – for example, *Abstract* instead of *Summary*, *Background* instead of *Introduction*, *Conclusions* instead of *Discussion*, and so on.

The *Summary* should be completely self-contained. It should state briefly the aim of the analysis, the method used, the key finding or findings, and the interpretation. It is usually written last, and should use largely non-technical language. The 'largely' in the previous sentence is a reflection of the fact that it is often simply not possible to provide an accurate summary of results without using some statistical terminology or referring to some statistical concepts. It is far better to give a slightly technical, but correct, summary than one apparently easily understood by all, but potentially misleading.

The *Introduction* should contain a brief description of the question or hypothesis to be investigated, the setting in which the data were collected, and the data available. Note that, in this module, the starting point is always a problem or question, and some data relevant to that problem or question.

A rather beautiful blind alley in Florence, Italy

The *Methods* section should include a description of the model, the procedures used to check the model, the statistical tests employed, the method used for calculating confidence intervals, and any other relevant techniques you have used, such as data transformations. The key guide to this section is to include enough detail to allow other statisticians to evaluate your method, and to repeat your investigation if they had the same data. You should not include all the blind alleys and dead ends you travelled (we all travel them) before settling on your preferred solution. However, if you found two equally plausible models that give appreciably different results, then you should include both.

The *Results* section should contain descriptive summaries of your data (for example, graphical and numerical summaries), evidence that your model is appropriate and, finally, the numerical results of statistical tests or confidence interval calculations. Sometimes, it might be appropriate to

round your numerical results further when reporting them in the *Results* section. It is important to remember that this section, as all others, should be written in prose: a collection of numbers and graphs is not sufficient.

The *Discussion* should contain your own assessment of the statistical evidence relating to the original question or hypothesis. In particular, you should discuss any evidence of lack of fit of your model, any problems with the data (for example, outliers), or any other matter that might have a bearing on the interpretation of the results.

There is no set order in which to write the sections of a report but you should present the sections in the order just described. Many readers will not read all the sections – for example, many will read only the *Introduction* and *Discussion* – so it is important to structure your report so that they can find the sections they are interested in quickly. In some sense, the *Results* section forms the heart of the report. The *Methods* section is organised in such a way as to explain how you obtained your results, while the *Discussion* is your interpretation of the results. Some authors prefer to write the *Results* section first, followed by the *Methods* section. You should use whatever order you feel most comfortable with. In any case, you will probably find yourself going back over previous sections to make sure everything fits together in a coherent whole.

Finally, one important general rule: the shorter, the better. If you can describe something accurately in one sentence rather than two, then so much the better! (But, of course, two short sentences are better than one long rambling sentence.)

Activity 13 *Organising the report*

The following are a few notes from a statistical analysis that you wish to write up as a statistical report. It includes one hypothesis test that you might not have carried out yourself because we have not discussed the details of it in this module. However, this should be no barrier to carrying out the task required in this activity. Organise the material into an outline report under the headings *Introduction*, *Methods*, *Results*, *Discussion*.

> Two groups, continuous variable. Checked that sample variances are similar. Did two-sample t-test, $p = 0.16$. Normality seemed OK in each group (probability plot). 95% two-sample t-interval $(-3.92, 17.63)$. Conclude means could be the same for both groups. Sample sizes 24 and 32.

Note that you are not expected to write the report, just to reorganise the information in the sentences or parts of sentences under the four headings.

5.2 Writing the report

In this subsection, an analysis of the data on head shape that were introduced in Activity 10 is used to illustrate what might be included in a

statistical report. Only the comparison of head shapes of first and second sons will be considered. Example 14 provides a brief account of the analysis of these data.

Example 14 *Head shapes of first and second sons*

Head measurements were taken for the first and second sons from 25 families. For each son, the head shape index was calculated as head breadth divided by head length (both in the same units), multiplied by 100. It is assumed that this index does not vary substantially over the age range of the data. The question is whether there is any difference in shape indices between first and second sons.

The data are paired, so it makes sense to look at the differences between the head shape indices of first and second sons. The mean difference is 0.19, so it appears that the head shape index is slightly greater for first sons than for second sons. In Activity 10, you may have suggested that a normal model might be appropriate for the differences. A histogram and a normal probability plot of the differences are shown in Figure 38.

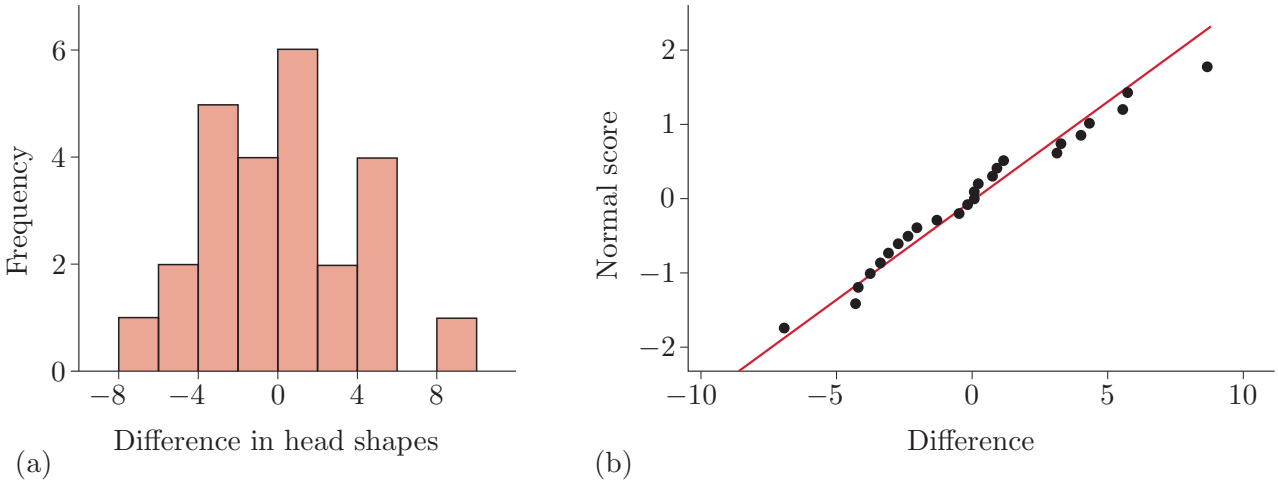


Figure 38 Head shape differences: (a) histogram; (b) normal probability plot

(Source: Frets, G.P. (1921) 'Heredity of head form in man', *Genetica*, vol. 3, no. 3, pp. 193–384)

The plots in Figure 38 both serve to confirm the reasonable validity of the normal model.

The next step is to calculate a t-interval for the mean difference: the 95% t-interval is $(-1.35, 1.72)$. Finally, a t-test of the null hypothesis of zero mean difference might be carried out. This gives $t = 0.25$ on 24 degrees of freedom; the p-value for the test is 0.803. There were no problems with outliers.

Example 15 *Writing the report*

The information in Example 14 can be organised into a statistical report as follows.

The *Introduction* states the problem and describes the data available, including the source of the data. You should not include extraneous material such as theories of heredity or genetics here (though, of course, the scientists who gave you the data might wish to do so in their report – but that is up to them). So the following (taken from the description of Activity 10 as well as Example 14) is adequate.

> *Introduction*
> In 25 families where there were at least two sons, measurements were taken on the head length and head breadth (both measured in mm) of the first and second sons. The head shape index is defined as $100 \times$ (breadth/length). The issue addressed in this report is whether there is a difference between the head shapes of first and second sons. The data for this analysis were taken from Frets, G.P. (1921) 'Heredity of head form in man', *Genetica*, vol. 3, no. 3, pp. 193–384.

The next step is to write the *Methods* section. You should state the variables and describe the methods that you used to reach your conclusions, but not all the blind alleys that you might have explored in the process! For example, a normal model was chosen, so it should be mentioned that a probability plot was used to check the adequacy of the model. This information is required so that if other statisticians read your results, they will know that the model is valid. However, if you originally thought you might use, say, an exponential model, but dropped the idea once you looked at the data, this information should not go in the *Methods* section. Readers are not interested in your thought processes, or all the mistakes you might have made along the way, but simply want to know how you obtained your results, and whether your methods were appropriate.

The *Methods* section is generally aimed at a statistical readership, and hence you can quote standard methods without describing them in any detail. For example, it is perfectly appropriate to say '95% t-intervals were calculated' without explaining what a confidence interval is or what the t-distribution is. Indeed, you should most definitely *not* describe what they are! Finally, it is often useful to include details of the software you used, usually the name and version. Here is a suggested *Methods* section for the report on head shapes.

> *Methods*
> As the data are paired, the analysis was based on differences between the head shape indices of first sons and second sons. A normal model was used for the differences; this assumption was checked using a normal probability plot. A 95% t-interval was calculated and a t-test was used to test the hypothesis that the mean difference is zero. All analyses were performed using Minitab Version 17.

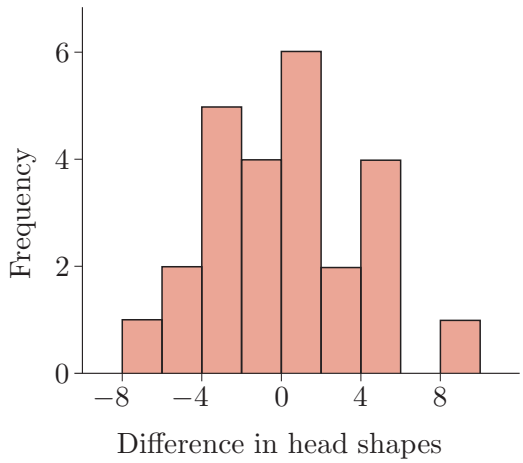
Next comes the *Results* section. A general rule is to separate the results into descriptive summaries and analytic results. Descriptive summaries might include some relevant numerical summaries (for example, median and interquartile range) or graphs. The aim is to convey some feel for the data. However, you should beware of including too many descriptive summaries: the aim is to highlight aspects of the data that are relevant to

the question of interest. For example, for the head shape data, you might include a boxplot or a histogram of the differences in head shape indices; one such graph is enough here.

The analytic results include those that directly address the original question set out in the *Introduction*. The original question relates to the difference between the head shapes of first and second sons. Thus you should report the mean difference and the 95% t-interval. In addition, you should report the result of the t-test. Finally, you need to provide some evidence that your methods are justified. In this case, a probability plot was used to test the normality assumption. It is not essential to show this plot. To save space, it is quite reasonable simply to state that you used this method to check the assumption.

Results

The distribution of the 25 differences between the head shape indices of first and second sons is shown in the histogram below. The data were approximately normally distributed, as confirmed by a probability plot (not shown).



The mean difference in head shape indices was 0.19, with 95% t-interval $(-1.35, 1.72)$. A t-test of the hypothesis of zero mean difference gave $t = 0.25$ on 24 degrees of freedom, with p-value of 0.803.

The next section is the *Discussion* section, in which you give your interpretation of the results in the light of the original question. This is also the place where you should comment on the possible impact of any other factors (such as missing data or outliers) on the interpretation. In this example there are no such factors. The section can thus be suitably brief: there is no evidence of a difference. However, it is worth qualifying this conclusion by reminding the reader that the sample size was rather small.

In general, it is important to write concisely and to the point.

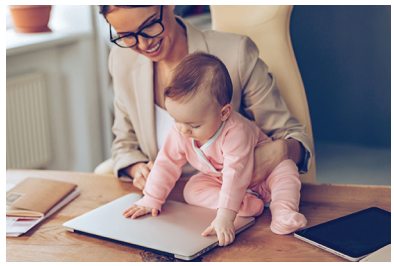
Discussion

We conclude that there is little or no evidence against the hypothesis of no difference between the head shape indices of first and second sons. However, the sample size for this study was quite small, being only 25.

Finally, having assembled and re-read the report, you can now write the *Summary*. This states briefly the purpose of the analysis, the method used, the key finding and its interpretation. It should be largely non-technical.

> *Summary*
> The aim of this analysis was to compare head shapes of first and second sons, using a head shape index based on the ratio of head breadth to head length. Data on 25 pairs of first and second sons were obtained from a published source and analysed using a normal model. We found no significant difference between the head shapes of first and second sons.

This completes the report. The final step is to read through the report and check it.



Yes, mum, you've finished your report ...

The sections of the report on head shapes of first and second sons that were written in Example 15 are assembled in the following box.

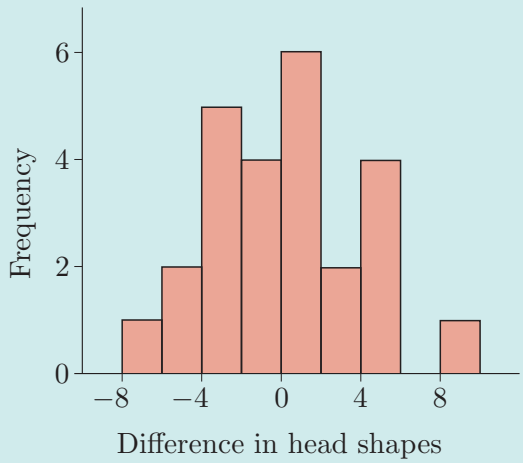
A complete statistical report

Summary
The aim of this analysis was to compare head shapes of first and second sons, using a head shape index based on the ratio of head breadth to head length. Data on 25 pairs of first and second sons were obtained from a published source and analysed using a normal model. We found no significant difference between the head shapes of first and second sons.

Introduction
In 25 families where there were at least two sons, measurements were taken on the head length and head breadth (both measured in mm) of the first and second sons. The head shape index is defined as $100 \times$ (breadth/length). The issue addressed in this report is whether there is a difference between the head shapes of first and second sons. The data for this analysis were taken from Frets, G.P. (1921) 'Heredity of head form in man', *Genetica*, vol. 3, no. 3, pp. 193–384.

Methods
As the data are paired, the analysis was based on differences between the head shape indices of first sons and second sons. A normal model was used for the differences; this assumption was checked using a normal probability plot. A 95% t-interval was calculated and a t-test was used to test the hypothesis that the mean difference is zero. All analyses were performed using Minitab Version 17.

Results
The distribution of the 25 differences between the head shape indices of first and second sons is shown in the histogram below. The data were approximately normally distributed, as confirmed by a probability plot (not shown).



The mean difference in head shape indices was 0.19, with 95% t-interval $(-1.35, 1.72)$. A t-test of the hypothesis of zero mean difference gave $t = 0.25$ on 24 degrees of freedom, with p-value of 0.803.

Discussion
We conclude that there is little or no evidence against the hypothesis of no difference between the head shape indices of first and second sons. However, the sample size for this study was quite small, being only 25.

Activities 14 and 15 will give you some practice at writing short statistical reports.

Activity 14 *Used windows on reusable envelopes*

This activity is based on the data on used windows on reusable envelopes, which were discussed in Example 11 and analysed in Activity 9. (Source: Sutherland, W. (1990) 'The great pigeonhole in the sky', *New Scientist*, 9 June, vol. 1720, pp. 73–4.) The aims of a slightly extended version of the analysis in Activity 9 were as follows:

- to estimate the mean number of used windows
- to find a suitable model for the distribution of used windows.

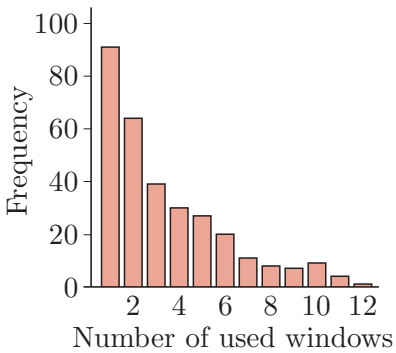


Figure 39 Numbers of used windows

A bar chart of the numbers of used windows is shown in Figure 39. (This is a repeat of Figure 34.) The mean number of used windows for this sample of 311 envelopes was 3.412, with large-sample 95% confidence interval for the mean $(3.122, 3.701)$.

In Activity 9, you may have concluded that the shape of the distribution was suggestive of a geometric distribution, even though the maximum possible number of used windows is 12. The goodness-of-fit of the geometric model was assessed using a chi-squared goodness-of-fit test; after suitable combination of cells with low expected frequencies, a chi-squared test statistic of 13.646 on 9 degrees of freedom was obtained, $p = 0.135$.

Write a short report of this analysis.

Activity 15 *The effect of caffeine on finger-tapping*

This activity is based on the data on the effect of the stimulant caffeine on alertness, as measured by speed of finger-tapping of student subjects, described and analysed in Activities 12, 16 and 18, and Example 15, of Unit 11. (Source: Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd edn, New York, John Wiley and Sons, p. 425.) To remind you, 30 male college students were trained in finger-tapping; they were then randomly divided into three groups of ten, and the students in each group received different doses of caffeine (0 mg, 100 mg and 200 mg). Two hours after treatment, each student was required to do finger-tapping, and the number of taps achieved per minute was recorded. For the purposes of this activity, the aims of the analysis of this dataset were as follows:

- to understand the relationship between caffeine dose and number of taps per minute by using a suitable model
- to predict the number of taps per minute that might be expected if a student had received a dose of 40 mg of caffeine.

A scatterplot of the number of taps per minute against caffeine dose is shown in Figure 40. (This is a repeat of Figure 19 of Unit 11.)

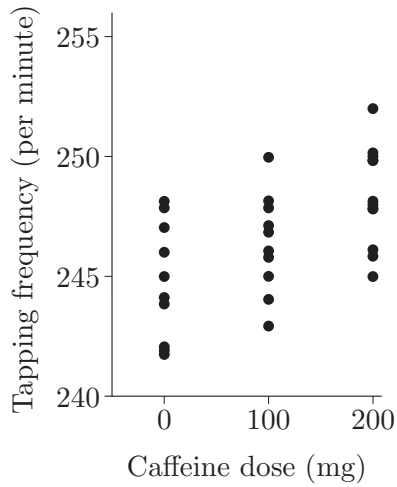


Figure 40 Tapping performance against caffeine dose

A regression line was fitted to the data using least squares; this line has the formula

$$\text{taps} = 244.75 + 0.0175 \times \text{caffeine dose}.$$

Examination of a residual plot and a normal probability plot of the residuals showed no evidence that the simple linear regression model was not suitable for these data. A hypothesis test of whether there is a regression relationship resulted in a p-value of 0.0013.

For a student taking a caffeine dose of 40 mg, the predicted number of taps per minute, using this regression model, is 245.45 with 95% prediction interval $(240.85, 250.05)$.

Write a short report of this analysis.

Exercise on Section 4

Exercise 4 *Fish traps*

In this activity, you are invited to write a report on a further analysis that was carried out of the fish traps data that were considered in Exercises 1 and 2 of Unit 8. (Source: David, F.N. (1971) *A First Course in Statistics*, 2nd edn, London, Griffin.) These data arose from an experiment in which $n = 100$ fish traps were set and the number of fish caught in each trap were counted. The aims of this further analysis were as follows:

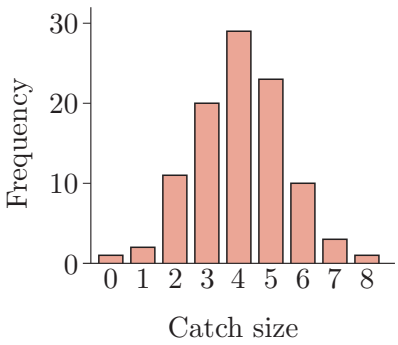


Figure 41 Fish trapped

- to describe the distribution of catches
- to estimate the mean number of fish caught per trap
- to estimate the proportion of traps with no catch.

Out of the 100 fish traps, 72 of the traps contained between 3 and 5 fish, and the maximum number of fish caught in any trap was 8. The distribution of catches is shown in the bar chart in Figure 41.

The mean number of fish per trap was 4.04, with 95% z-interval $(3.76, 4.32)$. One trap failed to catch any fish. The proportion of traps with no catch was thus 0.01. The exact 95% confidence interval for this proportion – obtained using a method not covered in this module – is $(0.00025, 0.054)$.

Write a short report of this analysis.

Summary

This unit is in two parts.

The first part consists of an investigation of the role of transformations of the data in statistics.

In the one-sample case, transformations to normality of continuous variables have been considered and, for use with positive data, the ladder of powers was introduced. In the regression situation, you have seen that sometimes a general regression model can be simplified to a linear regression model by an appropriate transformation of the data: a transformation may be applied to the explanatory variable to linearise the relationship, or a transformation may be applied to the response variable to improve the behaviour of the random terms in the model.

In the second part of this unit, the methods that you have learned so far in the module have been integrated into a statistical modelling process.

Some basic principles for thinking about data and models, even before looking at the data, have been reviewed. Key issues include:

- whether the data are discrete or continuous
- whether the setting in which the data were collected conforms to any of the standard settings
- what is the likely range and shape of the distribution.

These basic principles can help to formulate a starting point for choosing a model, which can be revised in the light of the data. The handling of outliers has been discussed briefly.

You have undertaken extended analysis of a dataset using Minitab, starting from a scientific question, progressing through the various stages of exploratory analysis, model and method choice, model checking, and performing the relevant statistical calculations.

Finally, you have learned how to structure and write a statistical report. A convenient structure includes paragraphs entitled *Summary*, *Introduction*, *Methods*, *Results* and *Discussion*.

Learning outcomes

After you have worked through this unit, you should be able to:

- choose a transformation of a single continuous variable to make its distribution more normal, if necessary
- use the ladder of powers to help choose a transformation in the case of positive random variables
- use a transformation of the explanatory variable to straighten out a non-linear relationship between the variables in a general regression model, so that a linear regression model can be fitted to the transformed data
- use a transformation of the response variable to improve the behaviour of the random terms in the general regression model, so that a linear regression model can be fitted to the transformed data
- fit quadratic and cubic functions of a single explanatory variable using multiple regression
- appreciate that statistical analysis is a process, beginning with a question or problem of interest, ending with a statistical report, and involving data exploration, model choice and model checking, in a cycle that may be repeated several times
- appreciate that the aim of statistical modelling is to draw valid and relevant inferences, not to find a perfect model
- use information about the setting of a problem and the type of data collected to set out an initial modelling framework
- choose appropriate statistical techniques to address a specific problem or question
- identify outliers and explore their influence
- combine the data manipulation, calculation, statistical and graphical facilities of Minitab to undertake a complete statistical analysis
- structure and write a statistical report; such a report comprises a non-technical summary, an introduction, a methods section, a results section, and a discussion.

Solutions to activities

Solution to Activity 1

A continuous uniform distribution is not a good model for these data because the p.d.f. is clearly not constant over whatever suitable range for the distribution of the whole dataset might have been chosen. In addition, the range of a continuous uniform distribution is finite; here, the lower limit, a, of the range could be 0 but what could one justifiably choose for the upper limit, b, the highest interspike time that could possibly occur? Also, the data being clearly right-skew indicates that a normal model is not appropriate either (even if the probability of a negative interspike interval were sufficiently small under a normal model).

Solution to Activity 2

The only available transformation of the four is $y = (2 + x)^2$.

The transformation $y = \sqrt{x}$ is not available because it is not defined for the negative values in the range of x.

The transformation $y = x^4$ is defined over the required range but, like x^2 in Example 4, is neither increasing nor decreasing over the entire range. This is illustrated in Figure 42, where the graph of $y = x^4$ is shown for $-1 < x < 1$: $y = x^4$ is decreasing between -1 and 0, then increasing between 0 and 1. Mathematically, if $h(x) = x^4$, then $h'(x) = 4x^3$, so $h'(x) < 0$ for $-1 < x < 0$ and $h'(x) > 0$ for $0 < x < 1$.

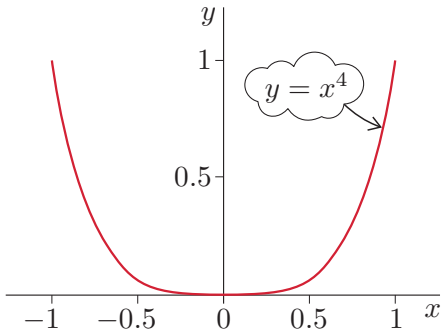


Figure 42 The graph of $y = x^4$ for x between -1 and 1

The transformation $y = (2 + x)^2$ is both defined and increasing over the required range. To see the latter, observe that if $h(x) = (2 + x)^2$, then, using the chain rule (Subsection 3.1 of Unit 7), $h'(x) = 2(2 + x)$. This is a linear function taking the value 2 when $x = -1$ and 6 when $x = 1$, and so is positive for all $-1 < x < 1$. The transformation is shown in Figure 43.

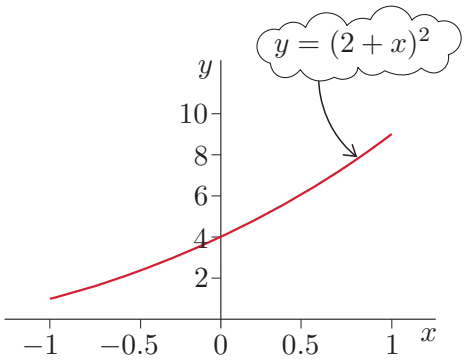


Figure 43 The graph of $y = (2 + x)^2$ for x between -1 and 1

The transformation $y = -\log x$ is not available because log, and hence $-\log$, is not defined for the negative (or zero) values in the range of x.

Solution to Activity 3

(a) The data are, in this case, left-skew. To tackle left-skew, it is necessary to go up the ladder of powers, and it would be worth trying a transformation such as x^2 or x^3. (As it happens, the simulated nature of these data mean that there is a correct answer, that the transformation x^3 would lead to a normally distributed dataset. However, you cannot ascertain this without further exploration that you were not asked to take on.)

(b) The histogram already looks quite symmetric, so no transformation is required. (Because the data are simulated, this is known to be the correct answer, even if you *might* have doubted the normality assumption from the histogram!)

(c) The data are, in this case, right-skew. To tackle right-skew, it is necessary to go down the ladder of powers, and it would be worth trying a transformation such as $z^{\frac{1}{2}}$ or $\log z$. (In fact, in this case, the log transformation is the one that would lead to a normally distributed dataset.)

Solution to Activity 4

(a) Transformations from the ladder of powers with powers less than 1 (including log) would be expected to be the most appropriate when attempting to transform the interspike interval data to normality. This is because the data are right-skew.

(b) The original data are markedly right-skew. Both the log and the square root transformations have reduced the skew by 'pulling in' the values to the right of the mode and 'stretching out' those to the left. This 'stretching out' effect is more marked for the log transformation as can be seen from the fact that low values deviate from the line in the normal probability plot. However, it is rather difficult to decide which of the two transformed datasets is better fitted by the normal distribution. One *might* argue that the log transformation has overcompensated for the right-skew, introducing some points on the

normal probability plot that are out of line with the others but 'in the opposite direction' at either extreme. Arguably, the points on the normal probability plot of the square-root-transformed data better follow a straight line (but a slight bend can still be perceived in the line of points). Has this transformation not quite done enough?

Solution to Activity 5

The first and third regression functions can be linearised by employing the transformations $x' = x^3$ and $x' = \log(x/(1-x))$, respectively. The other two regression functions cannot be linearised in this way because the potential transformations depend on further unknown parameters (λ in the case of the second regression function, both μ and γ in the case of the fourth regression function).

Solution to Activity 6

The prediction of the *square root* of the number of matings of an elephant of age 40 years is

$$\sqrt{y} = -0.812 + 0.0632 \times 40 = 1.716.$$

Therefore the predicted number of matings is

$$y = 1.716^2 = 2.945 \simeq 3.$$

Solution to Activity 7

(a) The regression function in the fitted model is not quite right. This is because the points in the residual plot display some structure. Inspection of the figure reveals a distinctly curved shape.

(b) For the cubic model, the points in the residual plot can be argued to show no clear pattern, so the random terms plausibly have a constant, zero mean and constant variance. Also, the residuals in the normal probability plot lie roughly along a straight line, so the assumption of normality of random terms seems plausible. The model assumptions seem adequate in this case. (As so often, there is a caveat. It could alternatively be argued that there is more spread in the residual values at low fitted values than at high, but there are not really enough data points to be sure.)

(c) In order to make a prediction, it is useful to rewrite the fitted regression model

$$y = 5.65 + 3.58\,x_1 + 0.654\,x_2 - 0.0552\,x_3$$

in terms of the hardwood content x itself, as

$$y = 5.65 + 3.58\,x + 0.654\,x^2 - 0.0552\,x^3.$$

For paper with $x = 10$, the predicted value of tensile strength is

$$\begin{aligned} y &= 5.65 + 3.58 \times 10 + 0.654 \times 100 - 0.0552 \times 1000 \\ &= 51.65 \text{ p.s.i.} \end{aligned}$$

Solution to Activity 8

Five of the distributions in Table 1 are discrete and five are continuous.

The discrete distributions are

Bernoulli, binomial, discrete uniform, geometric, Poisson.

The continuous distributions are

chi-squared, continuous uniform, exponential, normal, t.

Solution to Activity 9

(a) Start with the binomial model. To fit this in with its standard setting, we need to identify a relevant Bernoulli trial. The obvious choice is whether or not a window has been filled. However, as mentioned in Example 11, the probability that a window is filled depends on the position of the window. So, say, if we were to think of each window going down the envelope as a Bernoulli trial, the probability, p, of a 'success' (the window being filled) would change from window to window, and hence not be constant. Thus the standard setting for the binomial model does not seem appropriate. Also, the range of the binomial distribution includes zero, which is not possible in these data. Unless p is such that $P(X = 0)$ is very small under the binomial model, it seems that the binomial distribution is not likely to be a suitable model for these data. (Here, X is a random variable representing the number of used windows.)

The discrete uniform distribution can have the finite range $1, 2, \ldots, 12$. However, it is not clear at all from the setting that each outcome has the same probability; indeed, one would probably expect windows higher up the envelope to have higher probabilities of being filled than windows towards the bottom. However, the jury is out until we see some data!

Our inability to define what would constitute a constant-probability Bernoulli trial in this context rules out the standard setting for the geometric distribution. The range of the geometric distribution starts at 1, which is appropriate to these data, but continues beyond 12 ('to infinity'!), which is not. However, the range constraint might not be a problem if p is such that, under this model, $P(X > 12)$ is sufficiently small.

The standard setting for the Poisson model requires an 'event' that occurs at random, and some fixed interval of time within which such events occur. This doesn't really seem to fit the envelope situation at all. The range of the Poisson distribution includes both 0 and values beyond 12, so it has the difficulties in this respect of the binomial and geometric distributions combined! (However, a way forward if no simpler model proves useful might be to try modelling $X - 1$ using a Poisson distribution.)

From these considerations, there is no compelling reason to opt for any of the standard models, though some – the discrete uniform and geometric – seem more likely to be appropriate than others.

(b) It is clear that envelopes with few used windows are more frequent than envelopes with many used windows. So the uniform distribution is not appropriate.

The shape of the data appears to be consistent in general terms with either a Poisson model or a geometric model. Because its range starts at 1, a geometric model seems to be the more suitable.

(c) The p-value of 0.135, being greater than 0.1, means that there is little or no evidence against the null hypothesis that the geometric distribution provides a good model for these data.

Solution to Activity 10

All three variables are continuous, so a continuous model should be chosen in each case.

(a) The continuous uniform distribution is a *possible* model, but it requires all head sizes to be equally likely over some range of values of head size (and impossible otherwise), which doesn't seem especially plausible for this measurement. Head size is necessarily positive. So the exponential distribution is a possible model for these data. However, the shape of the exponential model does not seem appropriate, with its high probability of head sizes close to zero! Indeed, values of head length plus head breadth are likely to cluster around some typical value some distance from zero, so that a normal model is more likely to be appropriate. Note, however, that the normal model is not ideal since it theoretically allows negative values. If the data on head size are not very normally distributed – for example, they are skew – then a transformation to normality might be considered.

(b) Much the same considerations as in part (a) for head size apply to data on head shape also.

(c) Differences between head shapes can reasonably be expected to take both negative and positive values, perhaps clustered close to zero. A normal model again seems appropriate here, as a first choice; an exponential model is certainly not now appropriate, because of the negative values. A transformation to normality remains a possibility, the transformation being different from that used in part (a) because of the different ranges of the distributions of the data involved.

Solution to Activity 11

(a) The shape of the histogram suggests that the exponential model might, on the face of it, be a reasonable model for these data. (By the way, the normal model seems out of the question, owing to the substantial skewness of the histogram.)

(b) The sample mean and sample standard deviation do not differ greatly. This suggests that an exponential model might be a reasonable choice.

Solution to Activity 12

An estimate of the age of the site based on all eight observations is given by the sample mean, which is approximately 2621.8 years. However, this mean is rather unsatisfactory, as it lies above seven of the eight data points. This is because it is greatly influenced by the value for sample number C-367, which is 3433 years. When this point is omitted, the mean of the remaining seven points is approximately 2505.9 years. Sample number C-367 clearly has a big influence on the mean. You might therefore report the calculations both including and excluding sample C-367, and perhaps suggest further investigation of the outlier. Additionally, you might note that the median of all eight observations is 2530.5 years, which is broadly in line with the sample mean of the data without the outlier. (And the resistant nature of the sample median is reflected in it taking the similar value of 2521 years when the outlier is omitted.)

Solution to Activity 13

Reorganising the material should produce something like the following.

Introduction
Compare means of a continuous variable in two groups given samples of sizes 24 and 32.

Methods
Check normality in each group using probability plots.
Check assumption of equality of variances.
Calculate 95% two-sample t-interval.
Perform two-sample t-test.

Results
Normal model reasonable.
95% t-interval was $(-3.92, 17.63)$.
Two-sample t-test gave $p = 0.16$.

Discussion
Little or no evidence that the means are different in the two groups.

Solution to Activity 14

Here is a possible report.

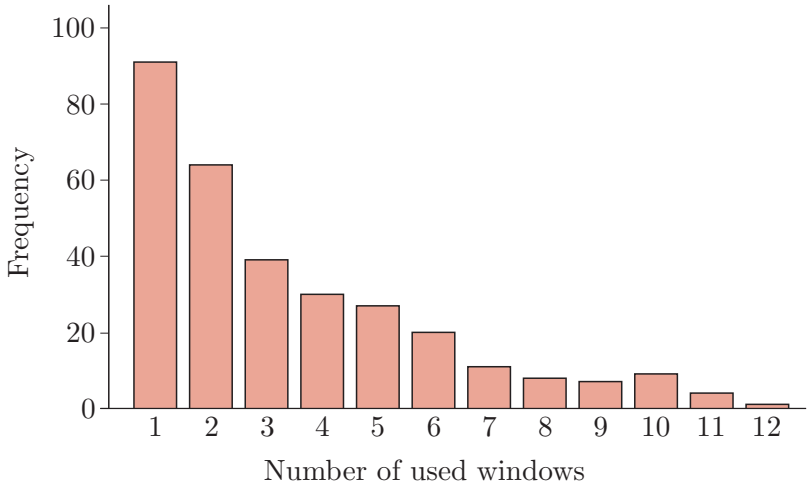
Summary
The main aim of this analysis is to develop a model for the distribution of the number of windows used on a sample of reusable envelopes. Data on a sample of 311 envelopes were obtained from a published source. A geometric model was found to provide a good fit to the data.

Introduction
The numbers of used windows on a sample of 311 reusable envelopes used for internal circulation of documents within an office were counted. The aim of the analysis is to describe the distribution of the number of used windows, estimate the mean number of used windows, and obtain a model for the distribution of used windows. The data for this analysis were

obtained from Sutherland, W. (1990) 'The great pigeonhole in the sky', *New Scientist*, 9 June, vol. 1720, pp. 73–4.

Methods
A 95% z-interval for the mean number of used windows was calculated. The fit of the geometric model was assessed using the chi-squared goodness-of-fit test. Most calculations were performed using Minitab Version 17.

Results
The numbers of used windows on the 311 envelopes were distributed as shown in the bar chart below. The mean number of used windows was approximately 3.41, with approximate 95% confidence interval $(3.12, 3.70)$.



A geometric model with $p = 1/3.412 \simeq 0.29$ provided a good fit to the data. The chi-squared test statistic was 13.646 on 9 degrees of freedom, $p = 0.135$.

Discussion
The average number of used windows on the envelopes is about 3.41. A geometric distribution provided a very good fit to the data. However, an approximate aspect of the model is that it does not allow for the fact that the number of windows on each envelope is restricted to 12.

Solution to Activity 15

Here is a possible report.

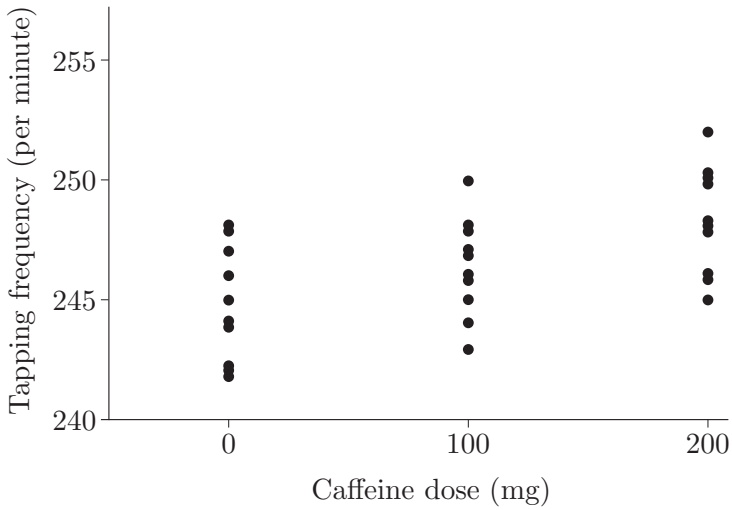
Summary
The main aim of this analysis is to develop a model for the effect of the stimulant caffeine on the alertness of subjects, as measured by their rate of finger-tapping. Data from an experiment on a sample of male subjects trained in finger-tapping were obtained from a published source. A linear regression model was found to provide a good fit to the data, allowing for interpretation and prediction of the effect of caffeine on finger-tapping.

Introduction
The tapping rates per minute of a trained sample of 30 male subjects,

each given one of three doses of caffeine (0 mg, 100 mg and 200 mg) were recorded. The aim of the analysis is to develop a model for the effect of caffeine on finger-tapping, and to use the model to interpret and predict this effect. The data for this analysis were obtained from Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd edn, New York, John Wiley and Sons, p. 425.

Methods
A linear regression model was fitted to the data using least squares. The fit of the regression model was assessed using a residual plot and a normal probability plot of the residuals. A hypothesis test of whether there is any regression relationship was performed. Point and interval predictions of finger-tapping rates for someone taking 40 mg of caffeine were obtained. Most calculations were performed using Minitab Version 17.

Results
A scatterplot of the number of taps per minute against caffeine dose is shown below.



The regression line fitted to these data is

$$\text{taps} = 244.75 + 0.0175 \times \text{caffeine dose}.$$

Examination of a residual plot and a normal probability plot of the residuals showed no evidence that the simple linear regression model was not suitable for these data. A hypothesis test of whether there is a regression relationship resulted in a p-value of 0.0013; there is strong evidence of the existence of a relationship.

The model suggests that, on average, the number of taps per minute increases by 0.0175 for each extra mg of caffeine taken.

For a student taking a caffeine dose of 40 mg, the predicted number of taps per minute, using this model, is about 245.45 with 95% prediction interval $(240.85, 250.05)$.

Discussion
There is strong evidence of a linear regression relationship between

caffeine dose and number of taps per minute, a small increasing effect being observed. This relationship has been used for prediction. It is unclear how far a linear relationship would continue to hold for caffeine doses greater than those (up to 200 mg) used in the experiment.

Solutions to exercises

Solution to Exercise 1

Any member of the ladder of powers other than log can be written as $h(x) = x^p$ where p takes values $\ldots, -2, -1, -1/2, 1/2, 1, 2, 3, 4, \ldots$. The derivative of $h(x)$ is

$$h'(x) = px^{p-1}.$$

For $x > 0$, $x^{p-1} > 0$ for all values of p on the ladder. It follows that $px^{p-1} > 0$ for all ladder values of $p > 0$ and $px^{p-1} < 0$ for all ladder values of $p < 0$. It then follows that $h(x) = x^p$ is an increasing function of positive x for all ladder values of $p > 0$ (including $p = 1/2, 1, 2, 3, 4$) and that $h(x) = x^p$ is a decreasing function of positive x for all ladder values of $p < 0$ (including $p = -2, -1, -1/2$).

Solution to Exercise 2

(a) While a normal model is not an indefensible one for these data, it does seem from the histogram that there is a certain amount of left-skew.

(b) Transformations from the ladder of powers with powers greater than 1 would be expected to be the most appropriate when attempting to transform the glass fibre strength data to normality. This is because the data are left-skew.

(c) The left-skew in the data is visible in Figure 16(a) as a bend in the line of points which 'faces the other way' in comparison with the bend due to right-skew in the normal probability plot of Figure 13, for example. All three transformations have reduced the left-skew by 'pulling in' the values to the left of the mode and 'stretching out' those to the right. It appears that the square transformation (Figure 16(b)) may not have done enough to remove the bend in Figure 16(a), while the fourth power transformation (Figure 16(d)) seems to have transformed the data 'too far': while the central dots in the latter plot follow a reasonable straight line, there are now outlying points at both extremes. Of the candidates offered, the cube transformation (Figure 16(c)) has, arguably, set the best balance in achieving something approaching normality. That said, one might still reject normality as a suitable model even for the cubes of the data.

Solution to Exercise 3

Figure 32(a) is a residual plot which is of the type that you might expect to obtain when the assumptions are justified. There is therefore no need to employ any transformations.

Figure 32(b) displays a strong pattern indicating a systematic discrepancy from the assumed mean of the model. The relationship between the response and explanatory variables appears to be non-linear. Transforming the explanatory variable might therefore provide a remedy. (Indeed, the quadratic/cubic nature of the main trend in the residual plot might indicate the need for a treatment like that of the kraft paper in Subsection 2.3 using multiple regression, involving certain transformations of the explanatory variable.)

In Figure 32(c), the pattern is indicative of a variance that is not constant, but increasing as the fitted value increases. Transformation of the response variable is an approach open to you to try to accommodate such behaviour.

Finally, the residual plot in Figure 32(d) is well-behaved except for a single outlier. Transformations are not, typically, the answer here (but see Subsection 3.3 for a little more on dealing with outliers).

Solution to Exercise 4

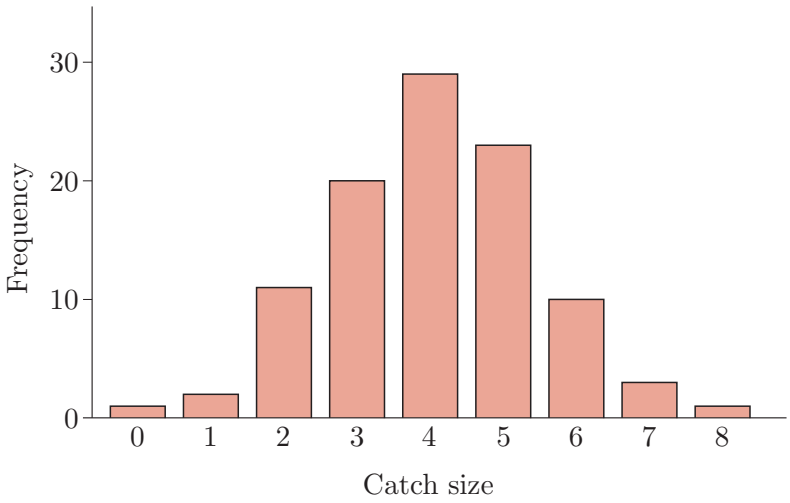
Here is a possible report.

Summary
The observed distribution of catch sizes from 100 fish traps is described and some aspects of the population distribution of catch sizes are estimated using data obtained from a published source.

Introduction
The numbers of fish caught in 100 traps were counted. The aims of this analysis are to describe the observed distribution of fish catches, and to estimate the population mean catch per trap and the population proportion of traps with zero catch. The data for this analysis were obtained from David, F.N. (1971) *A First Course in Statistics*, 2nd edn, London, Griffin.

Methods
The observed data were graphed. An approximate 95% confidence interval for the mean catch per trap was calculated using large-sample methods. A 95% confidence interval for the proportion of traps with zero catch was calculated using exact methods. All calculations were performed using Minitab Version 17.

Results
A bar chart of the numbers of catches is shown below.



Of the 100 traps, 72 contained between 3 and 5 fish. The minimum catch was 0, the maximum was 8. The average catch per trap was 4.04 fish, with approximate 95% confidence interval for the population mean $(3.76, 4.32)$. Out of the 100 traps, only one produced a zero catch. The proportion of traps with no catch is thus 0.01, with exact 95% confidence interval for the population proportion $(0.00025, 0.054)$.

Discussion
The distribution of fish catches was approximately symmetric with mean about 4. Only one out of one hundred traps had zero catch.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 81: © Katarzyna Bialasiewicz / Dreamstime.com

Page 82: Courtesy of Richard Mooney, Duke University

Page 85: © 2017 MSP Communications, Inc

Page 96: © LUZI et al. (2004) www.sswm.info This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 97: © Fernando Gregory Milan / www.123rf.com

Page 99: © NSIL/Dick Roberts, Visuals Unlimited / Science Photo Library

Page 101: © Sicnag This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 103: © Duncan Noakes / www.123rf.com

Page 106: © Mikhail Olykaynen / www.123rf.com

Page 111 top: © hxdbzxy / www.123rf.com

Page 111 bottom: © John Sommer / iStock / Getty

Page 112: © The Keasbury-Gordon Photograph Archive / Alamy Stock Photo

Page 115: Taken From: https://www.amazon.co.uk/New-Guardian-Intertac-Envelopes-Resealable/dp/B000I2BZI6

Page 118: © Guy Corbishley / Alamy Stock Photo

Page 119: © Roger This file is licensed under the Creative Commons Attribution-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 122: © Courtesy Lamoka Waneta Lakes Association

Page 124: © alkanc / www.123rf.com

Page 127: © Andrei Zaripov / www.123rf.com

Page 129: © Andrei Zaripov / www.123rf.com

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.

Unit 13

Applications

Introduction

This short unit considers a number of datasets, one per section (except the last), to illustrate the application of some of the statistical methods developed in the module. Each application is explored through a set of activities designed to give you practice in answering questions of the sort you are likely to meet in the examination. As such, this module also serves as useful revision, although obviously there is not enough space in a single unit to illustrate everything that has been covered in the module.

All the calculations, and even graphical representations, that you are asked to make in this unit can, and should, be done 'by hand', perhaps using your calculator and/or statistical tables; you do not need to use your computer at all.

1 Chocolates

Cadbury Heroes are miniature chocolate bars. At the time of writing, a 712 g tub of Cadbury Heroes contains seven different types of chocolate bar: Caramel, Creme Egg Twisted, Dairy Milk, Eclair, Fudge, Twirl and Wispa. In this section, we will investigate the distribution of the different types of chocolate bar found in 712 g tubs of Cadbury Heroes. Because it is convenient for random variables to take numerical values, we will code the different types of chocolate bars so that each has a numerical value. So, let Caramel be coded as 1, Creme Egg Twisted as 2, Dairy Milk as 3, Eclair as 4, Fudge as 5, Twirl as 6, and Wispa as 7.

Activity 1 *Which distribution?*

For each chocolate bar in a tub of Cadbury Heroes, let X be its type.

(a) What is the range of X?

(b) It seems reasonable to expect there to be equal numbers of each type of chocolate bar within a 712 g tub of Cadbury Heroes. In this case, suggest a suitable distribution for X and write down its probability mass function and cumulative distribution function.

The solution to Activity 1 argued that if a tub of Heroes has equal numbers of each type of chocolate bar, then a suitable distribution for X is the discrete uniform distribution with parameters $m = 1$ and $n = 7$. The numbers of the different types of chocolate bar in one particular 712 g tub of Cadbury Heroes (purchased in December 2016) are given in Table 1 (overleaf).

Table 1 Number of chocolates of different types

Type of chocolate	Frequency
1 (Caramel)	17
2 (Creme Egg Twisted)	10
3 (Dairy Milk)	12
4 (Eclair)	15
5 (Fudge)	5
6 (Twirl)	4
7 (Wispa)	8

(Source: C.M. Queen, The Open University)

This tub certainly didn't have equal numbers of each type of chocolate bar. Does that mean that the discrete uniform distribution with parameters $m = 1$ and $n = 7$ is not a suitable model for X? Or could the frequencies observed in Table 1 have arisen by chance from this discrete uniform distribution? You will use a chi-squared goodness-of-fit test to answer this question in the next activity.

The chi-squared goodness-of-fit test was covered in Section 2 of Unit 10.

Activity 2 *Does the discrete uniform distribution fit the data?*

(a) There were a total of 71 chocolate bars in the tub of Cadbury Heroes which provided the data in Table 1. Calculate the expected frequencies for X taking values $1, 2, \ldots, 7$, for a tub of 71 Cadbury Heroes when X is modelled by the discrete uniform distribution with parameters $m = 1$ and $n = 7$.

(b) Carry out a chi-squared goodness-of-fit test to test whether the discrete uniform distribution with parameters $m = 1$ and $n = 7$ is a suitable model for the data in Table 1.

The solution to Activity 2 concluded that there is moderate evidence to suggest that the discrete uniform distribution with parameters $m = 1$ and $n = 7$ is not a suitable model for X. This means that there is evidence to suggest that the different types of chocolate bars in a 712 g tub of Cadbury Heroes do not occur in equal numbers. Of course, these data were only for one single tub of chocolates, and the result from this test doesn't necessarily mean that *all* such tubs do not have equal numbers of each type of chocolate bar. To investigate this further, we would need more data and would need to see the contents of more 712 g tubs of the chocolates. This might be a simpler and more attractive exercise in gathering further data than statisticians are usually faced with!

2 Paralympic Games 2016

Since 2004, five-a-side football has been an event in the Paralympic Games played by athletes with visual impairment, including blindness. A special football is used which makes a noise when it moves so that players can locate it through sound. In order that no player has an unfair advantage, all players apart from the goalkeepers wear eye shades. Each team has 5 players, including the goalkeeper, who may be sighted and is also allowed to act as a guide during the game.

The Rio 2016 Paralympic Games saw eight five-a-side football teams compete. The teams were divided into two groups, A and B, of four teams each. All teams within a group played each other; the two best teams in each group went on to play in the semi-finals, and the winners of the two semi-finals played in the final. There were also matches to decide who came in 3rd, 4th, ..., 8th place. In all, 18 matches were played.

For each match, the number of goals for each team was recorded: these ranged from 0 goals to 3 goals. Table 2 gives the number of times that a team scored 0, 1, 2 and 3 goals for the 18 matches (including the scores of both teams for each match). Games in the knockout stage of the competition which ended in a draw were decided by having a 'penalty shoot-out', in which each team takes the same number of penalties and the team which scores the most penalties is declared the winner. Only the original scores for the games before any penalty shoot-outs are given in Table 2.

Table 2 Number of goals scored

Number of goals	Frequency
0	21
1	8
2	6
3	1

(Source: Wikipedia, http://en.wikipedia.org/wiki/Football_5-a-side_at_the_2016_Summer_Paralympics, 5 October 2016)

We are interested in finding a suitable probability model for the random variable X, the number of goals scored in individual matches by individual teams in five-a-side football matches at the 2016 Paralympics. We immediately make one strong – and clearly wrong – assumption: that the number of goals scored by any team in any match is independent of the numbers of goals scored by any team (including the same one) in any other match (or indeed of the number scored by the opposing team in the same match). Accommodating dependence between the numbers of goals scored by different teams (some are stronger, some are weaker), accounting for different opposing teams (again, some are stronger, some are weaker) and other factors, makes the modelling exercise much more difficult and beyond the scope of this module. However, models being models, a model

developed under an assumption of independence can still be useful for some purposes, provided that we remember the strong assumptions made.

Activity 3 *A model for goals scored*

(a) Present the data in Table 2 by means of a suitable simple graphical display.

(b) Suggest a modelling distribution for random variable X. Give two reasons to support your choice of distribution.

The solution to Activity 3 suggested the Poisson distribution as a plausible model for X, the number of goals scored in individual matches by individual teams in five-a-side football matches at the 2016 Paralympics. Let $\theta > 0$ denote the parameter of this distribution. In the next activity, you will obtain the likelihood of θ for these data assuming a Poisson model, and use this to find the maximum likelihood estimate $\widehat{\theta}$ of θ.

The Poisson distribution was covered in Section 4 of Unit 3, and the likelihood and maximum likelihood estimation were covered in Unit 7.

Activity 4 *Maximum likelihood estimate*

(a) Show that the likelihood of θ for these data is

$$L(\theta) = \frac{e^{-36\theta}\theta^{23}}{384}.$$

(b) By differentiating the likelihood, find the maximum likelihood estimate $\widehat{\theta}$ of θ, giving its value correct to three decimal places.

You showed in Activity 4 that the MLE of θ for these data is $\widehat{\theta} \simeq 0.639$. So if a Poisson model is a sensible model for the data, then the Poisson model which fits the data best is Poisson(0.639). In the next activity you will test whether in fact the Poisson(0.639) distribution is a good fit to the data in Table 2.

Activity 5 *Is the Poisson model a good fit?*

(a) Calculate the expected frequencies for scoring 0, 1 and ≥ 2 goals in 36 scores when the number of goals scored is modelled by the Poisson(0.639) distribution.

(b) A chi-squared goodness-of-fit test is to be carried out using the categories: '0 goals', '1 goal' and '≥ 2 goals'. Explain why the category '≥ 2 goals' is to be used rather than separate categories for '2 goals', '3 goals', and so on.

The chi-squared goodness-of-fit test was covered in Section 2 of Unit 10.

(c) Carry out a chi-squared goodness-of-fit test at the 5% significance level, using the data categorised as in part (b), to test whether the Poisson(0.639) model is a suitable model for the data in Table 2.

The solution to Activity 5 concluded that the Poisson(0.639) distribution seems to be a suitable model for the number of goals scored in individual matches by individual teams in five-a-side football matches at the Paralympics, using the 2016 Rio event for data. It would be interesting to investigate how widely such a Poisson model remains applicable, and how much there is to be gained by more sophisticated modelling taking into account strengths of teams, who's playing whom, etc.

3 Times between major tsunamis

On its website, *National Geographic* magazine describes a tsunami as: 'a series of ocean waves that sends surges of water, sometimes reaching heights of over 100 feet (30.5 meters), onto land. These walls of water can cause widespread destruction when they crash ashore' (http://www.nationalgeographic.com/environment/natural-disasters/tsunamis). Tsunamis are typically caused by underwater earthquakes, landslides or volcanic eruptions.

Table 3 contains the 29 'waiting times', in months, between 30 major tsunamis occurring worldwide between January 1950 and December 2015. The numbers should be read across the rows.

Table 3 Time interval between major tsunamis (months)

44	24	22	41	5	3	8	48	90	40	5	36	112	10	11
49	13	64	19	4	5	8	21	5	8	4	1	23	31	

(Source: Wikipedia, http://en.wikipedia.org/wiki/List_of_historical_tsunamis, 5 October 2016)

Because these data are waiting times, the two most likely distributions to model the data are the exponential distribution and the geometric distribution.

The exponential and geometric distributions for modelling waiting times were covered in Section 2 of Unit 5.

Activity 6 *A model for the waiting times between major tsunamis*

Explain why an exponential distribution would be more appropriate than a geometric distribution for modelling the variation in the waiting times between major tsunamis.

In the solution to Activity 6, the case was made that the exponential distribution is more appropriate than the geometric for modelling the variation in waiting times between major tsunamis. In the following activity, you will explore whether the exponential distribution seems to be a reasonable model for these data.

Activity 7 *Is an exponential model reasonable?*

Figure 1 shows a frequency histogram of the 29 waiting times between major tsunamis given in Table 3. For these data, the sample mean is (exactly) 26 months, and the sample standard deviation is (approximately) 27.01 months.

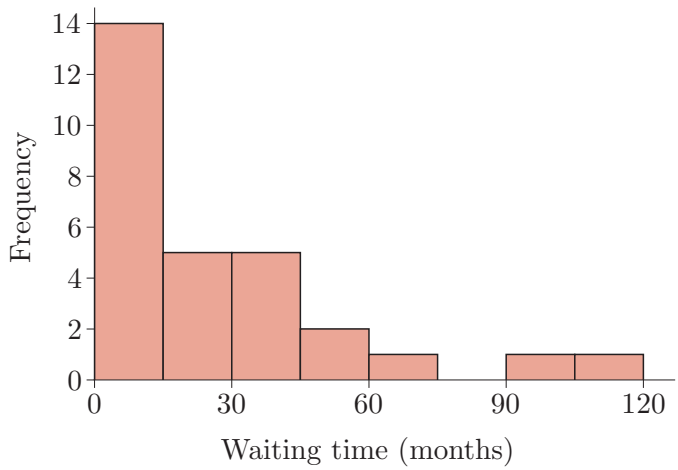


Figure 1 A histogram of waiting times between major tsunamis

Explain briefly whether or not an exponential distribution seems to be a reasonable model for the waiting times between major tsunamis.

The solution to Activity 7 suggested that an exponential distribution would appear to be a reasonable model for the variation in waiting times between major tsunamis. But which exponential distribution do the data suggest?

Activity 8 *Which exponential distribution?*

Use the information given in Activity 7 to find an estimate for the parameter λ of an exponential distribution to model the variation in waiting times between major tsunamis.

The solution to Activity 8 suggested an exponential distribution with parameter $\lambda = 1/26 \simeq 0.038$ for modelling the waiting times between major tsunamis. You will use this model to calculate some probabilities in the next activity.

Activity 9 *Calculating probabilities*

In this activity, use an exponential distribution with parameter $\lambda = 1/26$ to model the waiting times between major tsunamis.

(a) According to the model, what is the probability that the waiting time between two successive major tsunamis is at least one year?

(b) According to the model, what is the probability that the waiting time between two successive major tsunamis is less than 6 months?

(c) For the assumed exponential model, out of 29 waiting times, how many waiting times would be expected to be at least one year? How many would be expected to be less than 6 months? How do these expected values compare with the observed numbers of waiting times that were at least one year and less than 6 months, respectively?

In the next activity, we will make the assumption that the occurrences of major tsunamis may be modelled by a Poisson process, and in the final activity of this section you will investigate whether the data in Table 3 justify this assumption.

The Poisson process was covered in Section 3 of Unit 5.

Activity 10 *Assuming a Poisson process*

Assume that the occurrences of major tsunamis may be modelled by a Poisson process.

(a) Given that the mean time between major tsunamis for the data in Table 3 was 26 months, what is an estimate of the rate λ of the process per month?

(b) Let X be the number of major tsunamis that occur in one year. What is the distribution of X?

(c) Hence calculate the probability that:

(i) exactly two major tsunamis will occur in one year

(ii) at least one major tsunami will occur in one year.

Activity 11 *Is the assumption of a Poisson process reasonable?*

(a) If the occurrences of major tsunamis are modelled by a Poisson process, what assumptions about the occurrences of major tsunamis are being made?

(b) In Figure 2, the number of major tsunamis that had occurred since the start of the period of observation is plotted against the times at which the major tsunamis occurred. By considering all of the Poisson process model assumptions that you identified in part (a), is it reasonable to assume that the occurrence of major tsunamis may be modelled by a Poisson process?

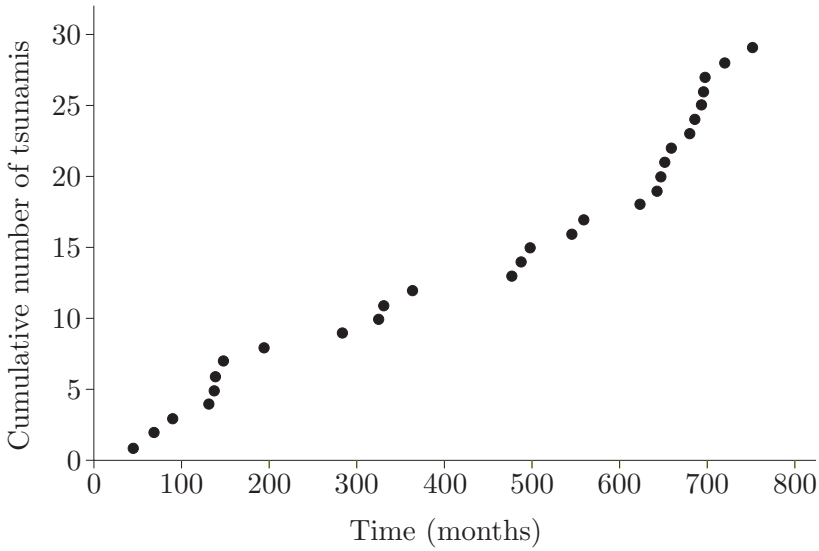


Figure 2 Scatterplot of cumulative number of major tsunamis against their times of occurrence

Activity 11 questioned whether a Poisson process is a reasonable model for the occurrences of major tsunamis. On the one hand, the data suggest that it may not be a reasonable model because Figure 2 suggests that the occurrences are becoming more frequent over time; if so, it is to the earth sciences that we must turn to find an explanation if such an increase in rate is real. But on the other hand, it is possible that the Poisson process *is* a reasonable model for the occurrences of major tsunamis more generally, but other factors concerning these particular data (such as having data for waiting times only in months rather than days, or the possibility that recording of tsunamis is improving over time which makes them appear to occur more frequently) are casting doubt on the model. This illustrates the need for a statistician to keep an open mind about the data available when carrying out a statistical analysis.

4 Pneumonia risk for smokers with chickenpox

Pneumonia can be a serious complication of chickenpox in adults. A study was conducted to determine whether smoking is a risk factor for

pneumonia for patients with chickenpox. The data in Table 4 are measurements of the carbon monoxide (CO) transfer factor levels in seven smokers with chickenpox who were admitted to a hospital. CO transfer factor is a measure of lung function; high values are good. CO transfer factor levels were recorded with a view to determining each patient's risk of contracting pneumonia. The measurements were taken when the patients entered the hospital and were repeated one week later. Here we are interested in investigating whether the data suggest that there is a difference in the CO transfer factor levels on entry and one week later. (On admission, patients were treated with intravenous acyclovir at 10 mg/kg, eight-hourly for five days. It is not recorded whether they were required to abstain from smoking.)

CO transfer being measured using modern equipment

Table 4 CO transfer factor levels in smokers with chickenpox

Patient	On entry	One week later
1	40	73
2	50	52
3	56	80
4	58	85
5	60	64
6	62	63
7	66	60

(Source: Ellis, M.E., Neal, K.R. and Webb, A.K. (1987) 'Is smoking a risk factor for pneumonia in patients with chickenpox?', *British Medical Journal*, vol. 294, no. 6578, p. 1002)

With so few data points, a histogram cannot usefully be drawn, but boxplots may be informative; you will investigate these in the next activity.

Activity 12 *A comparative boxplot*

(a) Calculate the median and the interquartile range for each set of CO transfer factor levels in Table 4.

(b) A comparative boxplot for the data is shown in Figure 3.

The median and interquartile range were covered in Section 4 of Unit 1, and comparative boxplots were covered in Subsection 5.3 of that unit.

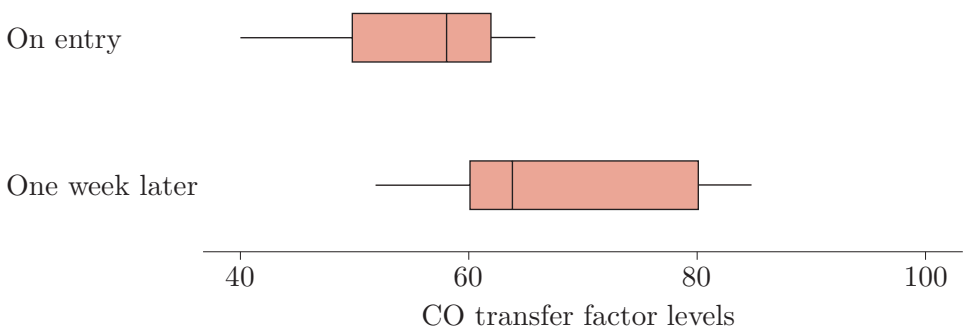


Figure 3 A comparative boxplot for CO transfer factors

Describe two main features of these data as revealed by the comparative boxplot in Figure 3.

As concluded in the solution to Activity 12, the comparative boxplot in Figure 3 suggests that the CO transfer levels are higher one week later than on entry. We can test more formally whether the levels are in fact different. Measurements have been repeated on the same patients, so the data are paired. Therefore, in order to test whether the CO transfer factor levels on entry are different to the CO transfer factor levels one week later, tests involving the differences between the measurements are required.

Activity 13 *Which tests?*

(a) Obtain the differences between measurements in Table 4, calculated as CO transfer factor level 'one week later' minus CO transfer factor level 'on entry'.

(b) Explain how you would investigate whether a normal model for the differences between the CO transfer levels on entry and one week later is plausible.

(c) If the assumption of a normal model is plausible, what test would you use to test whether the CO transfer factor levels on entry and one week later are different? For this test, what are the null and alternative hypotheses?

(d) If the assumption of a normal model is untenable, what alternative test would you use to test whether the CO transfer levels on entry and one week later are different? What are the null and alternative hypotheses in this case?

The solution to Activity 13 suggested two tests for testing whether the CO transfer factor levels on entry and one week later are different: the (one-sample) t-test when the assumption of normality of the differences is plausible, and the Wilcoxon signed rank test when the normality assumption is untenable. In Activities 14 and 15, you will carry out the t-test and Wilcoxon signed rank test, respectively. You will investigate whether the normality assumption is indeed plausible or not in Activity 16.

The t-test was covered in Subsection 3.1 of Unit 9, and the Wilcoxon signed rank test in Subsection 1.1 of Unit 10. Investigating normality was covered in Section 5 of Unit 6.

Activity 14 *Testing when the assumption of normality is justified*

In this activity, we assume that the assumption of normality of differences is plausible.

(a) The mean of the differences for the data in Table 4 (taking the values 'one week later' minus 'on entry') is 12.14, and the standard deviation of the differences is 15.38. Obtain the value of the test statistic for the

The differences themselves are in Table 8 in the solution to Activity 13.

t-test for testing the hypotheses

$$H_0 : \mu_D = 0, \quad H_1 : \mu_D \neq 0,$$

where μ_D is the (population) mean of the differences.

(b) What is the null distribution of the test statistic?

(c) The p-value for the test as calculated by Minitab is 0.082. What do you conclude?

Activity 15 *Testing when the assumption of normality is untenable*

In this activity, we assume that the assumption of normality of differences is untenable.

The differences between measurements in Table 4 (calculated as 'one week later' minus 'on entry') were given in Table 8 in the solution to Activity 13, and are repeated in Table 5 for convenience.

Table 5 Differences 'one week later' minus 'on entry'

Patient	1	2	3	4	5	6	7
Difference	33	2	24	27	4	1	−6

(a) Calculate the Wilcoxon signed rank test statistic for testing the hypotheses

$$H_0 : m_D = 0, \quad H_1 : m_D \neq 0,$$

where m_D is the (population) median of the differences.

(b) The p-value for the test as calculated by Minitab is 0.108. What do you conclude?

The t-test in Activity 14 and the Wilcoxon signed rank test in Activity 15 concluded, respectively, that there was only weak, and little or no, evidence that the CO transfer factor levels were different on entry and one week later. The t-test, however, assumed that the differences were normally distributed, and if that assumption is untenable, then the test is not valid. So is the assumption of normality plausible or not?

Activity 16 *Is the assumption of normality plausible?*

The normal probability plot for the seven differences is given in Figure 4 (overleaf).

Do you think that the assumption that the differences in CO transfer factor levels are normally distributed is plausible or not?

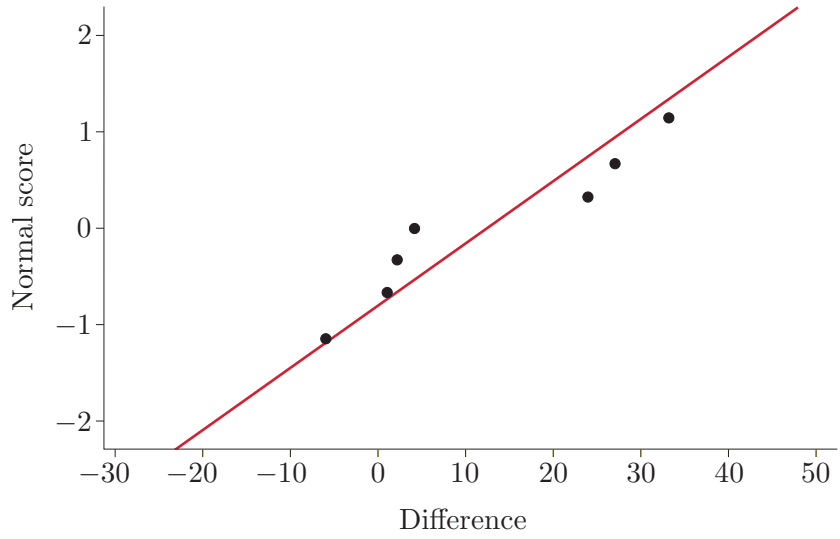


Figure 4 Normal probability plot for CO transfer factor differences

The solution to Activity 16 failed to come to a satisfactory conclusion as to whether the assumption of normality is plausible or not, so we are unsure of the validity of the t-test. In such cases, it is sensible to consider both tests: if both tests give the same conclusion, then you can be fairly confident that you have come to the right conclusion. In the case being considered here, neither test provided convincing evidence to suggest that there is any difference in CO transfer factor levels on entry and one week later.

5 The teleportation parameter

Computer scientists are interested in how many web links are followed by individuals surfing the internet. One parameter of interest is the so-called 'teleportation parameter', α, which can be defined as the probability that someone follows up on the information given on a web page by clicking on one of the links on that page.

An article considered, amongst other things, information collected automatically on web-user behaviour which gave a dataset of proportions of websites with links followed by each user. (Source: Gleich, D.F. et al. (2010) 'Tracking the random surfer: empirically measured teleportation parameters in PageRank', *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pp. 381–90.) As the data are collected automatically, the sample size is very large: $n = 257\,664$. Using these data, the authors of the cited work modelled the distribution of the proportion of web page visits from which a link was followed, X say, by the distribution with probability density function (p.d.f.) of the form

$$f(x) = 12x^2(1 - x), \quad 0 < x < 1. \tag{1}$$

Here, the range of X is determined by the fact that it is a proportion. This p.d.f. is shown in Figure 5.

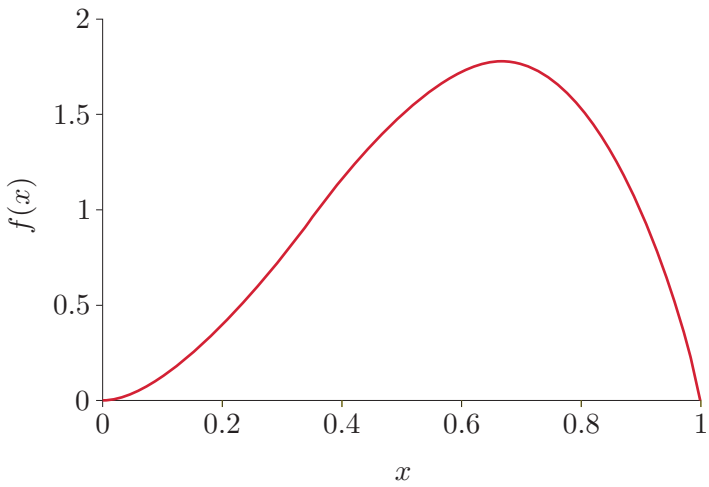


Figure 5 The p.d.f. $f(x) = 12x^2(1-x)$ on $0 < x < 1$

The teleportation parameter, α, can be defined as the mean of this distribution. In this section, you will use this model to calculate several properties of the distribution of proportions of websites with links followed, including the value of α associated with this model. The integration techniques that you will need in this section were reviewed in Subsection 3.1 of Unit 2.

The first task is to check that the claimed p.d.f., $f(x)$, really is a valid p.d.f. This you will do in the next activity.

Teleportation is really the theoretical notion of transferring matter from one place to another without going through the space between them; this is the fictional *Star Trek* transporter/teleporter!

Activity 17 *Checking that f is a p.d.f.*

Show that $f(x)$ given by Equation (1) really is a p.d.f.

The requirements for a function to be a p.d.f. were given in Subsection 3.3 of Unit 2.

Now that we are sure that $f(x)$ given by Equation (1) really is a p.d.f., we can find its corresponding cumulative distribution function (c.d.f.). This you will do in the next activity.

Activity 18 *Finding the c.d.f.*

Find the formula for the c.d.f. $F(x)$ when the p.d.f. $f(x)$ is given by Equation (1).

The c.d.f. of a continuous distribution was discussed in Subsection 4.3 of Unit 2.

The c.d.f. of a distribution can be used to find probabilities connected with that model. Recall that X is the proportion of web page visits from which a link is followed. Find the probabilities required in the following activity

by using the c.d.f. associated with the p.d.f. in Equation (1), which was found in the solution to Activity 18 to be

$$F(x) = x^3(4 - 3x), \quad 0 < x < 1. \tag{2}$$

Activity 19 *Finding probabilities*

Use the c.d.f. $F(x)$ given by Equation (2) to calculate the following probabilities.

(a) $P(X < 0.5)$

(b) $P(0.3 \leq X \leq 0.7)$

(c) The probability that the proportion of web page visits from which a link was followed is at least 0.6.

Now, the teleportation parameter itself, α, is defined as the mean of the distribution of proportions of website visits from which a link is followed, that is, $\alpha = E(X)$. In the next activity, you will find the value of α for the model given by Equation (1).

Activity 20 *The value of the teleportation parameter*

The mean of a continuous distribution was covered in Section 2 of Unit 4.

Calculate the value of the teleportation parameter $\alpha = E(X)$ for the model given by Equation (1).

In addition to the mean of the distribution whose p.d.f. is given by Equation (1), you can also evaluate some measures of the spread of the distribution; this is the topic of Activity 21.

Activity 21 *Measures of spread*

The variance and standard deviation of a continuous distribution were covered in Section 4 and especially Subsection 5.2 of Unit 4.

For the model given by Equation (1), calculate the value of:

(a) the variance, $V(X)$

(b) the standard deviation, $S(X)$.

You should make use of the value $E(X) = 0.6$ that you found for this model in Activity 20.

Remember that, in this section, we have been using a model obtained, from data, by others in order to make some statements concerning the distribution of the proportions of website visits for which a link on that web page is followed. We did this assuming that the model those researchers obtained is a good representation of the practical situation. In addition to using the models that they or others come up with,

statisticians spend a lot of their time producing, fitting, checking and refining the details of appropriate models for the data in the first place.

6 Daily steps

Along with the recent development of smartphones and smartwatches – but not, at the time of writing, teleportation devices! – there has been a surge in electronic devices and software for tracking personal activity. These range from specialist fitness tracking devices, many of which look rather like watches, to software which can be installed on smartwatches or smartphones.

One of the fitness trackers on the market in 2016

The dataset used in this section comprises some data collected by the author of this section about herself! As such, it has been written in the first person.

I don't have a special fitness tracking device, but I do have software on my phone which counts my steps. Obviously, steps are counted only when I carry my phone. I have extracted the data from my phone of the daily steps for 51 days between 1 October and 30 November 2015. This was a period when I regularly carried my phone for most of the day, although there were some days when I didn't carry my phone, resulting in very few or no steps being recorded for those days. I have excluded the data for such days from the analysis, since these very low counts would be unrepresentative of my usual number of daily steps.

Activity 22 *Histogram of the data*

Figure 6 shows a frequency histogram of daily steps on 51 days.

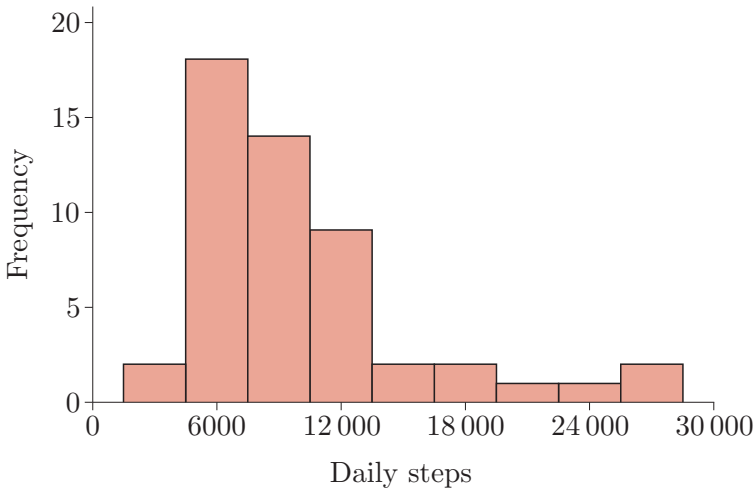


Figure 6 A histogram of daily steps

(Source: data provided by C.M. Queen, The Open University)

(a) Briefly describe the shape of the histogram.

(b) I would like to calculate a confidence interval to give a plausible range of values for my average number of daily steps. Which confidence interval would be more appropriate: a z-interval or a t-interval? Explain your answer.

z-intervals were considered in Subsection 1.2 of Unit 8, and t-intervals in Subsection 4.3 of that unit.

The solution to Activity 22 argued that a z-interval would be an appropriate confidence interval for my average number of daily steps. You will calculate the z-interval in the next activity.

Activity 23 *Confidence interval for the mean daily steps*

The sample mean number of steps for the 51 days with data was 9820, and the sample standard deviation was 5415. Each day's activity can be considered to be independent of any other day's activity.

(a) Calculate an approximate 95% confidence interval for my average number of daily steps.

(b) Fitness trackers usually set a default goal of 10 000 steps per day. Can I conclude that, on average, I am meeting the daily goal of 10 000 steps?

You saw in the solution to Activity 23 that it's plausible that, on average, I meet the goal of 10 000 steps per day. (Hooray!) However, as is clear from Figure 6, I don't achieve this goal every day. I would, however, like to meet the daily goal at least half of the time. For the 51 daily step counts shown in the histogram of Figure 6, the step count was at least 10 000 on 22 of the days: this is less than half of the daily counts. (Oh dear!) Does this mean that the proportion of days that my daily steps meet the goal of 10 000 is actually less than 0.5? A hypothesis test will be used to test this in the next activity.

Activity 24 *Proportion of days goal not met*

(a) Letting p be the proportion of days that I meet the daily goal of 10 000 steps, specify suitable null and alternative hypotheses for testing whether this proportion is less than 0.5.

(b) Given that I met the goal of 10 000 steps on 22 of the 51 days, use a p-value to carry out a large-sample test of the hypotheses that you specified in part (a). State your conclusions clearly.

Testing a proportion was covered in Subsection 3.2 of Unit 9, and p-values were covered in Section 4 of that unit.

Well, it's a bit of a relief that the solution to Activity 24 concluded that the data do not suggest that the proportion of days that I meet the daily goal of 10 000 steps is less than 0.5. However, this analysis has shown me that I am actually less active than I thought I was.

Since writing this analysis (in December 2016), I have made a conscious effort to be more active, and I recorded my daily steps for 1–31 January 2017.

Activity 25 *Comparing daily step data*

Figure 7 shows a comparative boxplot for my daily steps in October–November 2015 and in January 2017.

Comparative boxplots were covered in Subsection 5.3 of Unit 1.

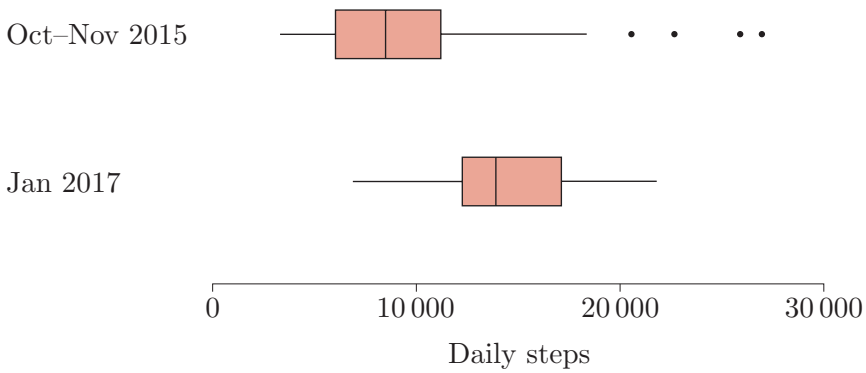


Figure 7 A comparative boxplot of daily steps

Use the comparative boxplot to compare the daily steps for the two time periods.

From the solution to Activity 25, it looks like I have indeed become more active. (Hooray!) Since I have made a conscious effort to be more active, I am confident that the average number of daily steps is greater than 10 000. However, I have now become more ambitious and I would like to test whether my average number of daily steps is actually now greater than 12 000. You will investigate this test problem in the next activity.

Activity 26 *More than 12 000 daily steps on average?*

Figure 8 shows a histogram of daily steps for the 31 days in January 2017.

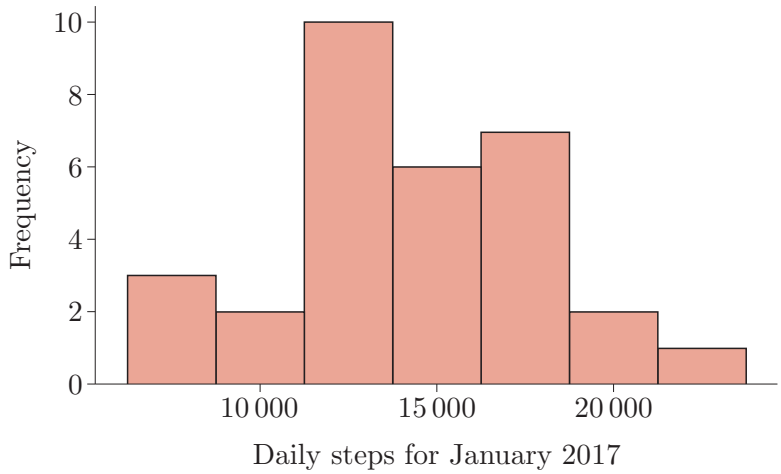


Figure 8 A histogram of daily steps for January 2017

The t-test and z-test were covered in Subsection 3.1 of Unit 9, while Wilcoxon's signed rank test was covered in Subsection 1.1 of Unit 10.

(a) In my view, the histogram in Figure 8 and the comparative boxplot in Figure 7 suggest that neither the t-test nor Wilcoxon's signed rank test may be appropriate for testing whether my average number of daily steps is now greater than 12 000. What aspect of the distribution of the data as portrayed in Figures 7 and 8 have I perceived to lead me to this conclusion? Why would I be content to employ the z-test instead?

p-values were covered in Section 4 of Unit 9.

(b) The sample mean for the daily step data for January 2017 was 14 121 while the sample standard deviation was 3719. Use a p-value to carry out a z-test of the hypotheses

$$H_0 : \mu = 12\,000, \quad H_1 : \mu > 12\,000,$$

where μ is my (current) mean number of daily steps. State your conclusions from this test clearly.

(c) A colleague looked at the comparative boxplot in Figure 7 and the histogram in Figure 8, and took a different view: he thought that any skewness in the distribution of the data is sufficiently small that an assumption of normality of these data is justified. This implies that a t-test may be appropriate for testing whether my average number of daily steps is now greater than 12 000.

Carry out a t-test of the hypotheses

$$H_0 : \mu = 12\,000, \quad H_1 : \mu > 12\,000,$$

where μ is my (current) mean number of daily steps. State your conclusions from this test clearly.

You should be able to obtain a range of possible values rather than an exact value for the p-value of this test.

Activity 26 highlights a couple of points that bear repeating about statistical analysis:

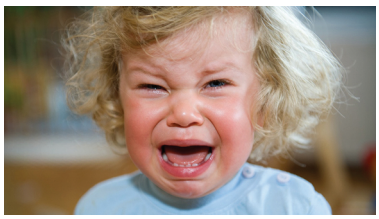
- there is an element of subjectivity in statistical modelling such that even experienced statisticians will sometimes make different modelling assumptions
- results of a statistical analysis often do not depend crucially on precise modelling assumptions, especially when, as above, the choice between different assumptions is somewhat marginal.

In addition, in the case of Activity 26, results using z- and t-tests are especially similar because the sample size, $n = 31$, is not too small. In this particular situation, the effect of different modelling assumptions on the outcome of interest is minor and would have remained so even if the data were more clearly non-normal.

Anyway, I'm delighted to see that there is strong evidence to suggest that my average number of daily steps is now greater than 12 000! And then, quite genuinely, in February 2017, this happened!

7 Expressed emotion

The *expressed emotion index* is a measure of the emotional climate of families with mentally ill members. Expressed emotion can be high or low: in families with high expressed emotion there is yelling, shouting, fighting, or critical or hostile comments. Studies suggest that patients living with



Toddlers frequently express emotion!

relatives scoring low on the expressed emotion index are less likely to relapse than those living with relatives who score high.

In a study of the relationship between expressed emotion and schizophrenia in Spain, a sample of 60 patients was followed up for two years after a psychiatric evaluation. One patient dropped out of the study after twelve months, leaving only 59.

At the initial evaluation, the families of the patients were scored on an expressed emotion scale. Families were then categorised as being either high expressed emotion families or low expressed emotion families. Table 6 shows the number of patients who relapsed during the two-year follow-up period for each of the groups of high and low expressed emotion families.

Table 6 Family expressed emotion and patient relapse

	Family expressed emotion	
	High	Low
Relapse	16	17
No relapse	12	14

(Source: Montero, I. et al. (1992) 'The influence of family expressed emotion on the course of schizophrenia in a sample of Spanish patients', *British Journal of Psychiatry*, vol. 161, no. 2, pp. 217–22)

We are interested both in the overall proportion of patients that relapse within two years of evaluation, and in the difference between the proportions relapsing in the two types of families. In Activity 27, you will calculate an approximate confidence interval to find a plausible range of values for the overall proportion of patients who relapse. You will then consider the difference between the two proportions in Activity 28.

Activity 27 *Overall proportion relapsing*

Confidence intervals for proportions were covered in Subsection 3.2 of Unit 8.

Estimate the overall proportion of patients that relapse, and calculate an approximate 95% confidence interval for this proportion. Is it plausible that half of all patients relapse?

Activity 28 *Difference between proportions relapsing*

(a) Calculate the observed proportion of patients relapsing in the high expressed emotion families, and the observed proportion relapsing in the low expressed emotion families.

(b) Calculate an approximate 95% confidence interval for the difference between the proportions of patients who relapse in the two types of families. What do you conclude from this interval about the effect of family expressed emotion on propensity to relapse in people with schizophrenia?

Confidence intervals for differences between two proportions were covered in Subsection 3.3 of Unit 8.

From the solution to Activity 28 we could not conclude from these data that the proportions of patients relapsing is not the same for the high and low expressed emotion families. However, as mentioned at the start of this section, other studies suggest that patients are less likely to relapse in families with low expressed emotion, which would lead to the proportion relapsing for low expressed emotion being lower than that for high expressed emotion families. Differences in conclusions between different studies do sometimes occur. This can be purely due to random variation, especially if the sample sizes are not very large – after all, one sample is unlikely to be exactly identical to another. But differences in conclusions may also be due to underlying differences between the studies. For example, perhaps the age of patients is relevant to the outcome, or perhaps the patient's gender, or some other factor, is relevant. Without further information, we cannot draw any further conclusions.

8 Norway spruce

The Norway spruce is a large evergreen coniferous tree native to Northern, Central and Eastern Europe. The data in Table 7 were collected as part of a study into how the heights and ages of trees are related. The heights (in feet) and the ages (in years) of 25 Norway spruce are given. In this section we will use linear regression to predict the height of a Norway spruce from its age.

Cones developing on a Norway spruce tree

Table 7 Age and height of Norway spruce

Age (years)	Height (feet)	Age (years)	Height (feet)
15	18	10	13
12	24	13	16
12	15	6	11
10	12	4	9
14	27	4	14
8	15	20	29
16	23	25	39
5	14	24	30
6	5	36	35
9	28	38	41
9	12	22	26
11	20	20	34
10	24		

Activity 29 *Response and explanatory variable*

If we are interested in predicting the height of a Norway spruce given its age, which is the response variable and which is the explanatory variable?

A scatterplot of height against age is shown in Figure 9.

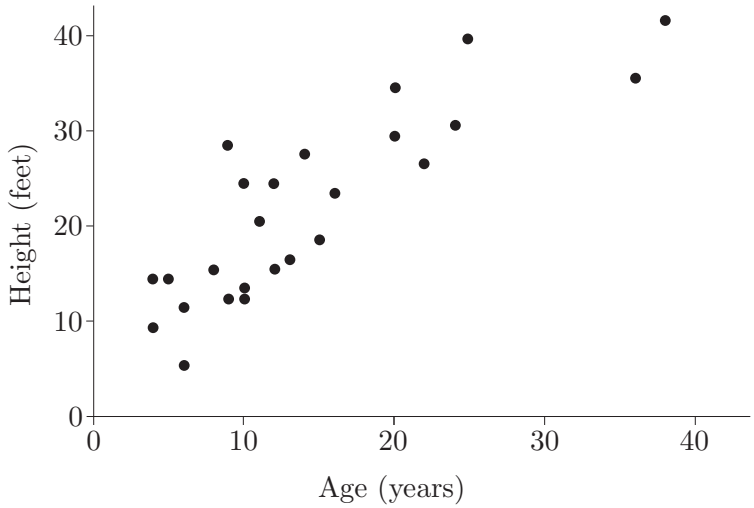


Figure 9 A scatterplot of height against age

The pattern in the plot suggests that it might be reasonable to fit a straight line to the data. You are asked to calculate the equation of the least squares line in Activity 30.

Activity 30 *The least squares line*

Suppose we wish to model the relationship between height Y and age x by a linear regression model. The summary statistics for the data in Table 7 are given by

$$n = 25, \quad \sum x_i = 359, \quad \sum y_i = 534,$$

$$\sum x_i^2 = 7135, \quad \sum x_i y_i = 9485.$$

Details of how to calculate the least squares line were given in Subsection 2.3 of Unit 11.

(a) Use the summary statistics to calculate S_{xx} and S_{xy}.

(b) Calculate the equation of the least squares line for the data.

In the solution to Activity 30, you saw that the fitted regression model for the data in Table 7 is

$$\text{height} = 8.18 + 0.92 \times \text{age}. \tag{3}$$

After fitting a regression model like this, it is important to check that the model assumptions are reasonable. You will do this in the next activity.

Activity 31 *Checking the model assumptions*

A residual plot is shown in Figure 10(a) and a normal probability plot of the residuals is shown in Figure 10(b).

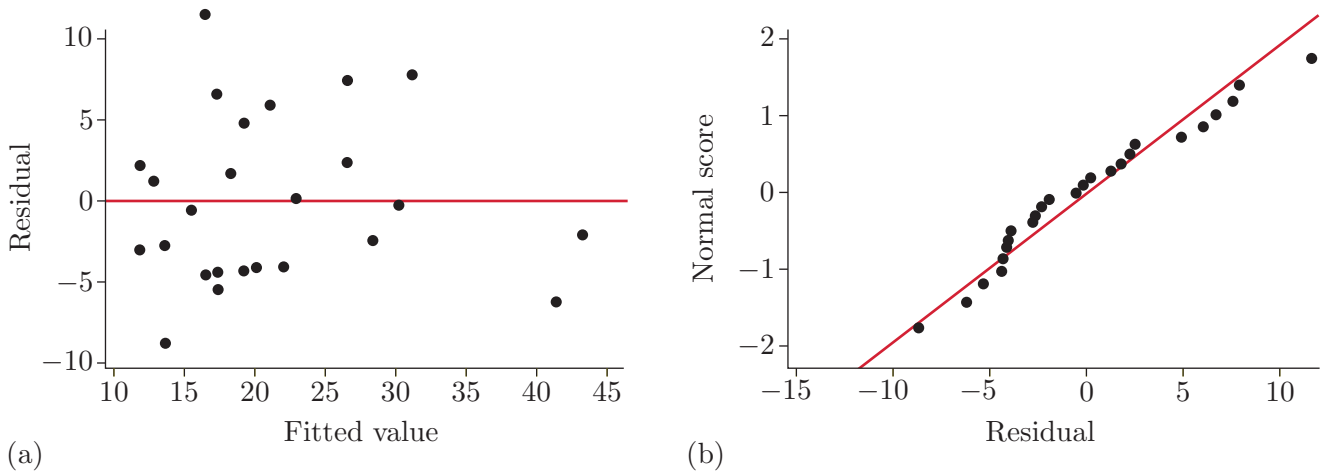


Figure 10 Checking the model assumptions: (a) residual plot; (b) normal probability plot

Comment on what these plots tell you.

Checking the linear regression model assumptions is covered in Section 3 of Unit 11.

The 'point' estimate of the slope parameter β in Equation (3) is 0.92. In the next activity, you will calculate a confidence interval for this parameter.

Activity 32 *A confidence interval for the slope β*

For the fitted regression model given in Equation (3), the residual sum of squares is given by

$$\sum(y_i - \widehat{y}_i)^2 = 626.580.$$

(a) Find an estimate of σ^2.

(b) Hence calculate a 95% confidence interval for the slope parameter β of the line.

How to find an estimate of σ^2 and a confidence interval for the slope parameter were covered in Subsections 4.1 and 4.3, respectively, of Unit 11.

One of the uses of a fitted regression model is to make predictions for specific values of the explanatory variable. You will use the model for this purpose next.

Activity 33 *Predicting the height*

According to the fitted regression model given in Equation (3), what is the predicted height of a 20-year-old Norway spruce?

In Activity 33, you used the least squares line for the data to predict the height of a 20-year-old Norway spruce. However, this doesn't give any indication of the uncertainty concerning the prediction. You will consider this uncertainty in the next activity.

Activity 34 *Confidence intervals and prediction intervals*

Confidence and prediction intervals in linear regression were covered in Subsection 4.3 of Unit 11.

(a) Calculate a 95% confidence interval for the mean height of 20-year-old Norway spruce trees.

(b) Calculate a 95% prediction interval for the height of an individual 20-year-old Norway spruce tree.

(c) Explain why, in the context of regression, a 95% prediction interval for the response must always be wider than a 95% confidence interval for the mean response at a specified value of the explanatory variable.

Activities 33 and 34 calculated a 'point' prediction and prediction interval, respectively, for a 20-year-old Norway spruce. So could the fitted model be used to predict the height of any Norway spruce given the tree's age? You will consider this question in relation to predicting the height of a particular Norway spruce in the next activity.

Activity 35 *Trafalgar Square Christmas tree*

The Trafalgar Square Christmas tree

Every Christmas since 1947, the city of Oslo, Norway, has given a Norway spruce tree as a gift to the people of Britain as a token of gratitude for British support for Norway in the Second World War. The tree is displayed in Trafalgar Square, London, and is decorated with hundreds of lights. The Trafalgar Square Christmas tree is typically 50–60 years old and over 20 metres (roughly $65\frac{1}{2}$ feet) tall.

(a) Suppose that the next Trafalgar Square Christmas tree is 58 years old. Explain why it may not be appropriate to use the fitted linear regression model for the data in Table 7 to predict the height of this tree.

(b) Now use the least squares line given by Equation (3) to make a point prediction of the height of a 58-year-old Norway spruce. Given what we said about the height of the Trafalgar Square Christmas tree, does the prediction surprise you?

As noted in the solution to Activity 35, it may not be appropriate to use the fitted linear regression model to predict the height for Norway spruce which are older than those in the sample used to fit the model. Equally, it may not be appropriate to use the model to predict the height of a Norway spruce which is younger than those in the sample. In particular, since the predicted height of a zero-year-old tree would be 8.18 feet according to the model, the model is clearly unreliable for small ages!

9 School performance

In England, school pupils aged 15–16 years take examinations giving GCSE (General Certificate of Secondary Education) or equivalent qualifications in a number of subjects. As well as their obvious importance to the pupils taking the exams, the results of these qualifications are used to measure school performance. In this section, we will consider two performance measures for schools in England which were introduced in 2016, called 'Attainment 8' and 'Progress 8'.

GCSEs with a numerical grading system have since been introduced in England.

Attainment 8 measures the achievement of 15–16-year-old pupils across 8 GCSE or equivalent qualifications including mathematics and English (both double-weighted), 3 further qualifications in specific subjects (namely, science subjects, computer science, history, geography and languages), and 3 further qualifications from a list of Department for Education approved qualifications. For the data considered here, each individual qualification is graded by a single letter, and each grade corresponds to a numerical 'score'. A school's Attainment 8 is calculated as the average of these grade scores across all pupils using their best 8 qualifications satisfying the subject criteria described above. Schools can calculate their Attainment 8 from their pupils' results.

Progress 8 is a type of value-added measure. Each individual pupil's Attainment 8 score is compared with the average Attainment 8 score of all pupils who had similar prior achievement: the higher the Progress 8 score, the greater the progress made by the pupil in comparison to other pupils with similar prior achievement. The school's Progress 8 score is the average of the individual pupil Progress 8 scores. Individual schools do not have the data available to be able to calculate their Progress 8 score, and the Progress 8 scores for all schools are calculated by England's Department for Education and are published around 3 months after the GCSE results are published.

In this section, we will investigate whether it is possible to use multiple regression to predict a school's Progress 8 score (the response variable Y) using two explanatory variables. We will use the Attainment 8 score as the first explanatory variable x_1, and the second explanatory variable, x_2, is the prior attainment of the cohort of 15–16-year-old pupils in each school as measured by the average point score obtained by these pupils at age 10–11 in national tests (the 'Key Stage 2 SATs').

The data used here were obtained from the Department for Education in England. (Source: https://www.compare-school-performance.service.gov.uk/download-data, data downloaded in February 2017.) We will consider data for 2016 for three types of state school:

- selective schools, which select their pupils based on academic achievement or aptitude
- comprehensive schools, which do not select their pupils based on academic achievement or aptitude
- secondary modern schools, which also do not select their pupils based on academic achievement or aptitude, but are usually found in areas where there are selective state schools – as such, secondary modern schools generally have a lower average prior attainment than comprehensive schools because some of the pupils with high prior attainment attend the local selective school instead.

Some schools do not have data available for all three variables (for example, the results for small schools are not published because individual pupils could be identified), and these schools have been excluded from the analysis.

Activity 36 *Scatterplots*

Figures 11 and 12 show scatterplots of the response Y (Progress 8) against x_1 (Attainment 8) and x_2 (attainment at age 10–11), respectively, for all three types of school. In each scatterplot, different colours are used for data points relating to the three different types of school.

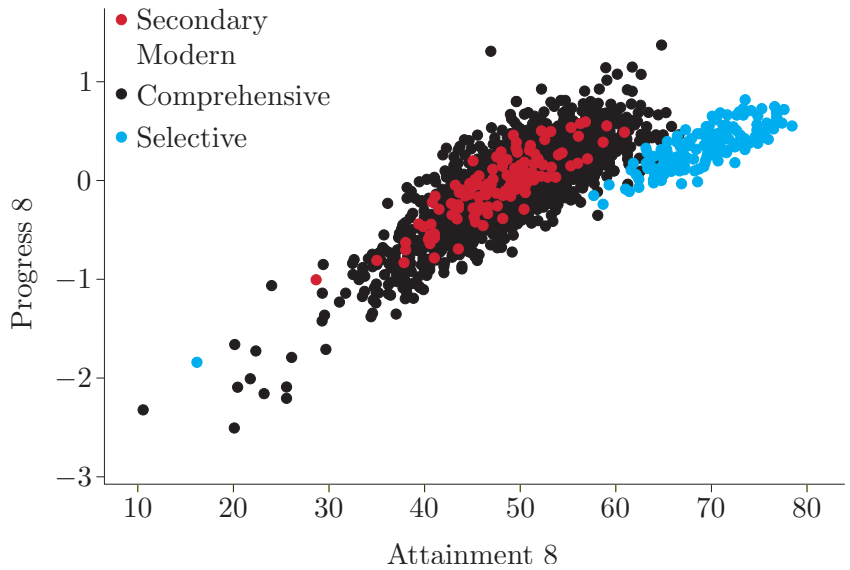


Figure 11 Scatterplot of Y (Progress 8) against x_1 (Attainment 8)

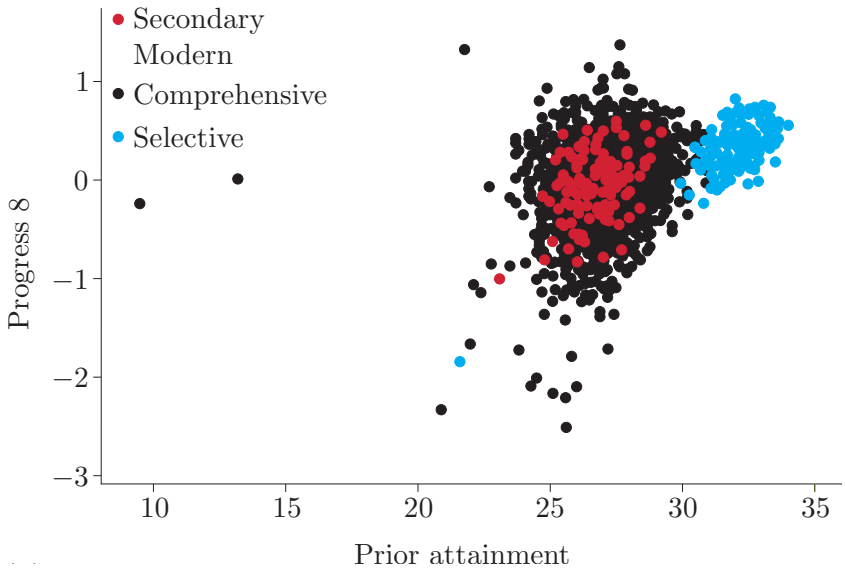


Figure 12 Scatterplot of Y against x_2 (prior attainment)

Explain why fitting a single multiple regression model using data from all three types of school wouldn't be appropriate.

You saw in the solution to Activity 36 that it would not be appropriate to use a single multiple regression model for all three types of school. We could, however, try fitting separate multiple regression models for the different school types. We will start by considering selective schools only: there are 164 such schools in the dataset.

Figure 13 shows scatterplots of the response Y (Progress 8) against x_1 (Attainment 8) and x_2 (prior attainment) for selective schools.

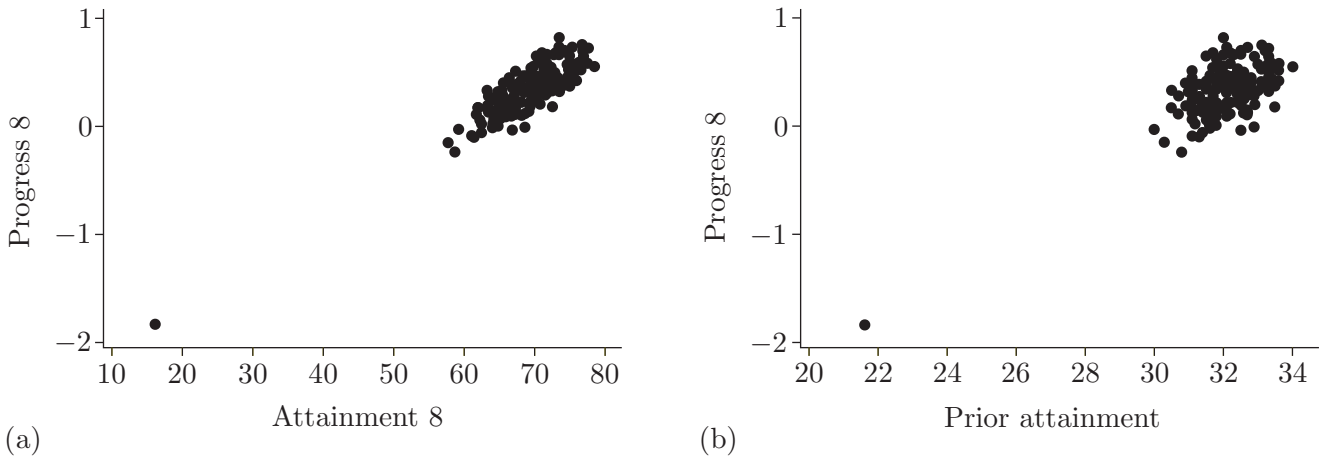


Figure 13 Scatterplots of (a) Y (Progress 8) against x_1 (Attainment 8), and (b) Y against x_2 (prior attainment), for selective schools only

Dominating both plots is an outlier in the bottom left-hand corner of each: this is a school with very low scores for each of Y, x_1 and x_2 in comparison to the other selective schools. Given that this school is so unlike the other selective schools, we will drop this outlier from the analysis in case this single anomalous data point influences the model unduly. (There turns out to be a good reason why this school is so unlike the others so that, unlike in Subsection 3.3 of Unit 12, it does not seem necessary to report results both with and without this school.)

Multiple regression was covered in Section 5 of Unit 11.

Some of the Minitab output when fitting the multiple regression model to the data for the remaining 163 selective schools is as follows.

```
Coefficients

Term                    Coef  SE Coef  T-Value  P-Value   VIF
Constant               5.104    0.139    36.68    0.000
Attainment 8         0.09831  0.00122    80.61    0.000  5.14
Prior attainment   -0.36004  0.00653   -55.16    0.000  5.14

Regression Equation

Progress 8 = 5.104 + 0.09831 Attainment 8 - 0.36004 Prior attainment
```

You will interpret what this output tells you about the multiple regression model in the next activity.

Activity 37 *Multiple regression for selective schools*

(a) Use the Minitab output to write down the fitted multiple regression model for selective schools in terms of x_1, x_2 and Y.

(b) Explain why this analysis suggests that Attainment 8 and prior attainment together influence Progress 8.

(c) Interpret the regression coefficients.

We now have a fitted multiple regression model and we established in Activity 37 that both explanatory variables have regression coefficients which are not zero and therefore should remain in the model. Next the model will be used to predict the Progress 8 scores for several schools.

Activity 38 *Fitted values and residuals*

(a) The 78th school in the dataset achieved a Progress 8 score of 0.00, which means that, on average, the attainment of pupils at this school is as expected compared with pupils of similar prior attainment. The Attainment 8 score was 64.8 and the prior attainment for this school was 31.8. Calculate the fitted value of Progress 8 for this school and

its associated residual. Did the model over- or under-predict the Progress 8 score for this school?

(b) The first school in the dataset achieved the highest Attainment 8 score of 78.5 and also had the highest prior attainment of 34.0. The Progress 8 score was 0.55. How well did the model predict the Progress 8 score for this school?

(c) The Attainment 8 score for the seventh school in the dataset is 69.4, which is the same as the median Attainment 8 score across selective schools. The prior attainment and Progress 8 scores for this school were 32.2 and 0.35, respectively. Calculate the residual for this school and comment on your result.

Of course, exam results can't measure *all* student achievement. Here, school pupils are on an expedition associated with the Duke of Edinburgh award scheme.

In Activity 39, you will consider the residual plot and normal probability plot of residuals for the fitted model to decide whether the model assumptions seem reasonable.

Activity 39 *Are the model assumptions reasonable?*

Figure 14 shows the residual plot and normal probability plot of residuals for the fitted model for selective schools.

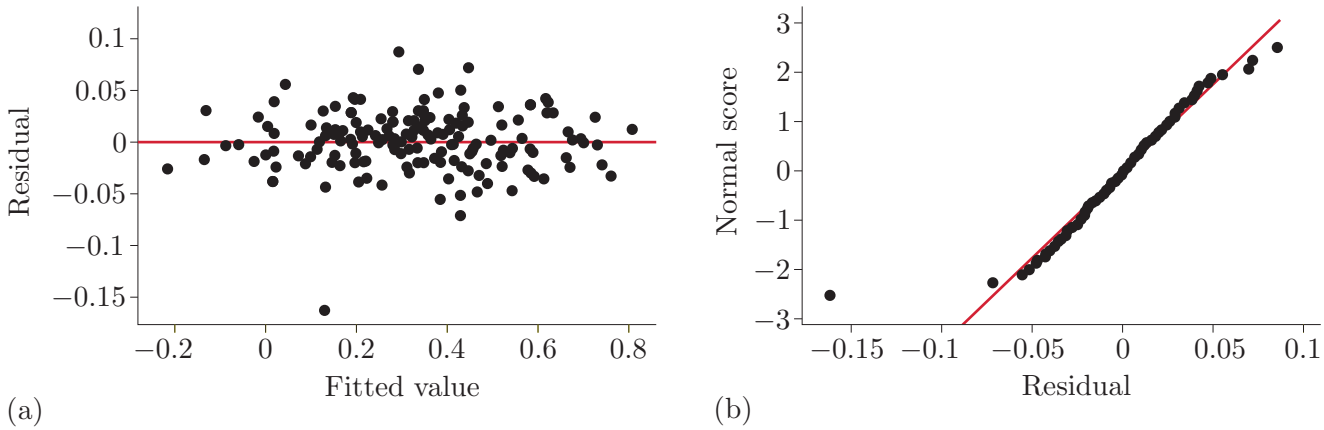


Figure 14 Checking the model assumptions: (a) residual plot; (b) normal probability plot of residuals

Do the model assumptions seem reasonable?

You have seen in Activities 37–39 how multiple regression can be used to model Progress 8 score using the Attainment 8 score and prior attainment as explanatory variables for selective schools. What about comprehensive schools and secondary modern schools? Do similar multiple regression models work well for those types of school as well?

Part of the Minitab output when a multiple regression model is fitted for the 2791 comprehensive schools in the dataset is as follows.

```
Coefficients

Term                    Coef    SE Coef  T-Value  P-Value   VIF
Constant              1.6982     0.0538    31.58    0.000
Attainment 8        0.088236   0.000569   155.11    0.000  2.17
Prior attainment    -0.22425    0.00260   -86.16    0.000  2.17

Regression Equation

Progress 8 = 1.6982 + 0.088236 Attainment 8 - 0.22425 Prior attainment
```

For the 117 secondary modern schools in the dataset, the corresponding Minitab output is as follows.

```
Coefficients

Term                    Coef  SE Coef  T-Value  P-Value   VIF
Constant               2.826    0.129    21.90    0.000
Attainment 8         0.09898  0.00127    77.95    0.000  2.82
Prior attainment    -0.28432  0.00645   -44.07    0.000  2.82

Regression Equation

Progress 8 = 2.826 + 0.09898 Attainment 8 - 0.28432 Prior attainment
```

You will consider the models for the comprehensive and secondary modern schools in the next activity.

Activity 40 *Fitted models for comprehensive and secondary modern schools*

Comment on the fitted multiple regression models for comprehensive and secondary modern schools in comparison to the fitted model for selective schools. In which ways are the models similar, and in which ways are they different?

Despite the similarities between the fitted models for comprehensive and secondary modern schools that were observed in the solution to Activity 40, there is actually a big difference between them: residual plots and probability plots of residuals (not shown) show that while the multiple regression model fits well for the secondary modern dataset, it does not for the comprehensive schools. We will therefore proceed, in the next activity, to consider further the fitted multiple regression models only for the secondary modern and selective schools.

Now, as you saw in the solution to Activity 40, the fitted multiple regression model for secondary modern schools is similar, although not identical, to that for selective schools. This means that schools with the same or similar Attainment 8 and prior attainment scores could have *different* predicted Progress 8 scores depending on the type of school. This is illustrated in the next activity.

Activity 41 *Predicting Progress 8 scores: secondary modern and selective schools*

The fitted multiple regression models for secondary modern and selective schools are as follows.

Secondary modern schools:

$$y = 2.826 + 0.09898\,x_1 - 0.28432\,x_2.$$

Selective schools:

$$y = 5.104 + 0.09831\,x_1 - 0.36004\,x_2.$$

(a) One of the secondary modern schools has a prior attainment score of 29.2. The Attainment 8 score for this school is 60.9, and the predicted Progress 8 score is approximately 0.552. Calculate the predicted Progress 8 score for a selective school with the same prior attainment and Attainment 8 scores.

(b) One of the selective schools has a prior attainment score of 30.0. The Attainment 8 score for this school is 59.3, and the predicted Progress 8 score is 0.133. Calculate the predicted Progress 8 score for a secondary modern school with the same prior attainment and Attainment 8 scores.

(c) Comment on your results.

The situation in which there is a slightly different model for the same variables under different circumstances (in our case here, different types of school) arises frequently in practice. We coped with it here by fitting three separate models for the three different types of school. There is, however, a neater way of doing things by using a single model which allows for differing parameter values for the different types of school by treating the type of school as a so-called factor. This model is beyond the scope of this module, but is something that you are likely to meet if you go on to study regression further.

10 And finally . . .

So this is the end of the unit, and indeed the end of the module! We very much hope that you have enjoyed it. The module has explored the fundamental statistical techniques and ideas used for analysing and interpreting data, and you should now have the basic skills required to start to make sense of data. The module also provided the necessary foundations required for studying statistics further if you wish to do so. In this age of data, there are a huge number of exciting datasets out there ready to be gathered and explored, and we hope that we have enthused you to do so!

Summary

In this unit, you have used various statistical methods developed in the module to explore several datasets. You have:

- considered various probability models for datasets, including the discrete uniform, Poisson, exponential and normal models, and one with a particular polynomial p.d.f., as well as the Poisson process
- derived a likelihood function and the associated maximum likelihood estimate
- calculated several confidence intervals, including a z-interval and confidence intervals for a proportion and the difference between two proportions
- carried out several tests, including the t-test, Wilcoxon signed rank test, chi-squared goodness-of-fit test and testing a proportion
- used linear regression, including obtaining a least squares line, checking model assumptions, calculating confidence and prediction intervals, and using multiple regression.

Learning outcomes

After you have worked through this unit, you should be able to:

- appreciate the wide variety of possible applications that the techniques developed in the module can be applied to
- be more confident in your application of the techniques developed in the module
- appreciate that there may be uncertainty as to the most suitable statistical technique to use when there is ambiguity over the validity of normality or other model assumptions
- appreciate that in the real world, a statistical analysis doesn't always give a clear-cut result
- appreciate that in the real world, it may not be possible to fully answer the question of interest with the data available.

Solutions to activities

Solution to Activity 1

(a) Each chocolate bar can be one of 7 types coded $1, 2, \ldots, 7$. The range of X is therefore $\{1, 2, 3, 4, 5, 6, 7\}$.

(b) If it is reasonable to expect equal numbers of each type of chocolate bar, then, since the range of X is $\{1, 2, 3, 4, 5, 6, 7\}$, a suitable distribution is the discrete uniform distribution with parameters $m = 1$ and $n = 7$. Therefore, from Equations (7) and (8) in Unit 3, respectively, X has probability mass function

$$p(x) = \frac{1}{7 - 1 + 1} = \frac{1}{7}, \quad x = 1, 2, \ldots, 7,$$

and cumulative distribution function

$$F(x) = \frac{x - 1 + 1}{7 - 1 + 1} = \frac{x}{7}, \quad x = 1, 2, \ldots, 7.$$

Solution to Activity 2

(a) From the solution to Activity 1(b),

$$p(x) = \frac{1}{7}, \quad x = 1, 2, \ldots, 7.$$

So in a tub of 71 chocolate bars, the expected frequency of each type of chocolate bar is

$$E_i = 71 \times \frac{1}{7} \simeq 10.14, \quad i = 1, 2, \ldots, 7.$$

(b) The chi-squared goodness-of-fit test uses the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where k is the number of categories, which in this case is 7. So the observed value of the test statistic is

$$\begin{aligned}\chi^2 &= \frac{(17 - 10.14)^2}{10.14} + \frac{(10 - 10.14)^2}{10.14} + \frac{(12 - 10.14)^2}{10.14} + \frac{(15 - 10.14)^2}{10.14} \\ &\quad + \frac{(5 - 10.14)^2}{10.14} + \frac{(4 - 10.14)^2}{10.14} + \frac{(8 - 10.14)^2}{10.14} \\ &\simeq 4.641 + 0.002 + 0.341 + 2.329 + 2.605 + 3.718 + 0.452 \\ &= 14.088 \simeq 14.09.\end{aligned}$$

If the discrete uniform model is correct, then the distribution of the test statistic is approximately $\chi^2(k - p - 1) = \chi^2(6)$, since k is the number of categories (in this case 7) and p is the number of estimated parameters (in this case 0, since the parameters of this discrete uniform distribution are known).

Comparing the observed value of 14.09 with the quantiles of the $\chi^2(6)$ distribution, from the table of chi-squared distribution quantiles in the Handbook, we see that the test statistic lies between the 0.95-quantile (which is 12.59) and the 0.975-quantile (which is 14.45). So the p-value lies between 0.025 and 0.05. (The exact p-value happens to be 0.029, but you cannot work this out from the table, nor do you need to.) There is therefore moderate evidence against the null hypothesis that the discrete uniform distribution with parameters $m = 1$ and $n = 7$ is a suitable model for X.

In the interpretational terms of Table 3 of Unit 9, it is therefore also the case that $0.01 < p \leq 0.05$.

Solution to Activity 3

(a) A suitable graph is a bar chart such as that in Figure 15. (This is because the data are discrete.) It is easy to draw by hand.

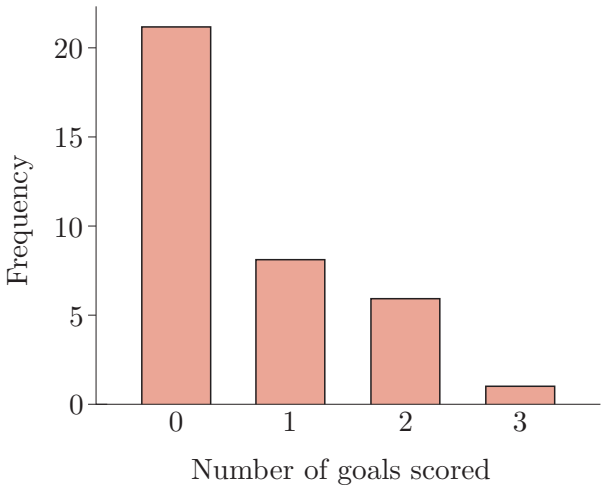


Figure 15 A bar chart for goals scored

(b) A Poisson model seems appropriate. Reasons include the following:

- the possible values for X are counts
- the range of X starts at 0 and doesn't have any theoretical upper value
- the bar chart of the data has a shape which is consistent with a Poisson distribution with a small mean; it is decreasing and right-skew.

Solution to Activity 4

(a) From Equation (6) of Unit 3, the Poisson p.m.f. is

$$p(x;\theta) = \frac{e^{-\theta}\theta^x}{x!}.$$

So, following Example 17 of Unit 7, the likelihood is

$$\begin{aligned} L(\theta) &= p(0;\theta)^{21} \times p(1;\theta)^8 \times p(2;\theta)^6 \times p(3;\theta)^1 \\ &= \left(\frac{e^{-\theta}\theta^0}{0!}\right)^{21} \times \left(\frac{e^{-\theta}\theta^1}{1!}\right)^8 \times \left(\frac{e^{-\theta}\theta^2}{2!}\right)^6 \times \frac{e^{-\theta}\theta^3}{3!} \\ &= e^{-21\theta} \times e^{-8\theta}\theta^8 \times \frac{e^{-6\theta}\theta^{12}}{2^6} \times \frac{e^{-\theta}\theta^3}{6} \\ &= \frac{e^{-36\theta}\theta^{23}}{384}, \end{aligned}$$

as required.

(b) Using Equation (9) of Unit 7 to differentiate a product, we have

$$\begin{aligned} L'(\theta) &= \frac{1}{384}\left(-36e^{-36\theta} \times \theta^{23} + e^{-36\theta} \times 23\theta^{22}\right) \\ &= \frac{e^{-36\theta}\theta^{22}}{384}\left(-36\,\theta + 23\right). \end{aligned}$$

To find $\widehat{\theta}$, we now need to solve $L'(\theta) = 0$. All but the linear term in brackets is irrelevant to solving $L'(\theta) = 0$ because, for $\theta > 0$, $e^{-36\theta}\theta^{22}/384 > 0$. The linear term has a single value of θ at which it is zero, so that value must be the MLE $\widehat{\theta}$: $\widehat{\theta}$ satisfies

$$-36\,\theta + 23 = 0,$$

so

$$\widehat{\theta} = \frac{23}{36} \simeq 0.639.$$

Solution to Activity 5

(a) The Poisson(0.639) p.m.f. is

$$p(x) = \frac{e^{-0.639}0.639^x}{x!}.$$

So

$$p(X = 0) = \frac{e^{-0.639}0.639^0}{0!} \simeq 0.5278,$$

and the expected frequency of scoring 0 goals in 36 scores, E_0, is, to two decimal places,

$$E_0 \simeq 36 \times 0.5278 \simeq 19.00.$$

Similarly,

$$p(X = 1) = \frac{e^{-0.639}0.639^1}{1!} \simeq 0.3373,$$

and the expected frequency of scoring 1 goal in 36 scores, E_1, is, to two decimal places,

$$E_1 \simeq 36 \times 0.3373 \simeq 12.14.$$

Finally,

$$P(X \geq 2) = 1 - \{P(X = 0) + P(X = 1)\}$$
$$= 1 - e^{-0.639} - 0.639e^{-0.639} \simeq 0.1349.$$

So the expected frequency of scoring ≥ 2 goals in 36 scores, E_2, is, to two decimal places,

$$E_2 \simeq 36 \times 0.1349 \simeq 4.86.$$

(b) For the chi-squared goodness-of-fit test to be valid, a rough rule of thumb is that the expected frequencies for each category need to be 5 or more. The expected frequency for the category '≥ 2 goals' is 4.86, so if there were separate categories '2 goals', '3 goals', and so on, the expected frequencies for these categories would be too small. As it is, the expected frequency is just under the rule of thumb value of 5. However, as this is a rule of thumb and the expected frequency is only just under 5, the test should be just about valid. (Further combining categories into just two groups would not work; we will mention why at the end of the solution to the next part of this activity.)

(c) The chi-squared goodness-of-fit test uses the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where k is the number of categories. So the observed value of the test statistic is

$$\chi^2 = \frac{(21 - 19.00)^2}{19.00} + \frac{(8 - 12.14)^2}{12.14} + \frac{(7 - 4.86)^2}{4.86}$$
$$\simeq 0.211 + 1.412 + 0.942 = 2.565 \simeq 2.57.$$

If the Poisson model is correct, then the distribution of the test statistic is approximately $\chi^2(k - p - 1) = \chi^2(1)$, since k is the number of categories (in this case 3) and p is the number of estimated parameters (in this case 1, since the parameter of the Poisson distribution has been estimated as 0.639).

We are using a 5% significance level, so the 0.95-quantile of the $\chi^2(1)$ distribution is required: from Table 18 in Unit 10 or the table in the Handbook, this is 3.84. Since the observed value of the test statistic is 2.56, which is less than the 0.95-quantile of $\chi^2(1)$, there isn't sufficient evidence to reject the hypothesis that Poisson(0.639) is a suitable model for the data in Table 2.

It seems that the Poisson model is a suitable model for the data in Table 2.

(Had we combined categories into just $k = 2$ groups, the number of degrees of freedom for the χ^2 test would have reduced to $k - p - 1 = 2 - 1 - 1 = 0$. This reflects the fact that there would be too little data left to test goodness-of-fit appropriately.)

Solution to Activity 6

Major tsunamis can occur at any time, and the waiting time between tsunamis need not be a whole number of months. Therefore an exponential distribution, which is continuous, is more appropriate for modelling the waiting times than a geometric distribution, which is discrete.

Solution to Activity 7

An exponential model does seem reasonable because:

- the shape of the histogram is not inconsistent with the data coming from an exponential distribution: there is a single peak at 0 and the frequencies generally decrease with increasing waiting times
- the sample mean and sample standard deviation are close together in value.

Solution to Activity 8

Since the mean time between major tsunamis is 26 months, an estimate of the parameter λ is

$$\widehat{\lambda} = \frac{1}{26} \simeq 0.038.$$

(This is, in fact, the maximum likelihood estimate of λ.)

Solution to Activity 9

(a) There are 12 months in one year, so, letting X denote the waiting time between two successive major tsunamis,

$$\begin{aligned} P(\text{waiting time is at least one year}) &= P(X \geq 12) = 1 - F(12) \\ &= 1 - \left(1 - e^{-\frac{12}{26}}\right) = e^{-\frac{12}{26}} \\ &\simeq 0.630. \end{aligned}$$

(The exponential distribution's cumulative distribution function is given in Equation (2) of Unit 5.)

(b) $$\begin{aligned} P(\text{waiting time is less than 6 months}) &= P(X < 6) = F(6) \\ &= 1 - e^{-\frac{6}{26}} \simeq 0.206. \end{aligned}$$

(c) The expected number of waiting times of at least one year is

$$e^{-\frac{12}{26}} \times 29 \simeq 18.28,$$

and the expected number of waiting times of less than 6 months is

$$\left(1 - e^{-\frac{6}{26}}\right) \times 29 \simeq 5.98.$$

For the observed data, 16 were waiting times of at least one year, and 8 were waiting times of less than 6 months. The expected values from the model are therefore not totally out of line with the data.

Solution to Activity 10

(a) The estimate of the rate λ of occurrence of major tsunamis per month is

$$\lambda = \frac{1}{26} \simeq 0.038 \text{ per month.}$$

(This is the same value as you calculated in Activity 8.)

(b) X has a Poisson distribution with parameter

$$\lambda t = \left(\frac{1}{26} \text{ per month}\right) \times (12 \text{ months}) = \frac{12}{26} = \frac{6}{13}.$$

(c) (i) From Equation (6) of Unit 3, the probability that exactly two major tsunamis will occur in one year is given by

$$P(X = 2) = \frac{e^{-\frac{6}{13}} \left(\frac{6}{13}\right)^2}{2!} \simeq 0.067.$$

(ii) The probability that at least one major tsunami will occur in one year is given by

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\frac{6}{13}} \simeq 0.370.$$

Solution to Activity 11

(a) There are three assumptions being made:

- major tsunamis occur singly
- the rate of occurrence of major tsunamis remains constant
- the incidence of future major tsunamis is independent of the past.

(b) It is not impossible that more than one major tsunami may occur simultaneously, but the probability of such an event seems to be so small that this assumption may well be reasonable. (The coarseness of the discretisation of the data in Table 3 is such that more than one tsunami might happen within a period of one month. If so, something needs to be done about this, but it does not mean that a Poisson process on the underlying continuous scale is not still a good model.)

However, Figure 2 suggests that the rate at which major tsunamis are occurring may not be constant since the plot doesn't follow a very straight line. It appears from Figure 2 that major tsunamis seem to be occurring more frequently towards the end of the time period. However, it is possible that the rate at which major tsunamis occur is in fact constant, but that the recording of them has improved in recent years, leading to major tsunamis being recorded more frequently.

It seems reasonable, but not unarguable, that the incidence of future major tsunamis is independent of the past, so the third assumption for a Poisson process identified in the solution to part (a) seems appropriate.

Overall, there is some doubt about the reasonableness of a Poisson process as a model for the occurrence of major tsunamis.

Solution to Activity 12

(a) On entry, the median is

$$m = x_{\left(\frac{1}{2}(n+1)\right)} = x_{(4)} = 58.$$

The quartiles are

$$q_L = x_{\left(\frac{1}{4}(n+1)\right)} = x_{(2)} = 50,$$
$$q_U = x_{\left(\frac{3}{4}(n+1)\right)} = x_{(6)} = 62.$$

So the sample interquartile range on entry is

$$q_U - q_L = 62 - 50 = 12.$$

One week later, the median is

$$m = x_{(4)} = 64.$$

The quartiles are

$$q_L = x_{(2)} = 60,$$
$$q_U = x_{(6)} = 80.$$

So the sample interquartile range one week later is

$$q_U - q_L = 80 - 60 = 20.$$

(b) The measurements after one week are generally higher than on entry since the boxplot for this set of measurements is located to the right of the other boxplot. The spread is also greater one week later. (These two features are apparent from the numerical work of part (a) also.)

Solution to Activity 13

(a) The required differences are given in Table 8.

Table 8 Differences 'one week later' minus 'on entry'

Patient	1	2	3	4	5	6	7
Difference	33	2	24	27	4	1	−6

(b) The normality of the differences can be investigated using a normal probability plot. If the points lie close to a straight line, then the normality assumption is plausible.

(c) A (one-sample) t-test on the differences could be used if the assumption of a normal model is plausible. For this test, because we're interested in detecting a difference between the two measurements in either direction, the hypotheses to be tested are

$$H_0 : \mu_D = 0, \quad H_1 : \mu_D \neq 0,$$

where μ_D is the population mean of the differences.

(d) If a normal model is untenable, then the Wilcoxon signed rank test could be used. (Another potential test if a normal model is untenable is the z-test, but that would require a larger sample than we have here – the rule of thumb we have been using is that we would need $n \geq 25$.)

Again, we are interested in detecting a difference in either direction, but for this nonparametric test the hypotheses involve the population median differences m_D (rather than the population mean differences as considered in the t-test). The hypotheses to be tested are

$$H_0 : m_D = 0, \quad H_1 : m_D \neq 0.$$

Solution to Activity 14

(a) The value of the test statistic is

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{12.14 - 0}{15.38/\sqrt{7}} \simeq 2.088.$$

(b) The null distribution of the t-test is $t(n-1)$, where n is the sample size. Therefore, for the data with sample size $n = 7$, the null distribution is $t(6)$.

(c) Since $0.05 < p < 0.1$, there is only weak evidence against H_0. We conclude that the data provide weak evidence to suggest that, on average, there is a difference between CO transfer factor levels on entry and one week later, and since t is positive, any difference is such that CO transfer factor levels are higher one week later than on entry.

Solution to Activity 15

(a) The Wilcoxon signed rank test statistic can be calculated using Table 9.

Table 9 Differences 'one week later' minus 'on entry'

Patient	1	2	3	4	5	6	7
Difference	33	2	24	27	4	1	−6
Sign of difference	+	+	+	+	+	+	−
Absolute value of difference	33	2	24	27	4	1	6
Rank of absolute value	7	2	5	6	3	1	4

The test statistic w_+ is the sum of the ranks associated with the positive differences, so

$$w_+ = 7 + 2 + 5 + 6 + 3 + 1 = 24.$$

(b) Since $p > 0.1$, there is little or no evidence to suggest that the median difference between CO transfer factor levels on entry and one week later is not zero. The data therefore suggest that there is little or no evidence that there is any difference in the CO transfer factor levels.

Solution to Activity 16

With only seven data points, it is very difficult to say whether the assumption of normality of the differences is plausible or not. The data do roughly lie about the straight line, so normality can't be ruled out, but equally, the points are certainly not so close to the line as to give overwhelming support to the assumption of normality either!

Solution to Activity 17

For a function f to be a valid p.d.f., it needs to be non-negative and to integrate to 1 over its range.

Non-negativity over its range is clear from Figure 5. Alternatively, notice that every element that is multiplied together to form f is itself non-negative for $0 < x < 1$.

As for its integral over its range, we have

$$\begin{aligned}\int_0^1 f(x)\,dx &= \int_0^1 12x^2(1-x)\,dx = 12\int_0^1 x^2(1-x)\,dx \\ &= 12\int_0^1 (x^2-x^3)\,dx = 12\left[\frac{x^3}{3}-\frac{x^4}{4}\right]_0^1 \\ &= 12\left(\frac{1}{3}-\frac{1}{4}-(0-0)\right) = 12\times\frac{1}{12} = 1,\end{aligned}$$

as required.

Solution to Activity 18

For $0 < x < 1$, the c.d.f. is given by

$$\begin{aligned}F(x) &= \int_0^x f(y)\,dy = \int_0^x 12y^2(1-y)\,dy = 12\int_0^x (y^2-y^3)\,dy \\ &= 12\left[\frac{y^3}{3}-\frac{y^4}{4}\right]_0^x = 12\left(\frac{x^3}{3}-\frac{x^4}{4}-(0-0)\right) \\ &= 12x^3\left(\frac{1}{3}-\frac{x}{4}\right) = x^3(4-3x).\end{aligned}$$

Solution to Activity 19

(a) Because X is continuous,

$$P(X < 0.5) = P(X \le 0.5) = F(0.5) = (0.5)^3(4-3\times 0.5) = 0.3125.$$

(b)
$$\begin{aligned}P(0.3 \le X \le 0.7) &= F(0.7) - F(0.3) \\ &= (0.7)^3(4-3\times 0.7) - (0.3)^3(4-3\times 0.3) \\ &= 0.6517 - 0.0837 = 0.568.\end{aligned}$$

(c) The required probability is $P(X \ge 0.6)$:

$$\begin{aligned}P(X \ge 0.6) &= 1 - P(X < 0.6) = 1 - F(0.6) \\ &= 1 - (0.6)^3(4-3\times 0.6) = 1 - 0.4752 = 0.5248.\end{aligned}$$

Solution to Activity 20

$$\begin{aligned}\alpha = E(X) &= \int_0^1 x\,f(x)\,dx = \int_0^1 12x^3(1-x)\,dx = 12\int_0^1 (x^3-x^4)\,dx \\ &= 12\left[\frac{x^4}{4}-\frac{x^5}{5}\right]_0^1 = 12\left(\frac{1}{4}-\frac{1}{5}-(0-0)\right) \\ &= 12\times\frac{1}{20} = \frac{3}{5} = 0.6.\end{aligned}$$

Solution to Activity 21

(a) We will use the relationship $V(X) = E(X^2) - \{E(X)\}^2$. To do so, we first need to calculate $E(X^2)$:

$$\begin{aligned} E(X^2) &= \int_0^1 x^2 f(x)\,dx = \int_0^1 12x^4(1-x)\,dx = 12\int_0^1 (x^4 - x^5)\,dx \\ &= 12\left[\frac{x^5}{5} - \frac{x^6}{6}\right]_0^1 = 12\left(\frac{1}{5} - \frac{1}{6} - (0-0)\right) \\ &= 12 \times \frac{1}{30} = \frac{2}{5} = 0.4. \end{aligned}$$

Therefore

$$V(X) = E(X^2) - \{E(X)\}^2 = 0.4 - (0.6)^2 = 0.04.$$

(b) We will use the relationship $S(X) = \sqrt{V(X)}$:

$$S(X) = \sqrt{V(X)} = \sqrt{0.04} = 0.2.$$

Solution to Activity 22

(a) The histogram is unimodal with a single peak around 6000 steps. It is also right-skew, since the bars on the right-hand side of the histogram fall away more slowly than those on the left-hand side.

(b) In order to use a t-interval, the data need to be normally distributed. However, a normal model is not plausible for these data because they are right-skew.

On the other hand, a z-interval can be used for any data regardless of their distribution, as long as the sample size is large enough. The sample size of $n = 51$ is reasonably large, so a z-interval would be more appropriate for these data.

Solution to Activity 23

(a) From Equation (3) of Unit 8, an approximate 95% confidence interval for the mean is the z-interval given by

$$(\mu^-, \mu^+) = \left(\overline{x} - z\,\frac{s}{\sqrt{n}}, \overline{x} + z\,\frac{s}{\sqrt{n}}\right),$$

where z is the 0.975-quantile of $N(0,1)$. So

$$\mu^- = 9820 - 1.96 \times \frac{5415}{\sqrt{51}} \simeq 8334,$$

$$\mu^+ = 9820 + 1.96 \times \frac{5415}{\sqrt{51}} \simeq 11\,306.$$

Thus an approximate 95% confidence interval for my mean number of daily steps is $(8334, 11\,306)$.

Because the numbers are in the 1000s, the confidence interval limits were rounded to the nearest whole number.

(b) The confidence interval calculated in part (a) contains the value 10 000, so it is one of the plausible values for my mean number of daily steps. Therefore it is plausible that, on average, I am indeed meeting the daily goal of 10 000 steps.

Solution to Activity 24

(a) The hypotheses to be tested are

$$H_0 : p = 0.5, \quad H_1 : p < 0.5.$$

(b) The sample size of $n = 51$ is reasonably large, so the hypotheses specified in part (a) can be tested using the test statistic

$$Z_p = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where the null distribution for Z_p is $N(0, 1)$. For these data, $\widehat{p} = 22/51$ so

$$z_p = \frac{\frac{22}{51} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{51}}} \simeq -0.980.$$

This is a one-sided test, so the p-value is calculated as

$$p = P(Z_p \leq -0.980),$$

where $Z_p \sim N(0, 1)$. So

$$\begin{aligned} p &= P(Z_p \leq -0.980) = P(Z_p \geq 0.980) = 1 - P(Z_p < 0.980) \\ &= 1 - \Phi(0.98) = 1 - 0.8365 = 0.1635. \end{aligned}$$

Since $p = 0.1635 > 0.1$, the p-value provides little or no evidence against H_0. We conclude that the data do not suggest that the proportion of days that I meet the goal of 10 000 steps is less than 0.5.

Solution to Activity 25

The daily steps for January 2017 are generally higher than those for October–November 2015 since the boxplot for January 2017 is located to the right of the other boxplot. The spread for January 2017 is generally similar to the spread for the first set of data, with the exception of four large outliers for the 2015 data.

Solution to Activity 26

(a) Both the histogram and the boxplot suggest to me that a symmetric model for the data may not be appropriate. However, symmetry of the underlying distribution is an assumption of Wilcoxon's signed rank test, while a normal model (and hence symmetry) is an assumption of the t-test. Thus neither of these tests may be appropriate for testing whether the average number of daily steps is now greater than 12 000.

The z-test doesn't have any such assumptions, and can therefore be used for these data because the sample size is large enough (i.e. ≥ 25 by the rule of thumb).

(b) The observed value of the z-test statistic is

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{14\,121 - 12\,000}{3719/\sqrt{31}} \simeq 3.175.$$

This is a one-sided test, so the p-value is calculated as

$$p = P(Z \geq 3.175),$$

where $Z \sim N(0, 1)$. So

$$p = P(Z \geq 3.175) = 1 - P(Z < 3.175) = 1 - \Phi(3.175)$$
$$\simeq 1 - \Phi(3.18) = 1 - 0.9993 = 0.0007.$$

Since $p < 0.01$ there is strong evidence against H_0, so there is strong evidence that the average number of daily steps is greater than 12 000.

(c) The observed value of the t-test statistic is the same as the z-test statistic of part (b): $t \simeq 3.175$. But in this case the p-value is calculated as

Both z- and t-test statistics have the formula $(\overline{x} - \mu_0)/(s/\sqrt{n})$.

$$p = P(t \geq 3.175),$$

where $T \sim t(30)$; here, the degrees of freedom parameter of the null distribution has been calculated as $n - 1 = 31 - 1 = 30$. Now, from the table of quantiles of the t-distribution in the Handbook, the observed value of t lies between the 0.995-quantile of the $t(30)$ distribution (which is 2.750) and the 0.999-quantile of the $t(30)$ distribution (which is 3.385). As we are conducting a one-sided test, the p-value lies between 0.001 and 0.005.

Again, since $p < 0.01$ there is strong evidence against H_0, so there is strong evidence that the average number of daily steps is greater than 12 000.

Solution to Activity 27

The overall proportion relapsing is

$$\widehat{p} = \frac{16 + 17}{16 + 12 + 17 + 14} = \frac{33}{59} \simeq 0.559.$$

An approximate 95% confidence interval for p, the underlying proportion that relapse, is then

$$\left(\widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) = \left(0.559 \pm 1.96\sqrt{\frac{0.559 \times 0.441}{59}}\right)$$

$$\simeq (0.559 - 0.127, 0.559 + 0.127) \simeq (0.432, 0.686).$$

Since the confidence interval contains the value 0.5, it is indeed plausible that half of all patients relapse.

Solution to Activity 28

(a) The proportion of patients in high expressed emotion families who relapsed was

$$\frac{16}{16 + 12} = \frac{16}{28} \simeq 0.571,$$

and the proportion in low expressed emotion families was

$$\frac{17}{17 + 14} = \frac{17}{31} \simeq 0.548.$$

(b) An estimate for $d = p_1 - p_2$, the difference between the proportions, where p_1 and p_2 are the proportions for high and low expressed emotion families, respectively, is

$$\widehat{d} = \widehat{p}_1 - \widehat{p}_2 = 0.571 - 0.548 = 0.023.$$

An approximate 95% confidence interval for d is

$$\left(\widehat{d} \pm 1.96\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}\right)$$

$$= \left(0.023 \pm 1.96\sqrt{\frac{0.571 \times 0.429}{28} + \frac{0.548 \times 0.452}{31}}\right)$$

$$\simeq (0.023 - 0.254, 0.023 + 0.254) \simeq (-0.231, 0.277).$$

This interval gives a range of plausible values for the true difference between the proportions relapsing. Since the interval contains zero, as well as both positive and negative values, we cannot conclude from these data that family expressed emotion is related to propensity to relapse.

Solution to Activity 29

If we wish to predict height from age, then age should be regarded as the explanatory variable and height as the response variable.

Solution to Activity 30

(a) Using the summary statistics provided, S_{xx} and S_{xy} are given by

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 7135 - \frac{359^2}{25} = 1979.76,$$

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 9485 - \frac{359 \times 534}{25} = 1816.76.$$

(b) The least squares estimates of β and α are

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1816.76}{1979.76} \simeq 0.9177 \simeq 0.92$$

and

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = \frac{534}{25} - \frac{1816.76}{1979.76} \times \frac{359}{25} \simeq 8.1823 \simeq 8.18.$$

The equation of the least squares line is therefore

$$y = 8.18 + 0.92\,x.$$

That is, the fitted model is

$$\text{height} = 8.18 + 0.92 \times \text{age}.$$

Solution to Activity 31

The points in the residual plot seem to be scattered about zero in a random, unpatterned fashion. Therefore the assumption that the residuals have constant, zero mean and constant variance seems reasonable.

The points lie very roughly along the straight line in the normal probability plot, indicating that the assumption that the data come from a normal sample might be reasonable. It could, on the other hand, be argued that there is a hint of some systematic variation around the line; however, this doesn't look serious enough to rule out the normality assumption.

Solution to Activity 32

(a) An estimate of σ^2 is given by

$$s^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2} = \frac{626.580}{23} \simeq 27.243.$$

(b) A 95% confidence interval for β is given by

$$\left(\widehat{\beta} - t\frac{s}{\sqrt{S_{xx}}}, \widehat{\beta} + t\frac{s}{\sqrt{S_{xx}}}\right),$$

where t is the 0.975-quantile of $t(n-2)$.

Here, $n = 25$ and the 0.975-quantile of $t(23)$ is 2.069. So the confidence interval is (using the unrounded value of $\widehat{\beta}$)

$$\left(0.9177 - 2.069\frac{\sqrt{27.243}}{\sqrt{1979.76}}, 0.9177 + 2.069\frac{\sqrt{27.243}}{\sqrt{1979.76}}\right)$$

$$\simeq (0.9177 - 0.2427, 0.9177 + 0.2427) \simeq (0.68, 1.16).$$

Solution to Activity 33

The predicted height (in feet) of a 20-year-old Norway spruce is

$$8.18 + 0.92 \times 20 = 26.58.$$

Solution to Activity 34

(a) A 95% confidence interval for the mean response at the value x_0 is given by

$$\left(\widehat{\alpha} + \widehat{\beta}\,x_0 - t\,s\sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n}}, \widehat{\alpha} + \widehat{\beta}\,x_0 + t\,s\sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n}}\right),$$

where t is the 0.975-quantile of $t(n-2)$.

When $x_0 = 20$,

$$\widehat{\alpha} + \widehat{\beta}\,x_0 = 8.1823 + 0.9177 \times 20 = 26.5363.$$

Also, $t = 2.069$ and $s = \sqrt{27.243}$ (from the solution to Activity 32(a)), $S_{xx} = 1979.76$ (from the solution to Activity 30(a)) and

$$\overline{x} = \frac{359}{25} = 14.36.$$

Hence the required confidence interval is

$$\left(26.5363 \pm 2.069\sqrt{27.243}\sqrt{\frac{(20 - 14.36)^2}{1979.76} + \frac{1}{25}}\right)$$

$$\simeq (26.5363 \pm 2.5571) \simeq (23.98, 29.09).$$

More decimal places were retained in the estimated regression coefficients here than in the solution to Activity 33 because of the greater complexity of the current calculations. The estimates to 4 decimal places can be found in the solution to Activity 30(b).

(b) The 95% prediction interval for the response at x_0 is given by

$$\left(\widehat{\alpha}+\widehat{\beta}\,x_0 - t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}}+\frac{1}{n}+1}, \widehat{\alpha}+\widehat{\beta}\,x_0 + t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}}+\frac{1}{n}+1}\right).$$

So the required prediction interval is

$$\left(26.5363 \pm 2.069\sqrt{27.243}\sqrt{\frac{(20-14.36)^2}{1979.76}+\frac{1}{25}+1}\right)$$

$$\simeq (26.5363 \pm 11.0977) \simeq (15.44, 37.63).$$

(c) A confidence interval is for the mean response at a given value of the predictor, whereas a prediction interval is for an individual response. Since the latter involves the additional variability of individual observations around the mean, the prediction interval must be wider.

Solution to Activity 35

(a) The ages of the trees in Table 7 range from 4 to 38 years. At 58 years, the particular Norway spruce being considered is much older than the trees in the dataset used to fit the linear regression model. As such, we do not know whether the fitted linear regression model will continue to be appropriate for larger values of x than those used to fit the model. For example, do older Norway spruce grow at a slower rate than younger Norway spruce? If so, then it's possible that the fitted model may not be appropriate for older trees.

(b) The predicted height (in feet) of a 58-year-old Norway spruce is

$$8.18 + 0.92 \times 58 = 61.54.$$

Perhaps surprisingly, the predicted height is rather less than the observed height. An explanation might be that the Trafalgar Square Christmas tree is, presumably, not any old, randomly chosen, 58-year-old Norway spruce but one carefully nurtured and/or chosen for its shape and height. (A height of $65\frac{1}{2}$ feet is, however, well within the 95% prediction interval associated with this point prediction, which is very wide.)

Solution to Activity 36

In the scatterplots in Figures 11 and 12, the data points for selective schools are in a different location to those for comprehensive and secondary modern schools. As such, it looks like it might not be appropriate to use a single multiple regression model for all three types of school, since the relationships between the response and explanatory variables are not the same for all types of school.

Solution to Activity 37

(a) The fitted multiple regression model is

$$y = 5.104 + 0.09831\,x_1 - 0.36004\,x_2.$$

(b) From the `Coefficients` table, the p-value for each individual two-sided test of the null hypothesis $H_0 : \beta_j = 0$, for $j = 1, 2$, is 0.000, which means that for each regression coefficient $p < 0.01$. There is therefore strong evidence that each regression coefficient is non-zero, which in turn implies that together Attainment 8 (x_1) and prior attainment (x_2) influence Progress 8 (Y).

(c) The regression coefficients can be interpreted as follows.

- Regression coefficient for x_1: If the value of Attainment 8 increases by one unit, and the value of prior attainment remains fixed, then Progress 8 would be expected to increase by 0.09831. The fact that there would be expected to be an increase in Progress 8 for fixed prior attainment makes sense because if two schools have the same prior attainment, then the school with the higher Attainment 8 score will have made more progress than the school with the lower Attainment 8 score.
- Regression coefficient for x_2: If the value of prior attainment increases by one unit, and the value of Attainment 8 remains fixed, then Progress 8 would be expected to decrease by 0.36004. The decrease is indicated by the negative coefficient. This makes sense because if two schools have the same Attainment 8, then the school with the higher prior attainment will have made less progress than the school with the lower prior attainment.

Solution to Activity 38

(a) The fitted value of Progress 8 for the 78th school is

$$\widehat{y}_{78} = 5.104 + 0.09831 \times 64.8 - 0.36004 \times 31.8 \simeq 0.0252 \simeq 0.03.$$

Therefore the residual for this school is

$$w_{78} \simeq 0.00 - 0.03 = -0.03.$$

Since the residual is negative, the model over-predicted the Progress 8 score for this school, although only slightly, since the residual is small.

(b) The fitted value of Progress 8 for the first school is

$$\widehat{y}_1 = 5.104 + 0.09831 \times 78.5 - 0.36004 \times 34.0 \simeq 0.5800 \simeq 0.58.$$

Therefore the residual for this school is

$$w_1 \simeq 0.55 - 0.58 = -0.03.$$

Since the residual is negative, the model over-predicted the Progress 8 score for this school as well, although also only by a small amount (in fact, by the same amount).

(c) The fitted value of Progress 8 for the seventh school is

$$\widehat{y}_7 = 5.104 + 0.09831 \times 69.4 - 0.36004 \times 32.2 \simeq 0.3334 \simeq 0.33.$$

Therefore the residual for this school is

$$w_7 \simeq 0.35 - 0.33 = 0.02.$$

Since the residual is positive, the model under-predicted the Progress 8 score for this school, but not by very much.

Solution to Activity 39

There is one very clear outlier with a low residual value in both of these plots. This school's actual Progress 8 score is much lower than that predicted by the model, given its Attainment 8 and prior attainment scores. Other than this value, the points in the residual plot seem to be scattered about zero in a random, unpatterned fashion, and in the normal probability plot the points lie quite close to the straight line (although there is perhaps the hint of a slight systematic 'S' shape in the normal probability plot). Overall, the model assumptions do seem reasonable.

Solution to Activity 40

The fitted multiple regression models for both comprehensive and secondary modern schools are similar to that for selective schools in that there is still a positive regression coefficient for x_1 (Attainment 8) and a negative one for x_2 (prior attainment), and these again are both significantly different from zero because their p-values are reported as being 0.000 for each model. However, the actual values of the regression parameters are different for each type of school to those in the fitted model for selective schools.

Solution to Activity 41

(a) The predicted Progress 8 score for a selective school with the same Attainment 8 and prior attainment scores is

$$\widehat{y} = 5.104 + 0.09831 \times 60.9 - 0.36004 \times 29.2 \simeq 0.578.$$

(b) The predicted Progress 8 score for a secondary modern school with the same Attainment 8 and prior attainment scores is

$$\widehat{y} = 2.826 + 0.09898 \times 59.3 - 0.28432 \times 30.0 \simeq 0.166.$$

(c) For the Attainment 8 and prior attainment scores of the secondary modern school considered in part (a), the predicted Progress 8 score increases from that given by the model for secondary modern schools when made under the model that assumes it to be a selective school.

For the Attainment 8 and prior attainment scores of the selective school considered in part (b), the predicted Progress 8 score increases from that given by the model for selective schools when made under the model that assumes it to be a secondary modern school.

It seems that both schools would benefit a little from swapping their statuses! There is nothing here to suggest that either secondary modern or selective status has a uniformly helpful effect on pupils from schools with these particular, similar, levels of Attainment 8 and prior attainment scores.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 149: © Nathan Jones

Page 151: © Vasilis Ververidis / www.123rf.com

Page 153: © FlyAkwa / https://commons.wikimedia.org/wiki/File:Tsunami_Phuket.jpg This file is licensed under the Creative Commons Attribution-Noncommercial-ShareAlike Licence http://creativecommons.org/licenses/by-sa/4.0/

Page 157: © BSIP / Universal Images Group

Page 161: © DVersiga84 / https://www.flickr.com/photos/61505200@N02/32531292440/in/photostream

Page 163: © Stephen VanHorn / www.123rf.com

Page 168: © arnoaltix / iStock / Getty Images Plus

Page 169: © Miroslav Pinkava / www.123rf.com

Page 172: © Diliff / https://commons.wikimedia.org This file is licensed under the Creative Commons Attribution-Noncommercial-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 173: © Photofusion / UIG

Page 177: © 2017 South Bromsgrove High Academy Trust. Reproduced by permission

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index