Chapter 22

Complex analyses and computers

chapter 22

Undertaking data analysis by computer

Throughout this book I have been explaining how to perform a range of statistical analyses. So the next piece of advice may seem a little unexpected: don't use up your valuable time undertaking statistical analysis when you can get a computer to do it for you! There are many excellent statistics programs, such as SPSS (see Hinton *et al*., 2004), the calculations are done quickly with a degree of consistent accuracy that we can rarely match as human beings. The key point I hope to have made in the book is that it is important to know why and how statistical analysis operates, the reasoning behind it, the assumptions made and the types of data that particular analyses can deal with. This knowledge not only allows you to perform the calculations with a calculator but it is also invaluable when using a computer. If you do not understand what you are doing then using a computer simply compounds the problem. When performing, say, a *t* test by hand you might learn something about the operation and logic of the test but with a computer the test gets 'magically' done and the result appears like a rabbit out of a hat. If you didn't know what you were doing beforehand, you certainly will not be any the wiser afterwards. It is only when we know what we are doing that the computer comes into its own. The person who understands statistical analysis can appreciate what the computer is doing, and more importantly, know when it is NOT DOING what is really wanted.

A key thing to remember when using a computer is the acronym GIGO – garbage in, garbage out. If you put a lot of nonsense into the computer you will get a lot of nonsense out! Computers do not know when you have made a mistake, in fact they do not 'know' anything, they simply do as they are told. If you choose the wrong analysis, or type in the wrong data, the computer program will still perform the analysis on that data. If you do not realise your mistake then you can unknowingly take away the results of an incorrect or inappropriate analysis. If this is for an important research programme with much depending on the results then the ramifications of your mistake may be profound.

Errors in data input

There are a number of checks you can perform to make sure that you have input the data correctly into a computer program or computer file for subsequent analysis. The first thing to do is to obtain a printout of the data after it has been input into the computer. You will then be able to make a check on the data that was actually analysed rather than the data you hoped was analysed. Look at the printout and ask yourself the following questions.

1 Are there any large numbers where you did not expect them? If you leave your finger on a key for too long you might input that digit twice by mistake. Check that numbers that should be 2 are not 22 or even 222.
2 Are there missing values where there should not be? When reading down a list of numbers to input, it is quite possible to miss one out. Check that the correct number of figures have been input.
3 Does the pattern of data look correct? Often with a large amount of data you can see patterns on the page of numbers, such as all 1s in a particular column. As you scan down the data is there an unusual figure somewhere? If so, check that it is correct rather than an error on input.
4 Has the data been input in the correct order for the analysis? This can be a very important question. If the analysis is complex such as a two factor mixed design ANOVA the data must be input in the correct order. If not, the computer might analyse the data for the independent factor as though it were the repeated measures factor and vice versa.

Interpreting output

Once the computer has performed the analysis the program will present a display or a printout of the results of the analysis. When interpreting this analysis keep in mind one question: is this what you expected when you input the data? If not then why were your expectations out? This illustrates why knowledge of statistical analysis is so useful. If you know that a certain analysis cannot produce what you have obtained then you know there is an error somewhere, whereas someone who has no knowledge of statistical analysis might simply accept the result as correct.

The first area to check is the means, totals, standard deviations, etc. You may already have worked out the means of the various conditions before performing the analysis. Does the computer come up with the same values? Has it the correct means in the correct conditions? A basic check of the simple calculations can confirm that the data has been input correctly and the correct numbers are in the appropriate conditions.

Next check that the statistical analysis is the one you wanted. Often the name of the analysis will appear on the output. Does it say 'related or repeated measures' when you really wanted independent? Does it say 'completely randomised or independent measures' when you wanted to perform a repeated measures analysis? Simply looking at the information at the top of the output can often be the most useful. But always make sure you know what analysis you want to perform before you ask the computer to do it!

Occasionally the computer program will have an error in it. The chances of a commercially available one containing a 'bug' are very small but if you are using a helpful little program you downloaded from the Internet (often written by academics then generously offered to others for free) then make sure that the results match your expectations. Recall that there are certain results you should never get, such as a negative value for a sums of squares in an ANOVA summary table. Always check the data first but do not always trust the program.

There are differences between the ways computer programs present the results and the ways it is done when working out the analysis by hand. The most common difference is in the presentation of the significance of a finding. Computer programs often give the actual probability of the result occurring by chance rather than whether it exceeds the significance level or not. For example, rather than stating '$p < 0.05$' or 'significant at $p = 0.05$' the computer might display '$p = 0.034215$', which is the actual probability of the result under the null hypothesis. It is up to you to decide whether this is significant at the significance level you have chosen. A result with a probability 0.034215 is less than 0.05 so is significant at the $p = 0.05$ level of significance but not at $p = 0.01$. Sometimes the computer will output the probability as $p = 0.000000$. This appears to indicate that the result could never occur under the null hypothesis, which is obviously impossible. The true explanation lies in the way the computer displays numbers. As there can never be a probability of zero that the data occurred by chance it must be that the probability is so small that there is not enough space for the computer to display enough decimal places. Therefore we should replace the last zero with a 1 so that we read 0.000000 as 0.000001.

We know that the probability is smaller than this so we are erring on the side of safety in reporting this probability. If we incorrectly reported a probability of zero other researchers would spot the error immediately whereas more correctly reporting a probability of 0.000001 clearly indicates a highly significant result. If you get a probability this low then check that your calculated value of the statistic under test is very large (or very small depending on the test) as we would expect with such a small probability value.

Always be wary of unusual figures, especially ones you did not expect. It is tempting to believe that a highly significant result must be true, particularly if it is a 'better' result than you were hoping for. Do not be seduced by the computer output. Is this really the result you would have expected by looking at the data? In this book, mainly for illustration purposes most of the statistical analyses have been found to be significant. It does not work like this in research. Often there are many non-significant findings. A significant finding is often cherished, particularly as it is more likely to be published than a non-significant finding. Yet we should still treat significant results with some scepticism as, if there is an error, the cost will be that much greater.

Complex analyses

There are a number of statistical analyses that are commonly used today which would have only been undertaken by a statistician in the past. This is due to the development of sophisticated computer programs for statistical analyses and the advance of computer technology. The computing power required to undertake complex analysis would have been owned only by major institutions (such as universities) only two decades ago. And prior to the advent of computers a statistician would have possibly taken days to carry out certain calculations. Now a standard personal computer can undertake these complex analyses in just a few seconds or less. The major time-consuming activity is inputting the data rather than carrying out the analysis. Thus it is outside the scope of this book to provide worked examples for complex analyses that would take forever by hand but which the modern computer can perform in considerably less time than it takes to boil a kettle!

However, the reason why certain complex analyses are now popular is that researchers are able to collect large amounts of data and then examine these data for underlying relationships between the various variables

under study. This is particularly the case when a number of participants are asked to provide scores on a wide range of variables. This might be a study in the laboratory where a group of people are tested on a number of skilled tasks such as logic, mathematical, spatial and verbal tasks. The research aim here is to find out which tasks are related, with the implication that they might rely on the same cognitive processing systems. Alternatively, a consumer questionnaire might be constructed where the questions ask for both a range of background information as well as finding out about the participants' product use and product preference. Indeed, the data layout in statistical computer programs often reflects this format:

	Variable 1	Variable 2	. . .	Variable *k*
Participant 1				
Participant 2				
⋮				
Participant *n*				

An example data input table

In the following analyses I am going to use the data in the table below for illustration purposes. In many real cases – for example questionnaire data – a researcher will have a lot more data, often hundreds if not thousands of data points. This is one reason why we usually would not contemplate undertaking these analyses by hand. However, to demonstrate the analyses the dataset will be small. I am also describing the data in rather a general way, labelling the variables Question 1, Question 2, etc. to again illustrate the wide applicability of the analyses. As long as the data satisfy the assumptions of the test then we can see that the analyses are very versatile and can be used in a number of different instances with a range of research topics.

Participant	Question 1	Question 2	Question 3	Question 4	Question 5
1	1	1	7	8	6
2	3	4	3	3	5
3	3	3	8	7	8
4	4	2	2	1	2
5	5	5	2	2	2
6	7	5	4	5	6
7	7	7	7	7	4
8	6	8	9	9	8
9	9	7	5	5	4
10	8	10	10	9	7
Mean	5.30	5.20	5.70	5.60	5.20
Stand.dev.	2.54	2.82	2.91	2.88	2.20
Variance	6.4556	7.9556	8.4556	8.2667	4.8444

Reliability

When we develop a questionnaire or other measure of a construct (such as 'honesty' or 'verbal ability') we want that measure to be both valid and reliable. A valid measure is one that genuinely measures the underlying construct. This is not always easy to achieve and often there is debate in the literature on the validity of a test, for example, do IQ tests really examine intelligence? Deciding on the validity of a measure is an academic issue rather than one for statistical analysis. However, reliability can be examined statistically.

When data are collected on a number of different measures we may be interested in examining their reliability. Reliability is defined as the ability

of a measuring instrument to measure the concept in a consistent manner. Imagine I had a tape measure and recorded a person's height as 1 metre 65 centimetres. It would be most odd if I measured them a second time with the same tape measure ten minutes later and read off a height of 1 metre 42 centimetres. The tape measure would be a highly unreliable measuring device. Similarly we want a questionnaire to be reliable across people and occasions. One way of testing reliability is to examine the 'test–retest' reliability. Does the test give the same results on different occasions? All we need to do is to give the test twice and correlate the findings. A high correlation indicates a high level of reliability. However, it is not quite as simple as that, as the participants may have remembered their answers from the first test and this might influence the way they respond on the second test. To avoid this some researchers construct two measures (version A and version B of their questionnaire) with slightly different questions which they hope are equivalent. However, this may double the work.

Within a questionnaire (or indeed similarly structured dataset) we can examine the internal reliability of the items within it. If the five questions in the above questionnaire are measuring different aspects of the concept of 'happiness' then we can examine whether participants are responding to the different items in a consistent manner. I have used the term item here rather than question as it is a more general term and the item could be a question or a score on any specific task. Thus, we can examine the internal reliability of our questionnaire by looking at the relationships between the answers to the different questions.

One measure of reliability is called 'split-half' reliability, where the answers on the first half of the questionnaire are compared to the answers on the second half of the questionnaire. So, if there is a high correlation between the two halves of the questionnaire we can argue that there is internal consistency in the questionnaire.

The most popular measure of internal consistency is Cronbach's alpha, which is a more sophisticated test of reliability than the split-half analysis as it examines the average inter-item correlation of the items in the questionnaire. It also takes into account the number of items in the questionnaire:

$$\text{Cronbach's } \alpha = \frac{k}{(k-1)}\left[1 - \frac{\sum \text{var}(i)}{\text{var}(sum)}\right]$$

where k is the number of items, var(i) is the variance of an item, and var(sum) is the variance of the totals for each participant. (In the above example participant 1 has a total of 23, and participant 2 has a total of 18).

Essentially, if all the items are measuring exactly the same thing (without any error), we can refer to this as the 'true score', and the scores will reflect this in the following way: all the individual item variances will be identical and var(*sum*) will simply be $k \times \text{var}(i)$. This will result in $\alpha = 1$. However, at the other extreme, if there is no shared variance in the items, then they are reflecting only 'error' rather than an underlying true score, resulting in var(*sum*) = $\sum \text{var}(i)$ and $\alpha = 0$.

In our example:

$$\alpha = \frac{5}{(5-1)}\left[1 - \frac{6.4556 + 7.9556 + 8.4556 + 8.2667 + 4.8444}{107.7778}\right]$$

$$= 0.8327$$

It is conventional to view an α of 0.7 or greater as indicating a reliable scale, so we would view this limited questionnaire data as reliable.

Interestingly, we can argue that if the items are measuring the same underlying dimension on the same scale then they should have the same variance. If we make this assumption then we can calculate a slightly different Cronbach's alpha, called the standardised Chronbach's alpha, based on the inter-item correlations rather than on item variances. This is expressed as follows:

$$\text{Standardised Cronbach's } \alpha = \frac{k\bar{r}}{1 + \bar{r}(k-1)}$$

where k is the number of items and $\bar{r}$ is the average inter-item correlation. The inter-item correlations for the questionnaire example are shown in the table below, referred to as the correlation matrix.

	Question 1	Question 2	Question 3	Question 4	Question 5
Question 1	1	0.8434	0.1790	0.1551	−0.0517
Question 2	0.8434	1	0.4958	0.4494	0.2434
Question 3	0.1790	0.4958	1	0.9675	0.8090
Question 4	0.1551	0.4494	0.9675	1	0.8217
Question 5	−0.0517	0.2434	0.8090	0.8217	1

There are 10 different correlations (of each question with another question), giving the average iter-item correlation $\bar{r}$, as 0.4913. Thus for our example:

$$\text{Standardised } \alpha = \frac{5 \times 0.4913}{1 + 0.4913 \times (5 - 1)} = 0.8284$$

Notice that there is a small difference between our two alpha values. This is due to the difference in the variances of the items rather than one alpha being 'better' than the other. We would use the standardised alpha when we have comparable items (i.e. measured on the same scale as in the example here) or we have standardised the data, but otherwise we would report the 'raw' value based on the item variances.

A further reason why we undertake the analysis by computer is that we can get a printout of the alpha value when a particular item is removed from the analysis. If we do this for each item in turn then we can see which combination of items gives the highest alpha value, and hence highest reliability. This allows us to refine a questionnaire and maximise its reliability.

Details on how to perform a reliability analysis using the SPSS computer statistical package can be found in Chapter 18 of Hinton *et al*. (2004).

Factor analysis

In the above example we found a high level of reliability of our items in the questionnaire ($\alpha = 0.83$) so we might wish to employ the questionnaire as it is. However, if we had found a low reliability then it would have informed us that the scores on the different items were not varying in a consistent manner. The reason for this might be that different questions are 'tapping' different underlying factors. For example in developing a cognitive test battery where a group of children are given four tests, of arithmetic, geometry, verbal reasoning and story comprehension, we might find that there is a high correlation between the scores on the arithmetic and geometry and a high correlation between the verbal reasoning and story comprehension scores but low correlations between the scores on arithmetic and verbal reasoning, arithmetic and story comprehension, geometry and verbal reasoning, geometry and story comprehension. Thus, arithmetic and geometry scores

are correlated and verbal reasoning and story comprehension are correlated indicating (possibly) two underlying factors that we might label 'mathematical ability' and 'language ability'.

Factor analysis is a procedure that examines the relationship between the scores on the different items and uses the correlations between them to specify where the relationships are strong enough to indicate underlying factors. This is not a procedure that we would wish to undertake by hand. In the past factor analysis would be the domain of statisticians who would take many hours of calculation in order to determine the factors underlying a dataset.

A factor analysis is essentially a data reduction technique as it is used to see whether there is a set of factors that can explain the variation of the variables under study. It is only useful if we can find fewer factors than variables which are able to explain the variation in the data. It can be undertaken for two reasons: exploratory (to discover underlying factors) or confirmatory (to confirm factors already proposed). We shall look at exploratory factor analysis in this example.

The first thing we need to consider is whether the data is suitable for a factor analysis. Essentially we need samples large enough to ensure that the correlations are a good representation of their population values. There are a number of 'rules of thumb' proposed to indicate what constitutes a large enough dataset: there should be at least 200 scores overall, with at least 10 scores per item and at least five times as many subjects as items. There clearly are not enough scores in our example data to satisfy these criteria but we shall continue for the purpose of illustration.

Two useful tests on the data are often carried out before a factor analysis. The Kaiser–Meyer–Olkin (KMO) test examines the data for sampling adequacy. This gives a measure of the common variance amongst the variables that the factors will be able to account for. The KMO statistic ranges from 0 to 1. In our example, the KMO value is 0.655. Any value over 0.6 is regarded as acceptable for a factor analysis as values below this would mean that the factor analysis will not be able to account for much of the variability in the data and so is not worth undertaking.

The second test is the Bartlett's test of sphericity. This examines the correlation matrix (see above). If there was no correlation at all between any of the variables then the values in the correlation matrix would have 1s down the diagonal with all the other values as zero. This is called an identity matrix. Our example gives a Bartlett $\chi^2 = 38.11$, $df = 10$, $p < 0.001$. This indicates that our correlation matrix is significantly different from an identity matrix so there are correlations worth investigating.

Now that we are confident that it is worth proceeding with the factor analysis we undertake a principal component analysis to find the factors. The scores on each item are standardised to a mean of 0 and a standard deviation of 1. Thus, the variance of every item becomes 1. With 5 items the total variance to explain is 5. Factors are then identified. Each factor has an eigenvalue which gives a value or 'weight' of each factor, in terms of the variance explained. These are shown in the following table.

Component	*Eigenvalue*	*Percentage of overall variance*	*Cumulative percentage of variance*
1	3.0838	61.6750	61.6750
2	1.5896	31.7925	93.4675
3	0.1956	3.9113	97.3789
4	0.1020	2.0393	99.4182
5	0.0291	0.5818	100.0000
Total	5.0000	100.0000	

We can see from the table that 5 components or factors have been identified. It is conventional to select only those factors with eigenvalues greater than 1 as an eigenvalue of 1 indicates that a factor can only explain as much variance as a single item. Only the first two factors are selected as their eigenvalues are greater than 1. Notice also that they can explain 61.6750 per cent and 31.7925 per cent of the variance in the items, so 2 factors can explain over 93 per cent of the total variability in the five items.

An alternative way of selecting the important factors is to produce a 'scree plot' of the components against eigenvalues. Imagine the profile of a mountain. If it was a real mountainside the scree falling down the slope would settle at a point where the slope flattens out. In Figure 22.1 this would be at component 3. We then take factors before this 'elbow' in the graph. So, in Figure 22.1, we can identify two factors as important from the scree plot, supporting the choice of factors from the table of eigenvalues.

We can now look at the correlation of each of the items with our two selected factors (shown in the following table, part (a), referred to as the

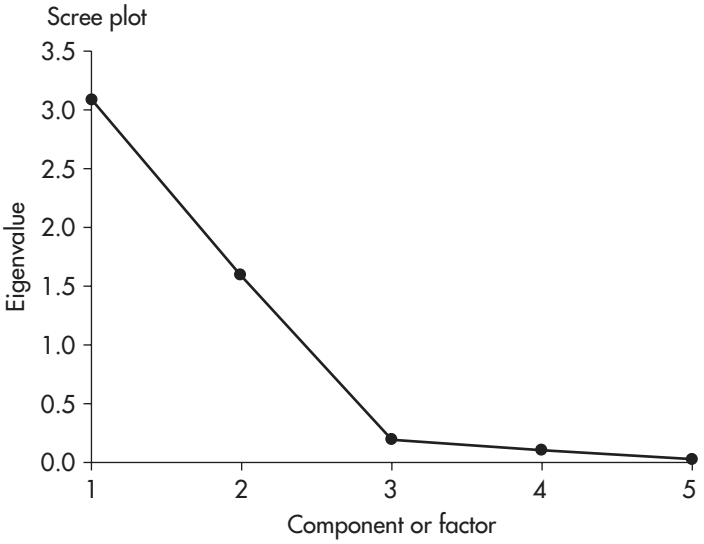

FIGURE 22.1 Scree plot of the eigenvalues

component matrix). Notice that Question 1 correlates 0.4105 with Factor 1 and 0.8850 with Factor 2.

(a)	Unrotated		(b)	Rotated	
	Factor 1	Factor 2		Factor 1	Factor 2
Question 1	0.4105	0.8850		0.0349	0.9750
Question 2	0.6893	0.6858		0.3040	0.9236
Question 3	0.9498	−0.2239		0.9482	0.2306
Question 4	0.9395	−0.2617		0.9561	0.1922
Question 5	0.8096	−0.4663		0.9330	−0.0491

The unrotated values give us some idea of the relationship between items and factors but we can make this much clearer by a procedure called rotation. This rotates our factors to 'line them up' better with the variables. Imagine placing a painting on a wall. You notice it is a little skewed so you rotate it to line it up straight. Rotating factors is a little like this: we are not changing the relationships – simply making them clearer. There are different methods of rotation, with the second version of the component matrix, (b) above,

showing the effect of a varimax rotation which endeavours to produce 1s and 0s in the Factor columns of the component matrix. Now we have a clearer picture with Questions 1 and 2 'loading' onto Factor 2 and Questions 3, 4, and 5 loading onto Factor 1. Question 2 does load onto both Factors but the rotation indicates that Factor 2 is the more important.

Finally we can ask how much of the variance in each of our items can be explained by the two factors we have produced. We can answer this by squaring the correlations in the component matrix and adding them for each item. Should we take the unrotated or the rotated correlations? The answer is that it does not matter: the rotation does not change the factors. I will take the rotated values but you can work them out for the unrotated values if you wish.

Question 1: Variance explained = $(-0.0349)^2 + 0.9750^2 = 0.9518$
Question 2: Variance explained = $0.3040^2 + 0.9236^2 = 0.9455$
Question 3: Variance explained = $0.9482^2 + 0.2306^2 = 0.9522$
Question 4: Variance explained = $0.9561^2 + 0.1922^2 = 0.9511$
Question 5: Variance explained = $0.9330^2 + (-0.0491)^2 = 0.8728$

Remember that the variance in each item has been standardised to 1, so our factors are able to explain a very large amount of the variability in the data. The figures in the final column above are referred to as the communalities, which provide a measure of the variability in that item shared with other items, in our case supporting the factors we have produced.

In conclusion factor analysis examines the correlations between the items in the dataset and produces a set of underlying factors. If we find factors that can explain a lot of the variability in the items then we can argue that our items can be reduced to the fewer factors we have elicited. In our example, the factor analysis revealed two factors, one underlying Questions 3, 4, and 5 and the second factor underlying Questions 1 and 2.

Details on how to perform a factor analysis using the SPSS computer statistical package can be found in Chapter 17 of Hinton *et al*. (2004).

Multivariate analysis of variance (MANOVA)

In many instances of data analysis we wish to compare different groups of participants on our measuring device, such as a questionnaire, to examine

hypotheses such as, 'Are younger adults going to score higher on happiness than older adults?' If we obtain an overall score on our measuring device then the data is suitable for a univariate analysis: that is, analysing a single dependent variable – the participant's score on the test. We can then undertake a univariate test such as a *t* test (if we have two groups of participants) or an analysis of variance (if we have more). However, we may not produce a composite score for the questionnaire but wish to analyse the different questions as separate dependent variables. In this case we could do lots and lots of univariate tests on each separate dependent variable. The problem with this is that we will undertake lots of tests and increase the risk of a Type I error. A solution to this is to perform a multivariate analysis of variance (MANOVA) which allows the analysis of more than one dependent variable. In the table below I have added an additional question from the questionnaire where participants indicate their income level.

Participant	*Income*	*Question 1*	*Question 2*	*Question 3*	*Question 4*	*Question 5*
1	Low	1	1	7	8	6
2	Low	3	4	3	3	5
3	Low	3	3	8	7	8
4	Low	4	2	2	1	2
5	Low	5	5	2	2	2
Group mean		3.20	3.00	4.40	4.20	4.60
6	High	7	5	4	5	6
7	High	7	7	7	7	4
8	High	6	8	9	9	8
9	High	9	7	5	5	4
10	High	8	10	10	9	7
Group mean		7.40	7.40	7.00	7.00	5.80
Overall mean		5.30	5.20	5.70	5.60	5.20

We now have a single independent variable of 'income' and we could examine the effect of this on the responses to each question by five separate *t* tests. However, an alternative is to analyse the data employing a MANOVA with the five questions as five dependent variables in the analysis.

Like the ANOVA the MANOVA requires the assumptions of normally distributed populations and homogeneity of variances. However, as we have a multivariate design we also have the assumption of homogeneity of covariance, that is, the intercorrelations are similar across the conditions of the variables.

The logic of the MANOVA follows that of the ANOVA but the calculations involve matrix algebra which is beyond the scope of this book (although see the end of Chapter 23). Our data table is actually a matrix of responses. In a MANOVA we analyse the dependent variables in combination to provide a composite dependent variable to test for the effect of the independent variable. In an ANOVA we work out the sums of squares for the 'treatment' and a sums of squares for the 'error'. We then calculate the mean square (or variance) for the treatment and the mean square for the error to produce a variance ratio (or F value). In a MANOVA we still work out the sums of squares but we also work out cross-products. With one dependent variable Y, the sums of squares is $\sum(Y - \overline{Y})^2$. When there is more than one dependent variable (Y_1, Y_2, etc.) we can still work out sums of squares for each one, i.e. $\sum(Y_1 - \overline{Y}_1)^2$ for Y_1, but we can also work out the cross-products, with the cross product of Y_1 and Y_2 being $(Y_1 - \overline{Y}_1)(Y_2 - \overline{Y}_2)$ and then we work out a sums of cross-products. We have seen this type of product before in the description of the Pearson correlation coefficient. Essentially a cross-product is a measure of how much two variables covary. A matrix called the 'sums of squares and cross-products' (SSCP) is at the heart of the MANOVA just as the sums of squares is at the heart of the ANOVA. So the MANOVA analyses the covariation of the dependent variables. Thus, it is able to determine the effect of the independent variable on the composite dependent variables.

Just like in an ANOVA, where we divide the total sums of squares into the sums of squares between groups and the sums of squares within groups, the total SSCP matrix (**T**) is calculated as well as an SSCP matrix for the treatment effect between groups (**B**) and for the 'error' or within groups (**E**). Now we would like to compare these last two: **B** and **E** in the same way as we compare the sums of squares in an ANOVA. (Actually we compare the mean squares in the ANOVA rather than the sums of squares but the principle is the same.) Unfortunately, **B** and **E** are not single values but matrices. However, there is a mathematical way of finding out the variation of the values in a matrix and this is referred to as the determinant of the matrix, with the notation $|\mathbf{B}|$ for the determinant of **B**, which returns a single figure. This now allows us to work out a statistic to evaluate the significance of the effect under investigation. (I appreciate that matrix

mathematics may be a new concept but I think you can appreciate from the above description the similarity in the logic of MANOVA and ANOVA.)

A number of different statistics have been produced for MANOVA but the most commonly used is Wilks' lambda which is calculated as follows:

$$\text{Wilks' lambda } \Lambda = \frac{|\mathbf{E}|}{|\mathbf{B} + \mathbf{E}|}$$

This will range from 0 when there is no error (and all the variation is due to the treatment effect) to 1 when the variation is due to error and there is no treatment effect. So we are looking for a small value of Λ to indicate a significant effect.

In comparison to the variance ratio (F) in an ANOVA, where the F value is the treatment effect plus error divided by the error, Λ is like an upside down F ratio. Indeed, Λ can be converted to an F value quite easily and so you will usually see an F value as well as a Λ value in a computer printout for a MANOVA. In the above example, with income as the independent factor and the five questions as the five dependent variables we obtain $\Lambda = 0.0542$, $p < 0.05$ (which converts to $F(5,4) = 13.9675$). Thus we have found an effect of income on the dependent variables.

We can then undertake separate one factor independent measures ANOVAs on each question to examine the effect of income on them individually. These give the following results:

Question 1	$F(1,8) = 25.20$	$p < 0.01$
Question 2	$F(1,8) = 16.69$	$p < 0.01$
Question 3	$F(1,8) = 2.28$	$p > 0.05$
Question 4	$F(1,8) = 2.86$	$p > 0.05$
Question 5	$F(1,8) = 0.72$	$p > 0.05$

From this array we can see that income is having a significant effect on the first two questions but not the remaining three.

When we undertake a number of tests on the same data we often correct the significance level for the increased risk of Type I errors. This is called a Bonferroni correction and involves dividing the significance level by the number of tests, so with five tests, instead of choosing the $p = 0.05$ level of significance we would choose $p = 0.01$ (see Chapter 12). In this example the pattern of results of the univariate ANOVAs remains the same even with the stricter criterion for significance.

Details on how to perform a MANOVA using the SPSS computer statistical package can be found in Chapter 12 of Hinton *et al*. (2004).

Discriminant function analysis

Whereas a MANOVA examines the effect of an independent variable or variables on a number of dependent variables, a discriminant function analysis works in the opposite direction by examining which combination of independent variables is best able to predict a dependent variable. Interestingly, a discriminant function analysis is a useful follow-up analysis after a significant independent measures MANOVA as it is actually employing the same sums of squares and cross-products matrices as the MANOVA calculations. For this reason it requires the same assumptions as a MANOVA.

Essentially the discriminant function analysis produces functions of the independent variables that discriminate between the conditions of the dependent variable. To undertake this analysis on the example the independent and dependent variables are swapped round. The five questions are treated as the independent variables in this analysis and income becomes the dependent variable. Can we find functions of our five questions that are able to predict a person's income level? With only two income levels (low and high) there will only be one function produced. If we had three or more income levels then more than one function might emerge. With more than one function each will explain a certain percentage of the variation in the data and the functions (like factors in factor analysis) can be examined to see how much variation they can explain (and whether this is a significant amount). Conventionally functions are seen as worthy of further consideration if their eigenvalue is over 1 and the canonical correlation is over 0.6. A canonical correlation is essentially the correlation of the function with the dependent variable – in this case the multiple correlation coefficient (R – see Chapter 21). In the current example there is evidence of the strength of the discrimination as the eigenvalue of the function is 17.4594 and the canonical correlation is 0.9725, both high values. The significance of the function is shown by Wilks' lambda, in this case 0.0542, $p < 0.01$, so the function is highly significant in being able to discriminate the two income conditions. Notice also that this is exactly the same value of Wilks' lambda we produced in the MANOVA above, illustrating the link between the two analyses.

As we have only one function, this function is actually the multiple regression equation. The unstandardised canonical discriminant function coefficients (produced in this analysis) provide the regression coefficients, so the function for our example can be expressed as:

$$\text{Discriminant function} = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$$

$$\begin{aligned}\text{Discriminant function} &= -10.6089 + 1.4508 \times \text{Question1} - 0.0639 \times \text{Question2} \\ &\quad - 1.1093 \times \text{Question3} + 1.5500 \times \text{Question4} + 0.1721 \times \text{Question5}\end{aligned}$$

The point about this function is that when we input the values of questions 1–5 for a participant in the equation it should provide us with an outcome that we can use to classify the person into the categories of the dependent variable (i.e. predict their income level). You can see from the following table that the function is able to classify all the participants correctly: by producing a negative value for all low income participants and a positive value for all the high income participants.

Participant	*Income group*	*Function*
1	Low	−3.5545
2	Low	−4.3295
3	Low	−3.0958
4	Low	−5.2579
5	Low	−2.4488
6	High	3.5726
7	High	2.8727
8	High	2.9278
9	High	4.8929
10	High	4.4202

The mean values of the function for each group, referred to as the group centroids, provide information to make a classification. In this case the group centroids are −3.7373 and +3.7373. As we have equal numbers of participants in each group we can choose our cut-off point at the middle position between them (i.e. their average = zero). (With unequal sample sizes we would weight them by their sample size to find a weighted average position for the

cut-off point.) We can now use the function to predict the income group of a new participant once we have their results for Questions 1–5. If the function gives a negative value we classify them as 'low income' and if the function produces a positive value we classify them as 'high income'. A person who scores 7, 4, 8, 3, 5 on Questions 1–5 will score −4.0728 on the function and hence we predict them to be in the low income group.

Finally, we can examine the structure matrix (the table below) that shows the correlation of each variable with the function which, as in factor analysis, allows us to see which variables correlate highly with the function. The structure matrix has the correlation coefficients for each of the questions in order of size, with Questions 1 and 2 showing the highest values, echoing what we showed above in the MANOVA analysis.

Question	*Function*
1	0.4284
2	0.3457
4	0.1431
3	0.1279
5	0.0718

In this particular example, we saw a simple case of discriminant function analysis. With a more complex design we might find two or more functions and therefore reveal a pattern of the underlying relationship between variables responsible for a significant Wilks' lambda.

Conclusion

The advent of fast computers, available to all, has meant that even the most complicated statistical analysis can be undertaken on research data at the touch of a button. However, the crucial point is not whether an analysis can be done but whether it should be undertaken. The question for the researcher is whether they have enough understanding of the analysis to decide if it is appropriate for their data and whether they are able to correctly interpret the output of the analysis when it is produced. It may well be that a relatively simple analysis is able to properly demonstrate the key findings of a piece of research in a clear and comprehensible manner.

Chapter 23

An introduction to the general linear model

chapter 23

FOR SOME PEOPLE it a surprise to learn that the basic principles underlying the *t* test, the analysis of variance, correlation and regression, plus the multivariate tests considered in the previous chapter, are the same – they all are examples of the general linear model. The tests seem to have different aims, the calculations appear to be different, the outcomes produce different statistics, such as *t*, *F* or *r*, so that superficially they appear not the same at all. However, underlying these different tests is a model of how we expect the data to behave in order for us to perform the tests. Indeed, you may have observed that the assumptions underlying the tests are very much the same.

Now it is quite possible that you find this all very interesting but not relevant to you. Just as a person can happily drive a car without understanding the workings of the engine we can undertake statistics without knowing about the general linear model. However, if the car breaks down and you know the basics of the engine you might be able to get it going again (especially if it's a simple blockage or a lack of fuel) whereas not knowing might lead to a costly wait for the breakdown truck. Similarly, a basic understanding of the general linear model provides an awareness of what is happening in a test and whether the data are appropriate to that test. Understanding the general linear model can lead to an understanding of why we have the assumptions of the statistical tests and what it means if those assumptions are not met.

Models

In everyday conversation, when we think of a model (and not a fashion model) we often think of a small object such as a model car or a model of the Eiffel Tower. Notice that these models are representations of the thing they are modelling. Some models are very good representations, such as a detailed scale model of the Eiffel Tower, and some are not, such as the fluffy pink models of the Eiffel Tower you can buy in the souvenir shops of Paris. Yet even the poor models have to resemble the original to some extent – even the fluffy models of the Eiffel Tower have four feet and a

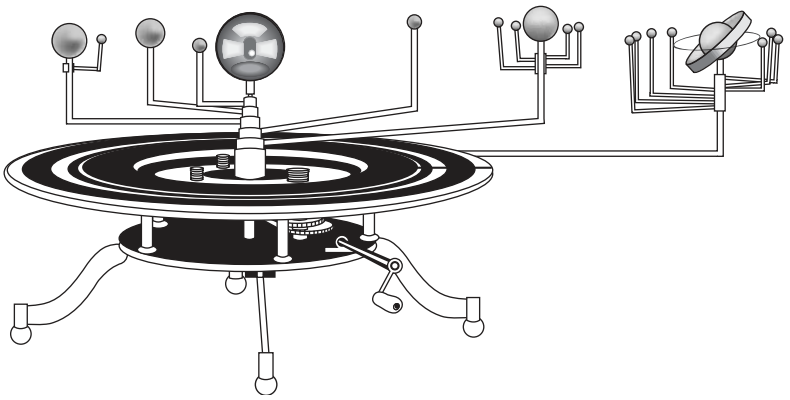

FIGURE 23.1 An orrery

pointed top. So models seek to represent the essential pattern of the thing they are modelling.

A classic example of a model is an orrery. This is a model of the solar system and you may have seen them in museums and collections of antiques. The first one was made by the clockmaker John Rowley in 1712 for Charles Boyle, the 4th Earl of Orrery (from whence it got its name). The one in Figure 23.1 is based on an orrery in the Smithsonian Institution in Washington DC.

To operate the model you turn the handle and the planets rotate around the sun. Notice that the model had the extremely useful function of being able to demonstrate in a simple manner the workings of the solar system – how the planets move relative to each other, what a year means and so on. In fact it is an extremely helpful teaching aid. However, at another level it is a very poor model. The objects are not to scale – the sun at the centre would need to be much, much bigger – and the real planets do not go round the sun in circular orbits but in ellipses. It is certainly not a model you could use to guide an astronaut in space.

Yet men have been to the moon and spacecraft have landed on other planets and the space centres have needed models of the solar system to get them there successfully. Clearly these models are enormously more complicated than the simple orrery but more importantly these models are no longer built by clockmakers in their workshops but are constructed by using mathematics. They are no longer physical objects but mathematical formulae, written down and stored on computer. If we want to estimate

where Mars and Venus will be in six months time we no longer turn the handle on the orrery and look at the new positions of the planets but input the time data into the mathematical model on the computer and print out the details of the new positions of the planets predicted by the model. If it is a good model then the positions will be accurate predictions.

Models share the same features in that they attempt to represent the relationships within a particular system (such as the movement of the planets). As soon as we decide a system is not random we can seek out a model to represent the pattern we observe. From the beginning of time people have noted the rising and setting of the sun and the change of the seasons and tried to make sense of the patterns they observe. Our current mathematical models are quite impressive as we can use them to land a spacecraft on another planet. But, who knows, in three hundred years time they may look as crude and simplistic to the people of the future as the orrery does to us today.

An example of a linear model

When we collect data we are not interested in the specific scores produced at a specific time but what the collection of scores can tell us about the relationships between variables in order to make predictions. The way we do this is by assuming that there is an underlying relationship between the variables and then we attempt to model that relationship. And, like the orrery, we can decide if the model is any good or not.

One specific type of model that is central to statistical analysis is referred to as a linear model. As we saw in Chapter 20, in its simplest case, with only the relationship between two variables, a linear model is a straight line. The mathematical formula for a straight line is $Y = a + bX$, where X and Y are the variables, 'a' is a constant (the value of Y when X is zero, the point at which the line crosses the Y axis) and 'b' is the slope of the line.

Imagine that you give a person a pack of playing cards and ask them to sort the pack as quickly as they can (but without making mistakes) into 2 piles, one of red cards and one of black cards. You shuffle the pack thoroughly and accurately measure the time it takes them to complete the task. It takes them 20.8 seconds. Now you shuffle the pack again and, this time, ask them to sort the cards into the four suits. This takes 31.2 seconds. Finally you ask them to sort the pack into 8 piles: low hearts (ace to seven), high hearts (8 to king), low diamonds, high diamonds, etc. This takes 41.6 seconds. We now plot these figures on a graph (Figure 23.2).

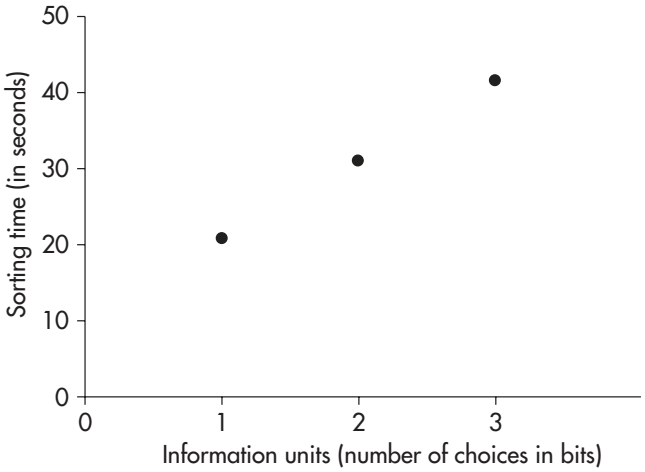

FIGURE 23.2 A graph of card sorting times

You can see that the *X*-axis is labelled 'information units' rather than 'number of piles'. A single choice (two options: e.g. on/off or red/black) contains one information unit (one 'bit' of information). Four choices involve two information units (two bits) and eight choices involve three information units (3 bits). The reason we use information units rather than number of choices is that the researchers who first did this study in the early 1950s noticed that the pattern of results when plotted on a graph in this way followed a straight line. They obviously collected considerably more data (which was much more varied) than the simple example I have given above. The resulting model, a linear relationship between amount of information and speed of processing, has immortalised the researchers who found it, and it is referred to as the Hick–Hyman Law.

For our participant, we can work out the formula for the straight line that passes through these three points, by putting the three points into the formula $Y = a + bX$ and working out 'a' and 'b' to give: $Y = 10.40 + 10.40X$, which states that:

$$\text{Sorting time} = 10.40 + (10.40 \times \text{Information units})$$

We can now use this model to predict what we do not know. If the person had to sort the pack into the high, middle and low numbers of each suit (12 choices or 3.585 bits) we would expect them to take $10.40 + (10.40 \times 3.585) = 47.68$ seconds.

Modelling data

Underlying most of our statistical techniques is the assumption that a linear model represents the pattern of the relationship between variables. Without this model we would not be able to draw the conclusions we do from our statistical analysis. Just as the space scientists need their models to land a spacecraft on Mars we need a model to make a statistical decision. In this example we shall be taking a more complex case than the three points considered in the card sorting example above, and in this new example our points will not all lie neatly along a straight line.

A researcher is interested in the relationship between a child's age and their general knowledge. We shall assume, for the sake of argument, that the researcher is able to appropriately select a suitable school and randomly selects 6 children from classes across three school years: Class 1 (roughly 8 years old), Class 2 (roughly 9 years old) and Class 3 (roughly 10 years old). Each child is given the same test of general knowledge and the scores are recorded. The results are shown in the table below.

Class	*Child's age in months*	*General knowledge score*
1	91	6
1	93	9
1	95	8
1	96	10
1	98	9
1	100	12
2	103	11
2	105	14
2	107	13
2	108	15
2	110	14
2	112	17
3	115	16
3	117	19
3	119	18
3	120	20
3	122	19
3	124	22

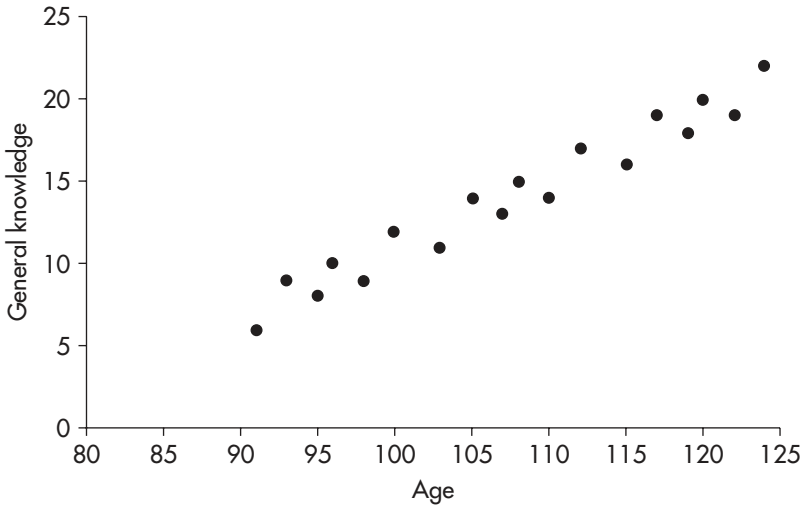

FIGURE 23.3 A plot of the children's general knowledge scores by age

Are the results random or is there a systematic relationship between age and general knowledge score? Certainly it looks from the table that the scores get larger as age increases. We can see this rather better if we plot the results, as in Figure 23.3.

I could look at the data in the graph and say that these are the results and that is that: what do we need a model for? I could claim that each point is a true representation of the child's age and score. However, this does not tell us anything we really want to know. We are not really interested in the finding that on Thursday February 21st John Peterson aged 8 years 4 months scored 12 on a general knowledge test. What we really wish to learn is whether there is an underlying relationship between age and general knowledge. If there is then we can use this relationship to make predictions about what level of general knowledge we can expect in children we have not tested. We can generalise our findings to a wider population.

When we look closely at the data it does look as though the scores more or less follow a straight line. Notice that they are all contained within a narrow band going from the bottom left to the top right of Figure 23.3 – with no scores in the top left or bottom right. So I could propose that the relationship between the general knowledge scores and age is linear (a straight line in this case) and that the underlying model for the data is a linear model. So, if the relationship between the variables really is a straight line, then that line should lie somewhere in the middle of the points, as in Figure 23.4.

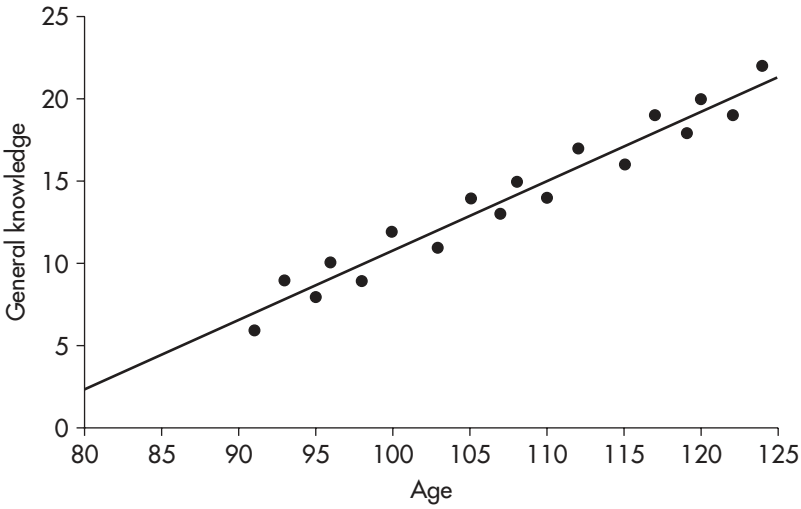

FIGURE 23.4 A proposed linear relationship between general knowledge and age

Now there is a problem here. None of the points actually lie on the line! Does this mean that this straight line is a poor model of the relationship between the general knowledge scores and age? Not necessarily. First the points seem pretty close to the line (which surely indicates that the model is not that bad). Second I could argue, the points would lie on the line had it been a perfect world but we live in a world of error and chance. Maybe one child under-performed due to having a cold and another did better than usual because they guessed an answer correctly. There are a number of factors in our everyday lives that make it messy rather than well ordered. Maybe if we took away the messiness (or random errors) then the underlying pattern would emerge (if there is one). I am suggesting that in an ideal world all the points would lie along the line. In this example, it could be that the scores have not quite fallen on the line due to these random errors that occur in any human activity, such as research, despite our best efforts at control (see Chapter 10 for a related argument).

I, therefore, argue that the underlying model of the relationship between the general knowledge scores and age is a straight line and that the reason the scores have not fallen exactly on the straight line is due to random error. Hence each observed score is made up of that predicted by the model (‘explained variation’) plus a random error (‘unexplained variation’).

Each point in Figure 23.5 shows a child’s general knowledge score. Notice that a very large proportion of each general knowledge score can be

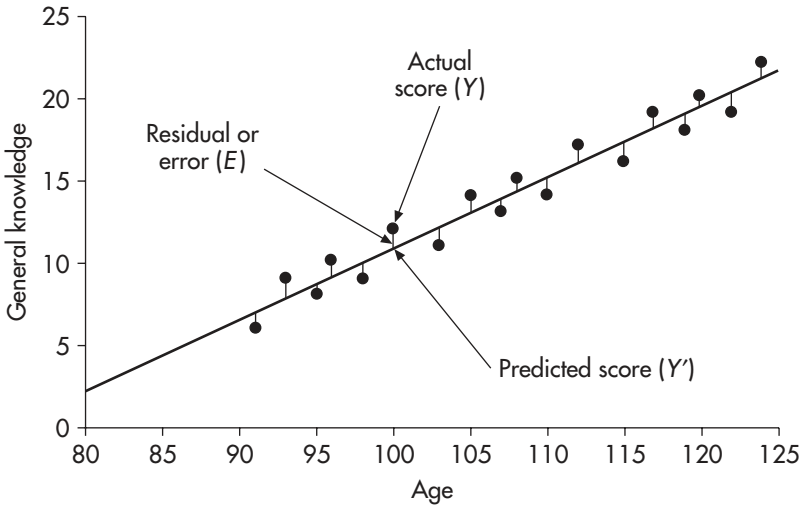

FIGURE 23.5 Separating each score into predicted score plus residual

accounted for by the model (the sloping line) as the points more or less follow the line. The 'error' scores – that is the difference between the actual score and that predicted by the model – seem quite small. The size of the error score is shown by the vertical bar joining the point to the line.

If we take one child's general knowledge score, that I will call Y, then we can explain most of that score by our model (i.e. a point on the straight line where we would predict the score would be) which we can call Y'. But, because the score does not lie on the straight line, Y is not equal to Y'. As a result I argue that E, the difference between Y and Y', is the 'error', as I believe that this is a result of random error, and cannot be explained by my model. Another term for E is residual as each of these values is the residual amount of the general knowledge score after we have taken away the amount explained by the model. So for each score:

$$\text{Actual score } (Y) = \text{Predicted score } (Y') + \text{Residual } (E)$$

The model: the regression equation

I predicted a straight line as a model for the relationship between age and general knowledge score. The problem is: which line? We can begin to work out the answer to this by looking at Figure 23.6.

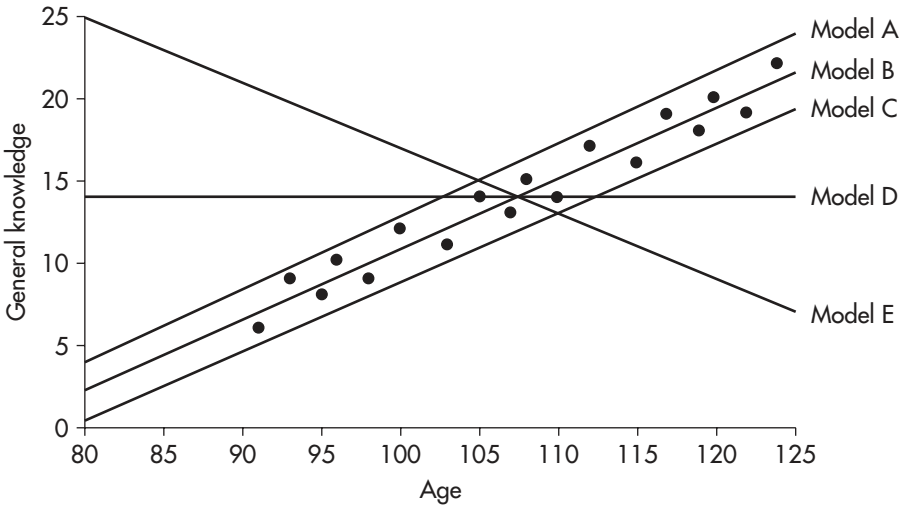

FIGURE 23.6 Different linear models

Model C is clearly not a good model as the line is way below the actual data points. We can demonstrate this mathematically as the residuals will all be positive values and their sum will be a large positive number. Similarly Model A does not fit the data very well as, again, the residuals will add to a large negative number. Adding them up will give us a large but negative sum. Model B not only looks to be the best model as it lies amongst the data points but also has smaller residual values than both Models A and C. With some of the residuals positive (the point lies above the line) and some negative (the point lies below the line), when we add up the residuals they will cancel each other out. So our best fitting model will be the straight line where the residuals add up to zero.

Another way of putting this is that the residuals have a mean of zero for our best fitting line. This makes sense as, in this case, the 'average' amount of error will be zero. Looking back to Models A and C on the above graph we can see that the residuals will not have a mean of zero as the models are not a good fit for the data, with their mean values telling us how far they are from the best model we can produce for the data. A mean of zero also indicates that the line passes through the mean values for age and general knowledge.

Unfortunately, if we now look at Models B, D and E we see all three pass through the mean values for age and general knowledge (107.5 months and a score of 14). All three models will have residuals that add up to zero

(you can work them out if you wish) and the mean of their residuals will also be zero. However, it does not require much observation to see that both Models D and E are a very poor fit to the data. The difference between Model B and Models D and E is that Model B is the model with the smallest residuals.

We now need to find the equation of the line with the smallest residual values, which add up to zero (Model B). We do this by working out the regression of age on the general knowledge scores (described in Chapter 20), which gives us the model of 'best fit' to the data. The linear regression technique is built upon the assumption of a linear model and relies on the in-built assumptions of linearity in order for it to produce its analysis. In this case it finds the linear model that minimises the size of the residuals and hence explains more variation in the data than any other linear model.

We are assuming that the observed general knowledge scores (Y) are a combination of the linear model (the regression line Y') plus the errors or residuals (E) then:

$$Y = Y' + E$$

As we know the formula for a straight line we have: $Y' = a + bX$ (where Y' is the predicted general knowledge score and X is the child's age), so:

$$Y = a + bX + E$$

This gives us a formula for E:

$$E = Y - a - bX$$

Now we can add up all the residuals:

$$\sum E = \sum (Y - a - bX)$$

This sum needs to be zero for the 'best fit' line. But we also need to find the values of 'a' and 'b' that result in the smallest residuals to get the best fitting model (Model B rather than Model D or E). There is no point simply adding up the residuals as they will cancel each other out to give a total of zero. So to find the smallest residuals we square all the residual values to get rid of the pluses and minuses and then find the line that gives us the smallest value for the sum of the squared residuals (the 'least squares method' – see Chapter 20 – that finds the minimised value for $\sum(Y - a - bX)^2$).

The outcome of this analysis gives us the following formula for the straight line that provides the best fitting straight line for the data:

$Y' = -31.77 + 0.43X$ (To be more accurate, $a = -31.7665$ and $b = 0.4257$)

This formula, our model, predicts:

General knowledge = $-31.77 + (0.43 \times \text{Age})$

We can now use this model to work out the values of the residuals by putting the age values in the equation and finding the predicted general knowledge scores. These are shown in the table below.

Child's age in months	*General knowledge score from the test*	*General knowledge score predicted by model*	*Residuals*	*Squared residuals*
91	6	6.98	−0.98	0.96
93	9	7.83	1.17	1.37
95	8	8.68	−0.68	0.46
96	10	9.10	0.90	0.81
98	9	9.96	−0.96	0.92
100	12	10.81	1.19	1.42
103	11	12.08	−1.08	1.17
105	14	12.94	1.06	1.12
107	13	13.79	−0.79	0.62
108	15	14.21	0.79	0.62
110	14	15.06	−1.06	1.12
112	17	15.92	1.08	1.17
115	16	17.19	−1.19	1.42
117	19	18.04	0.96	0.92
119	18	18.90	−0.90	0.81
120	20	19.32	0.68	0.46
122	19	20.17	−1.17	1.37
124	22	21.02	0.98	0.96
Total	252	252	0	17.71

The first point to note is that the residuals add up to zero, with some positive residuals and some negative residuals that cancel each other out when added up. Furthermore, the sum of the squared residuals (17.71) is smaller for this line than any other.

Selecting a good model

There are two qualities of a good model. The first is that the model follows the pattern of the data. If we plot the data on a graph and it follows an S-shaped curve then a straight line might not be a very good model to apply. We want to be convinced that a linear model is the appropriate model for the data. This is where the residuals come into play. The decision on what makes a good model and whether it is a good fit to the data is determined first by the characteristics of the residuals.

Second the model needs to explain as much of the data as possible. If the model can explain only 10 per cent of the variation in the scores we might not consider it as good a model as one that can explain 90 per cent of the variation. We shall be examining this second aspect later but first we consider the characteristics of the residuals.

Characteristics of the residuals

A good model is one where the error or residual values are random. If our model leaves systematic variation in the residuals then the implication is that there is a better model than the one we proposed that is able to take account of this systematic variation as well.

We want the model to explain the data equally well regardless of where we examine the data. If the model is a close fit to the data for the first few points (leaving small residuals) but then is a poor fit to subsequent points (resulting in large residuals) then it is not a good model. This is where the equality of variance assumption (or homoscedasticity) is required: the residuals should be randomly spread out at whichever point of the model we examine. Thus, we predict that the variance of the residuals at any point on the model should be the same – as there is no systematic reason why they should be larger or smaller at one point or another.

To be certain that our model is a good model and the residuals are truly random we make three further assumptions about them:

- they add to zero and have a mean of zero;
- they are from a normally distributed population; and
- they are independent of each other.

Characteristics of the residuals: they add up to zero

We have seen from the above analysis that only a model where the residuals add up to zero can provide an appropriate linear model for the data. Taking the reverse position, if the residuals do not add up to zero then we know that there is a better fitting model for the data. With the residuals summing to zero we guarantee that the model maps onto the mean values of the data.

Characteristics of the residuals: they are drawn from a normally distributed population

Given that we are assuming that the linear model underlies the data then the errors (i.e. the residuals) should be random with a normal distribution. Think about what a normal distribution means. If the errors are occurring randomly then we should occasionally get a large positive residual and occasionally we should get a large negative residual; however, most residuals should cluster round zero. So the assumption that the residuals are drawn from a normal distribution is the assumption that the residuals are indeed random (and there are no systematic patterns in the data that the model has not accounted for). If the residuals were not from a normally distributed population then the model we are proposing may be an inappropriate model for these data.

Characteristics of the residuals: they are independent of each other

If the sizes of the residuals were related to the order that the children were tested or the class they were in, then there would be a non-random element in the residuals. If the residuals got larger with increasing age of the child then the residuals would not be independent of each other. This is a concern because it demonstrates a relationship in the data not accounted for by the model.

However, if the residuals are independent of each other then there is no relationship between them and hence the ‘error’ remaining after we have

imposed the model is random, leaving no systematic variation to be explained. Thus a good model explains all the systematic variation in the data, leaving only random variation.

Conclusion

If we do not meet these assumptions then it is quite possible that the residuals are not completely random and there is still some systematic variation within them that could be accounted for by an alternative model. Indeed the common assumptions we make with our statistical tests (homogeneity of variance, etc.) arise from these assumptions concerning the residuals.

The variation in the data explained by the model

We have found, in our example, the linear model that best fits the data. No other linear model is as good as the one we have worked out. The characteristics of the residuals satisfy the assumptions. Now that we have found the best linear model we can ask a second question: how good is it? To explain what I mean, I'll rephrase the question: how much of the data is now explained by the model and how much of the data remains unexplained as error data (shown by the residuals)?

If we look at Figure 23.7 we can see that the same model (the same line) fits both sets of data (one indicated by the crosses + and the second indicated by the dots •). However, the crosses are more spread out around the line compared to the dots. We can restate this by saying that the residuals are larger for the first dataset than in the second. We can restate this again by saying that for the first dataset there is more data unexplained by the model than in the second.

This leads us on to the second judgement of a good model. A good model takes into account the variation in the data. From the data we see that as the children get older the general knowledge scores get higher. The model should predict this. There are two related methods for examining the amount of data explained by a linear model: linear correlation and the analysis of variance. Both make the assumption that the underlying model is linear, so they require the above assumptions concerning the residuals to be met.

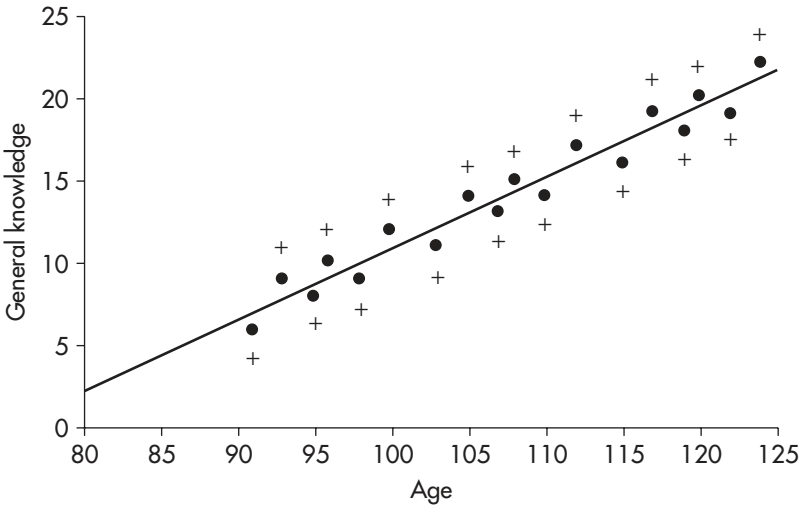

FIGURE 23.7 The same linear model for two sets of data

The linear model and correlation

We can examine whether a linear model is able to explain a lot or only a little of the variation in the data by working out the linear correlation coefficient. The technique it employs (described in Chapter 20) examines the variation in the data measured on one variable in relationship to variation on the second variable. However, it can only do that by assuming that the relationship between the two variables is linear, and then testing the strength of that relationship. It cannot detect a complex non-linear correlation – it will simply tell us that the data follows a linear relationship very badly.

In our example, the linear correlation of age and general knowledge scores is $r = 0.975$, which is an extremely high correlation ($p < 0.01$, for a two-tailed prediction, $df = 16$). Essentially, this is telling us that, assuming the relationship to be linear, the variation in the general knowledge scores can be accounted for by the variation in age to a large extent. Recall from Chapter 20 that r^2 tells us the amount of the variation in one variable explained by the other, so $r^2 = 0.951$, which means that 95.1 per cent of the variation in the general knowledge scores is explained by the variation in age.

Put simply, the linear correlation undertakes the following analysis: assuming an underlying linear relationship between the variables, how much of the variation in the data can be attributed to that relationship and how

much cannot? With 95.1 per cent of the variation in the data accounted for we can be confident that there is a linear relationship between these two variables.

Interestingly, the correlation coefficient is the slope of the 'best fit' regression line for the z scores for general knowledge and age (see Chapter 20 on z scores in the correlation calculation). A z score standardises a score so that the mean becomes zero and the standard deviation becomes 1. So instead of producing a regression for the actual scores we can produce a regression line for the z scores. This will have $a = 0$, as the line passes through (0, 0) because the means of the z scores will both be 0. It will have $b = 0.975$, as r is the slope of the line.

For the z scores:

$$z_Y = 0 + 0.975z_X$$

So:

The z score of general knowledge = $0.975 \times$ the z score of age

The good thing about this is that it shows a strong linear relationship. However, the formula is not very useful in making predictions about general knowledge scores from age, as it is couched in terms of z scores, which is why we use the standard regression equation.

The linear model and the analysis of variance

We can also provide an answer to the question about how much of the variation in the data is explained by the model by employing an analysis of variance. The analysis of variance technique is built on the assumption of a linear model. The ANOVA proportions the data into variance explained by the model and the variance that remains unexplained (the error variance). In the ANOVA we consider the variation of the scores from the mean to give a measure of the variation in the general knowledge scores. The mean general knowledge score is 14. If we take the first child's score of 6 we find that the model would predict 6.98 for this child. Thus, the model can explain $6.98 - 14 = -7.02$ of the variation of this child's score from the mean. We then square this difference (we always do this to give us a measure of the size of a difference and to get rid of the awkward minus signs at the same time). For the first child this value is 49.35. Finally we add up these squared

differences to give us a 'sums of squares' for the amount of variation in the data explained by our model. These figures are shown in the table below.

Class	*Child's age in months*	*General knowledge score*	*General knowledge score predicted by model*	*Explained variation from mean*	*Explained variation squared*	*Residuals: unexplained variation*	*Squared residuals: unexplained variation squared*
1	91	6	6.98	–7.02	49.35	–0.98	0.96
1	93	9	7.83	–6.17	38.11	1.17	1.37
1	95	8	8.68	–5.32	28.32	–0.68	0.46
1	96	10	9.10	–4.90	23.97	0.90	0.81
1	98	9	9.96	–4.04	16.36	–0.96	0.92
1	100	12	10.81	–3.19	10.20	1.19	1.42
2	103	11	12.08	–1.92	3.67	–1.08	1.17
2	105	14	12.94	–1.06	1.13	1.06	1.12
2	107	13	13.79	–0.21	0.05	–0.79	0.62
2	108	15	14.21	0.21	0.05	0.79	0.62
2	110	14	15.06	1.06	1.13	–1.06	1.12
2	112	17	15.92	1.92	3.67	1.08	1.17
3	115	16	17.19	3.19	10.20	–1.19	1.42
3	117	19	18.04	4.04	16.36	0.96	0.92
3	119	18	18.90	4.90	23.97	–0.90	0.81
3	120	20	19.32	5.32	28.32	0.68	0.46
3	122	19	20.17	6.17	38.11	–1.17	1.37
3	124	22	21.02	7.02	49.35	0.98	0.96
Total		252	252	0	342.29	0	17.71

(I have given the accurate total for the sixth column. If you added up the figures to only two decimal places you would get a figure of 342.32 due to rounding errors.)

Now we have both the 'sums of squares' for the general knowledge scores explained by age (342.29) plus the 'sums of squares' for the error term (the sum of the squared residuals: 17.71). Thus, of the total variation in the data (total sums of squares = 360.00) we can explain 342.29 of it by the linear model, leaving 17.71 unexplained. It is a simple matter to complete an ANOVA summary table – we just need to supply the degrees of freedom to finalise the calculations.

THE ANOVA SUMMARY TABLE

Source of variation	Degrees of freedom	Sums of squares	Mean square	*F*	Significance
Model (linear regression)	1	342.29	342.29	309.25	$p < 0.01$
Residual (error)	16	17.71	1.11		
Total	17	360.00			

The results of the analysis of variance tell us that the model can explain a highly significant amount of the variation in the data.

We have employed a regression, correlation and analysis of variance on our data. Each of these analyses assumes that there is a linear relationship between the two variables we have measured. In the example of age and general knowledge all three statistical techniques have supported a linear relationship between the two variables. A linear model is a good fit to the data and it can explain a considerable amount of the variation in the scores.

Comparing samples (the analysis of variance once again)

It is relatively easy to see the underlying assumption of a linear model in a linear regression and linear correlation. However, it is not always so clear that this assumption is also inherent in the analysis when we are comparing samples (e.g. in a *t* test or ANOVA). We can illustrate this assumption by once again looking at the general knowledge and age data. We can use a one factor independent measures ANOVA to compare the general knowledge scores for the different Classes (Class 1, Class 2 and Class 3). By placing them in the category of Class rather than taking their age we are placing all of the 6 children in each class at the same position on the *X*-axis. But the same logic that we employed above when looking at age still applies. We can see a plot of the data in Figure 23.8.

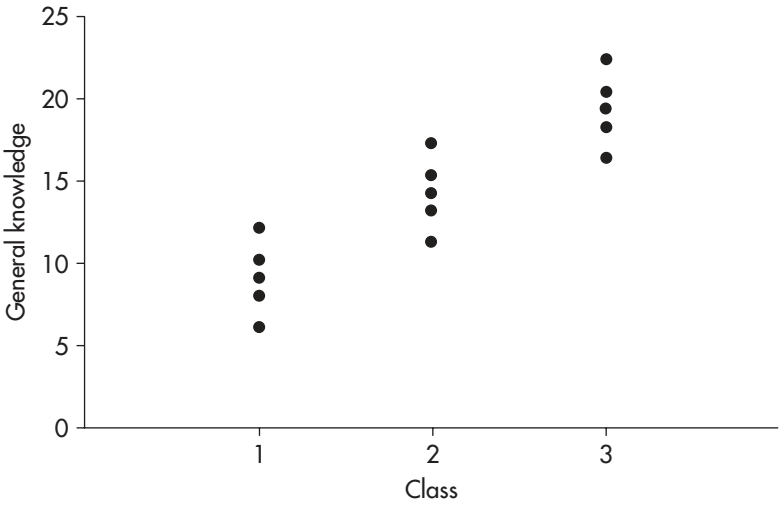

FIGURE 23.8 Plot of general knowledge scores for each class

Now we can do exactly as we did before with the scatterplot of general knowledge scores and age. Is there a linear model that underlies Figure 23.8? Although we do not normally think of undertaking a correlation with category data like this, computer statistical programmes will often print out the correlation coefficient r, or r^2, with the ANOVA summary table.

Correlating the general knowledge scores with Class produces a high linear relationship between the two variables ($r = 0.913$, $p < 0.01$ for a two-tailed prediction, $df = 16$) with $r^2 = 0.833$, indicating that the variable Class can explain 83.3 per cent of the variation in the general knowledge scores. Even though we are comparing the categories Class 1, Class 2, and Class 3, we are still examining the fit of a linear model.

We can find the best linear model to fit these data by performing a regression analysis. The result of this gives us the following formula:

$$Y = 4 + 5X$$

So our 'best fit' linear model predicts:

$$\text{General knowledge} = 4 + 5 \times \text{Class}$$

This model is shown in Figure 23.9.

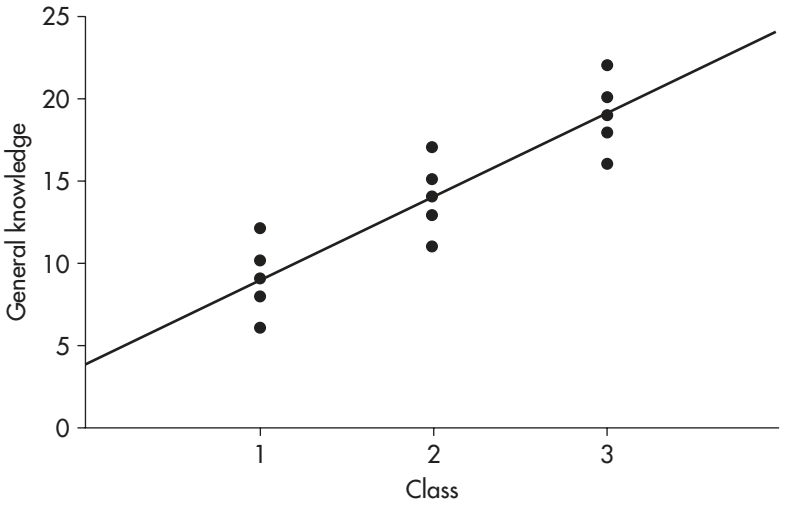

FIGURE 23.9 A linear model for the class data

Interestingly you can see, from Figure 23.9, why we have the equality of variance assumption with comparisons and the homoscedasticity assumption with correlations. We are assuming that the data is evenly spread around the regression line in both cases.

Now that we have our model we can work out the general knowledge scores as predicted by the model and the 'error' scores or residuals. Notice, from the third column in the following table, that the scores predicted by the model are the category means, so the predicted score for all the children in Class 1 is the mean of Class 1 (a score of 9). The residuals are shown in the sixth column. Just as we did in the previous analysis of variance, we work out the variation from the mean for each data point predicted by the model (as this is the variation the model is able to explain). We square these values to produce a measure of explained variation. If we add these up we get a 'sums of squares' for the explained variation. We also square the residuals to get a value for the size of the unexplained variation. Adding these up gives us the 'sums of squares' for the error variation. These are also listed in the table below.

Class	*General knowledge score*	*General knowledge score predicted by model*	*Explained variation from mean*	*Explained variation squared*	*Residuals: unexplained variation*	*Squared residuals: unexplained variation squared*
1	6	9	–5	25	–3	9
1	9	9	–5	25	0	0
1	8	9	–5	25	–1	1
1	10	9	–5	25	1	1
1	9	9	–5	25	0	0
1	12	9	–5	25	3	9
2	11	14	0	0	–3	9
2	14	14	0	0	0	0
2	13	14	0	0	–1	1
2	15	14	0	0	1	1
2	14	14	0	0	0	0
2	17	14	0	0	3	9
3	16	19	5	25	–3	9
3	19	19	5	25	0	0
3	18	19	5	25	–1	1
3	20	19	5	25	1	1
3	19	19	5	25	0	0
3	22	19	5	25	3	9
Total	252	252	0	300	0	60

We now have all the information to draw up the analysis of variance summary table:

THE ANOVA SUMMARY TABLE

Source of variation	Degrees of freedom	Sums of squares	Mean square	*F*	Significance
Model (linear regression)	2	300.00	150.00	37.50	$p < 0.01$
Residual (error)	15	60.00	4.00		
Total	17	360.00			

Analysing the data by Class rather than age divides the total sums of squares (360.00) into that explained by the model (300.00) and the remainder not explained by the model (60.00). It is clear that our underlying model can account for a significantly large proportion of the variation in the data ($p < 0.01$). Hence we can reject the null hypothesis that the means are drawn from the same population distribution.

Explaining variations in the data

You may have wondered why in both of the above analyses of variance we worked out the explained variation relative to the mean value. The answer arises from the way we calculate variation in the data. If we simply square the general knowledge scores and add them up we get a total of 3888. This value would only be a measure of the total variability of the scores in the data *if* the mean equals zero. Consider two scores 99 and 101. These scores vary by 2, with 99 one below their mean of 100, and 101 one above their mean of 100. Now consider two other scores 25 and 35. These vary by 10 with 25 five below their mean of 30, and 35 five above their mean of 30. It is obvious that there is greater variation in the second two scores compared to the first, despite the fact that 25 and 35 are smaller than 99 and 101. So when considering the variation of scores in our data we are not interested in their actual values but the amount of variation between them so that is why we compare them to the mean. Sometimes you will see, in the output of statistical computer programs, the sum of the squared scores referred to as the 'total' and the sum of the squared scores-minus-the-mean as the 'corrected total' as it is the second of these two sums that gives us the correct measurement of the total variability in the data. In our example the corrected total is 360 (the value we have used in the above calculations for the total variability in the scores). The difference between the total and the corrected total, $3888 - 360 = 3528$, is simply an indication of how far the mean value differs from zero.

We now have undertaken two analyses of variance on the general knowledge data. The first looked at the relationship between the general knowledge scores and age and the second compared the general knowledge scores across the three classes. In both cases the analyses were only possible because we had postulated an underlying linear model for the data. The models that best fitted the data were different in the two cases as the first analysis included the age information whereas the second included only the class information. However, given that the analysis was undertaken on the

same general knowledge scores it is no surprise to see that the total variation in the data (the 'sums of squares') added up to 360 in both cases. With the assumption of a linear model we were able to separate this into the 'variability explained by the model' and the unexplained variability in the data. In our first analysis the 'explained sums of squares' was 342.29 and the 'unexplained sums of squares' was 17.71. In the second analysis these figures were 300 and 60 respectively.

The general linear model

Up to now we have deal with the simplest case of a linear model, that is, a straight line relationship between two variables, shown by the formula $Y = a + bX$. However, this is the simplest case of a much more general model that can include not just one independent or X variable but many independent variables (indeed we have seen two independent variables in the two factor ANOVAs in Chapter 15). Furthermore, it also allows for multiple Y or dependent variables. To illustrate this we need first to display our model in terms of matrix representation.

Our two variable example using matrix representation

Consider once again the general knowledge and age data. To find our linear model we minimised the error for $Y = a + bX + E$, where Y is the test score and X the age and E the error or residual. So for our children we can put each of their scores into the formula, one at a time:

$$
\begin{array}{lcl}
Y_1 = a + bX_1 + E_1 & \blacktriangleright \text{ or with the} \blacktriangleright & 6 = a + b \times 91 + E_1 \\
Y_2 = a + bX_2 + E_2 & \text{values of general} & 9 = a + b \times 93 + E_2 \\
Y_3 = a + bX_3 + E_3 & \text{knowledge and} & 8 = a + b \times 95 + E_3 \\
\vdots & \text{age inserted} & \vdots \\
Y_{16} = a + bX_{16} + E_{16} & & 20 = a + b \times 120 + E_{16} \\
Y_{17} = a + bX_{17} + E_{17} & & 19 = a + b \times 122 + E_{17} \\
Y_{18} = a + bX_{18} + E_{18} & & 22 = a + b \times 124 + E_{18}
\end{array}
$$

(I have used the small dots to indicate that the rest of the values need to be included here, otherwise I would have had to list out the formulae for all eighteen children.)

We can represent this in matrix terms as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_{16} \\ Y_{17} \\ Y_{18} \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \vdots & \vdots \\ 1 & X_{16} \\ 1 & X_{17} \\ 1 & X_{18} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \vdots \\ E_{16} \\ E_{17} \\ E_{18} \end{bmatrix} \blacktriangleright \text{ or with the values of general knowledge and age inserted } \blacktriangleright \begin{bmatrix} 6 \\ 9 \\ 8 \\ \vdots \\ 20 \\ 19 \\ 22 \end{bmatrix} = \begin{bmatrix} 1 & 91 \\ 1 & 93 \\ 1 & 95 \\ \vdots & \vdots \\ 1 & 120 \\ 1 & 122 \\ 1 & 124 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \vdots \\ E_{16} \\ E_{17} \\ E_{18} \end{bmatrix}$$

Using large bold characters to represent matrices rather than the smaller letters we have been using to represent individual values we can replace the matrices as follows:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_{16} \\ Y_{17} \\ Y_{18} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \vdots & \vdots \\ 1 & X_{16} \\ 1 & X_{17} \\ 1 & X_{18} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} a \\ b \end{bmatrix}, \mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \vdots \\ E_{16} \\ E_{17} \\ E_{18} \end{bmatrix}$$

(We have a column of 1s in the **X** matrix to represent the intercept, '*a*', in our model. If we did not put in this column then our line would be forced to go through zero.)

So, in matrix terms:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

Employing matrix algebra (which is a little too complicated for this book) we can find the values of '*a*' and '*b*' that minimise the error quite easily as:

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

where $\mathbf{X}'$ is the transpose of **X** and is worked out by swapping the rows and columns of **X**, and the inverse matrix (a matrix raised to the power of −1) can be calculated by a mathematical formula.

I hope you will appreciate that we are able work out the appropriate values of '*a*' and '*b*' using matrix algebra (which I am not expecting you to know about) using the information we already have. And it turns out that:

$$\mathbf{B} = \begin{bmatrix} \overline{Y} - b\overline{X} \\ \\ \dfrac{SP}{SS_x} \end{bmatrix}$$

(You can see from Chapter 20 on regression that these are the formulae for '*a*' and '*b*'.)

Hopefully, even for readers not familiar with matrix algebra, it is clear that all we have done is represent the same model but in a different way.

Multiple *X* variables

We can extend the linear model by looking at more than two variables. If you refer back to Chapter 21 on multiple regression, we see that the formula we employ is of the form:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$$

This is still a linear model as it still contains the intercept '*a*' plus the slope '*b*' but here we have a *b*-value for each of the *X* variables. Essentially the linear model means that there are no squared or higher values of *X* in the formula. As we are no longer working with only two variables the linear model is no longer a straight line but a multidimensional space. However, we can use the same logic to examine as many independent or *X* variables as we wish in our analysis and perform multiple correlation and regression operations as well as performing multifactorial analyses of variance, such as a two factor analysis of variance.

This is still a linear model of the form:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

In this case the **X** matrix is now

$$\begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}$$

where *n* is the number of participants and *k* is the number of independent variables so, for example, X_{12} is the value of participant 2 on the first independent variable.

The general linear model and multivariate analysis

We can generalise the linear model further to allow more than one *Y* variable as well as more than one *X* variable. We refer to analysis involving multiple dependent variables as multivariate analysis as compared to the single *Y* variable or univariate analysis (as we saw in Chapter 22). But the matrix notation does not change. We still have:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

But in this case the **Y** matrix is now

$$\begin{bmatrix} Y_{11} & Y_{21} & \cdots & Y_{m1} \\ Y_{12} & Y_{22} & \cdots & Y_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{1n} & Y_{2n} & \cdots & Y_{mn} \end{bmatrix}$$

where m is the number of dependent variables and Y_{21} is the score of the first participant on the second dependent variable.

Thus, the model is now referred to as the General Linear Model as it can include multiple independent and multiple dependent variables. You may have noticed that in Chapter 22 on multivariate analyses there is mention of matrices at various points in the discussions of complex analyses (such as factor analysis and MANOVA).

The important point here is that, regardless of whether we are dealing with one independent and one dependent variable or many of them, we can map our data onto a linear model and, as long as we satisfy the assumptions of the model, we have a powerful tool for making sense of research findings. Just as scientists use their models of the solar system to predict the movements of the planets, we can use a linear model to predict the relationships between our variables. We may be excited by the prospect of exploring other planets but we need to get there safely first and we can only do that with a good model. Similarly we may wish to discover exciting relationships between variables in our own field of study and it is well worth appreciating the role of the general linear model in the processes of quantitative data analysis and how it helps us to reach conclusions to our studies.

I hope this brief account of the underlying linear model in statistical analysis has given you some insight into the construction and application of

many statistical tests. In particular an awareness of the importance of residuals is crucial to understanding the assumptions required for these tests. Unfortunately, further explanation requires a deeper foray into matrix algebra, which is beyond the scope of this book.