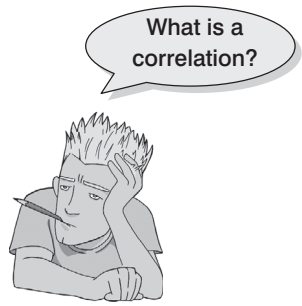
6 Correlation

FIGURE 6.1 I don't have a photo from Christmas 1981, but this was taken about that time at my grandparents' house. I'm trying to play an E by the looks of it, no doubt because it's in 'Take on the world'

6.1. What will this chapter tell me? ①

When I was 8 years old, my parents bought me a guitar for Christmas. Even then, I'd desperately wanted to play the guitar for years. I could not contain my excitement at getting this gift (had it been an *electric* guitar I think I would have actually exploded with excitement). The guitar came with a 'learn to play' book and after a little while of trying to play what was on page 1 of this book, I readied myself to unleash a riff of universe-crushing power onto the world (well, 'skip to my Lou' actually). But, I couldn't do it. I burst into tears and ran upstairs to hide.[1] My dad sat with me and said 'Don't worry, Andy, everything is hard to begin with, but the more you practise the easier it gets.'

[1] This is not a dissimilar reaction to the one I have when publishers ask me for new editions of statistics textbooks.



In his comforting words, my dad was inadvertently teaching me about the relationship, or correlation, between two variables. These two variables could be related in three ways: (1) *positively related*, meaning that the more I practised my guitar, the better a guitar player I would become (i.e. my dad was telling me the truth); (2) *not related* at all, meaning that as I practise the guitar my playing ability remains completely constant (i.e. my dad has fathered a cretin); or (3) *negatively related*, which would mean that the more I practised my guitar the worse a guitar player I became (i.e. my dad has fathered an indescribably strange child). This chapter looks first at how we can express the relationships between variables statistically by looking at two measures: *covariance* and the *correlation coefficient*. We then discover how to carry out and interpret correlations in SAS. The chapter ends by looking at more complex measures of relationships; in doing so it acts as a precursor to the chapter on multiple regression.

6.2. Looking at relationships ①

In Chapter 4 I stressed the importance of looking at your data graphically before running any other analysis on them. I just want to begin by reminding you that our first starting point with a correlation analysis should be to look at some scatterplots of the variables we have measured. I am not going to repeat how to get SAS to produce these graphs, but I am going to urge you (if you haven't done so already) to read section 4.7 before embarking on the rest of this chapter.

6.3. How do we measure relationships? ①

6.3.1. A detour into the murky world of covariance ①

The simplest way to look at whether two variables are associated is to look at whether they *covary*. To understand what **covariance** is, we first need to think back to the concept of variance that we met in Chapter 2. Remember that the variance of a single variable represents the average amount that the data vary from the mean. Numerically, it is described by:

$$\text{variance}(s^2) = \frac{\sum(x_i - \bar{x})^2}{N-1} = \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{N-1} \tag{6.1}$$

The mean of the sample is represented by $\bar{x}$, x_i is the data point in question and N is the number of observations (see section 2.4.1). If we are interested in whether two variables are related, then we are interested in whether changes in one variable are met with similar changes in the other variable. Therefore, when one variable deviates from its mean we would expect the other variable to deviate from its mean in a similar way. To illustrate what I mean, imagine we took five people and subjected them to a certain number of advertisements promoting toffee sweets, and then measured how many packets of those sweets each person bought during the next week. The data are in Table 6.1 as well as the mean and standard deviation (s) of each variable.

TABLE 6.1

Subject:	1	2	3	4	5	Mean	s
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92

If there were a relationship between these two variables, then as one variable deviates from its mean, the other variable should deviate from its mean in the same or the directly opposite way. Figure 6.2 shows the data for each participant (hollow circles represent the number of packets bought and solid circles represent the number of adverts watched); the dashed line is the average number of packets bought and the solid line is the average number of adverts watched. The vertical lines represent the differences (remember that these differences are called *deviations*) between the observed values and the mean of the relevant variable. The first thing to notice about Figure 6.2 is that there is a very similar pattern of deviations for both variables. For the first three participants the observed values are below the mean for both variables, for the last two people the observed values are above the mean for both variables. This pattern is indicative of a potential relationship between the two variables (because it seems that if a person's score is below the mean for one variable then their score for the other will also be below the mean).

So, how do we calculate the exact similarity between the pattern of differences of the two variables displayed in Figure 6.2? One possibility is to calculate the total amount of deviation, but we would have the same problem as in the single-variable case: the positive and negative deviations would cancel out (see section 2.4.1). Also, by simply adding the deviations, we would gain little insight into the *relationship* between the variables. Now, in the single-variable case, we squared the deviations to eliminate the problem of positive and negative deviations cancelling out each other. When there are two variables, rather than squaring each deviation, we can multiply the deviation for one variable by the corresponding deviation for the second variable. If both deviations are positive or negative then this will give us a positive value (indicative of the deviations being in the same direction), but if one deviation is positive and one negative then the resulting product will be negative

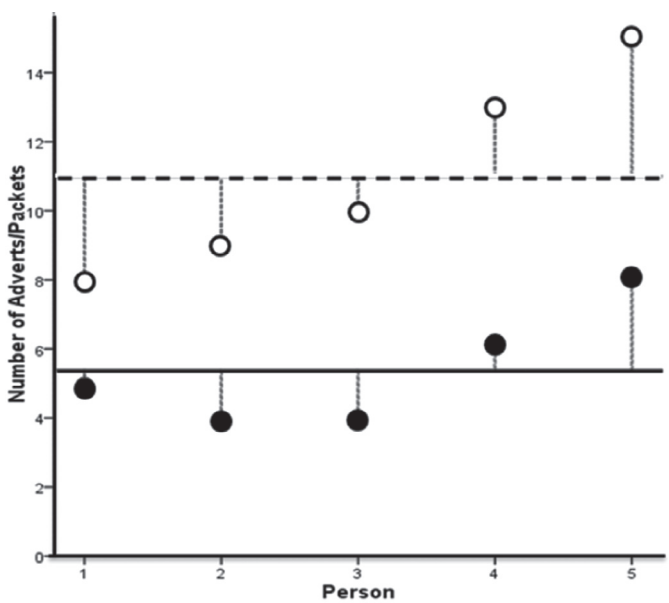


FIGURE 6.2
Graphical display of the differences between the observed data and the means of two variables

(indicative of the deviations being opposite in direction). When we multiply the deviations of one variable by the corresponding deviations of a second variable, we get what is known as the **cross-product deviations**. As with the variance, if we want an average value of the combined deviations for the two variables, we must divide by the number of observations (we actually divide by $N - 1$ for reasons explained in Jane Superbrain Box 2.2). This averaged sum of combined deviations is known as the *covariance*. We can write the covariance in equation form as in equation (6.2) – you will notice that the equation is the same as the equation for variance, except that instead of squaring the differences, we multiply them by the corresponding difference of the second variable:

$$\text{cov}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1} \tag{6.2}$$

For the data in Table 6.1 and Figure 6.2 we reach the following value:

$$\begin{aligned}\text{cov}(x, y) &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1} \\ &= \frac{(-0.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (0.6)(2) + (2.6)(4)}{4} \\ &= \frac{1.2 + 2.8 + 1.4 + 1.2 + 10.4}{4} \\ &= \frac{17}{4} \\ &= 4.25\end{aligned}$$

Calculating the covariance is a good way to assess whether two variables are related to each other. A positive covariance indicates that as one variable deviates from the mean, the other variable deviates in the same direction. On the other hand, a negative covariance indicates that as one variable deviates from the mean (e.g. increases), the other deviates from the mean in the opposite direction (e.g. decreases).

There is, however, one problem with covariance as a measure of the relationship between variables and that is that it depends upon the scales of measurement used. So, covariance is not a standardized measure. For example, if we use the data above and assume that they represented two variables measured in miles then the covariance is 4.25 (as calculated above). If we then convert these data into kilometres (by multiplying all values by 1.609) and calculate the covariance again then we should find that it increases to 11. This dependence on the scale of measurement is a problem because it means that we cannot compare covariances in an objective way – so, we cannot say whether a covariance is particularly large or small relative to another data set unless both data sets were measured in the same units.

6.3.2. Standardization and the correlation coefficient ①

To overcome the problem of dependence on the measurement scale, we need to convert the covariance into a standard set of units. This process is known as **standardization**. A very basic form of standardization would be to insist that all experiments use the same units of measurement, say metres – that way, all results could be easily compared.

However, what happens if you want to measure attitudes – you'd be hard pushed to measure them in metres! Therefore, we need a unit of measurement into which any scale of measurement can be converted. The unit of measurement we use is the *standard deviation*. We came across this measure in section 2.4.1 and saw that, like the variance, it is a measure of the average deviation from the mean. If we divide any distance from the mean by the standard deviation, it gives us that distance in standard deviation units. For example, for the data in Table 6.1, the standard deviation for the number of packets bought is approximately 3.0 (the exact value is 2.92). In Figure 6.2 we can see that the observed value for participant 1 was 3 packets less than the mean (so there was an error of –3 packets of sweets). If we divide this deviation, –3, by the standard deviation, which is approximately 3, then we get a value of –1. This tells us that the difference between participant 1's score and the mean was –1 standard deviation. So, we can express the deviation from the mean for a participant in standard units by dividing the observed deviation by the standard deviation.

It follows from this logic that if we want to express the covariance in a standard unit of measurement we can simply divide by the standard deviation. However, there are two variables and, hence, two standard deviations. Now, when we calculate the covariance we actually calculate two deviations (one for each variable) and then multiply them. Therefore, we do the same for the standard deviations: we multiply them and divide by the product of this multiplication. The standardized covariance is known as a *correlation coefficient* and is defined by equation (6.3) in which s_x is the standard deviation of the first variable and s_y is the standard deviation of the second variable (all other letters are the same as in the equation defining covariance):

$$r = \frac{\text{cov}_{xy}}{s_x s_y} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{(N-1)s_x s_y} \tag{6.3}$$

The coefficient in equation (6.3) is known as the *Pearson product-moment correlation coefficient* or **Pearson correlation coefficient** (for a really nice explanation of why it was originally called the 'product-moment' correlation see Miles & Banyard, 2007) and was invented by Karl Pearson (see Jane Superbrain Box 6.1).[2] If we look back at Table 6.1 we see that the standard deviation for the number of adverts watched (s_x) was 1.67, and for the number of packets of crisps bought (s_y) was 2.92. If we multiply these together we get $1.67 \times 2.92 = 4.88$. Now, all we need to do is take the covariance, which we calculated a few pages ago as being 4.25, and divide by these multiplied standard deviations. This gives us $r = 4.25/4.88 = .87$.

By standardizing the covariance we end up with a value that has to lie between –1 and +1 (if you find a correlation coefficient less than –1 or more than +1 you can be sure that something has gone hideously wrong!). A coefficient of +1 indicates that the two variables are perfectly positively correlated, so as one variable increases, the other increases by a proportionate amount. Conversely, a coefficient of –1 indicates a perfect negative relationship: if one variable increases, the other decreases by a proportionate amount. A coefficient of zero indicates no linear relationship at all. We also saw in section 2.6.4 that because the correlation coefficient is a standardized measure of an observed effect, it is a commonly used measure of the size of an effect and that values of ±.1 represent a small

[2] You will find Pearson's product-moment correlation coefficient denoted by both *r* and *R*. Typically, the uppercase form is used in the context of regression because it represents the multiple correlation coefficient; however, for some reason, when we square *r* (as in section 6.5.2.3) an upper case *R* is used. Don't ask me why – it's just to confuse me, I suspect.

JANE SUPERBRAIN 6.1

Who said statistics was dull? ①

Students often think that statistics is dull, but back in the early 1900s it was anything but dull with various prominent figures entering into feuds on a soap opera scale. One of the most famous was between Karl Pearson and Ronald Fisher (whom we met in Chapter 2). It began when Pearson published a paper of Fisher's in his journal but made comments in his editorial that, to the casual reader, belittled Fisher's work. Two years later Pearson's group published work following on from Fisher's paper without consulting him. The antagonism persisted, with Fisher turning down a job to work in Pearson's, group and publishing 'improvements' on Pearson's ideas. Pearson for his part wrote in his own journal about apparent errors made by Fisher.

Another prominent statistician, Jerzy Neyman, criticized some of Fisher's most important work in a paper delivered to the Royal Statistical Society on 28 March 1935 at which Fisher was present. Fisher's discussion of the paper at that meeting directly attacked Neyman. Fisher more or less said that Neyman didn't know what he was talking about and didn't understand the background material on which his work was based. Relations soured so much that while they both worked at University College London, Neyman openly attacked many of Fisher's ideas in lectures to his students. The two feuding groups even took afternoon tea (a common practice in the British academic community of the time) in the same room but at different times! The truth behind who fuelled these feuds is, perhaps, lost in the mists of time, but Zabell (1992) makes a sterling effort to unearth it.

Basically, then, the founders of modern statistical methods were a bunch of squabbling children. Nevertheless, these three men were astonishingly gifted individuals. Fisher, in particular, was a world leader in genetics, biology and medicine as well as possibly the most original mathematical thinker ever (Barnard, 1963; Field, 2005d; Savage, 1976).

effect, ±.3 a medium effect and ±.5 a large effect (although I re-emphasize my caveat that these canned effect sizes are no substitute for interpreting the effect size within the context of the research literature).

6.3.3. The significance of the correlation coefficient ③

Although we can directly interpret the size of a correlation coefficient, we have seen in Chapter 2 that scientists like to test hypotheses using probabilities. In the case of a correlation coefficient we can test the hypothesis that the correlation is different from zero (i.e. different from 'no relationship'). If we find that our observed coefficient was very unlikely to happen if there was no effect in the population then we can gain confidence that the relationship that we have observed is statistically meaningful.

There are two ways that we can go about testing this hypothesis. The first is to use our trusty z-scores that keep cropping up in this book. As we have seen, z-scores are useful because we know the probability of a given value of z occurring, if the distribution from which it comes is normal. There is one problem with Pearson's r, which is that it is known to have a sampling distribution that is not normally distributed. This is a bit

of a nuisance, but luckily thanks to our friend Fisher we can adjust r so that its sampling distribution *is* normal as follows (Fisher, 1921):

$$z_r = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) \tag{6.4}$$

The resulting z_r has a standard error of:

$$\mathrm{SE}_{z_r} = \frac{1}{\sqrt{N-3}} \tag{6.5}$$

For our advert example, our $r = .87$ becomes 1.33 with a standard error of 0.71.

We can then transform this adjusted r into a z-score just as we have done for raw scores, and for skewness and kurtosis values in previous chapters. If we want a z-score that represents the size of the correlation relative to a particular value, then we simply compute a z-score using the value that we want to test against and the standard error. Normally we want to see whether the correlation is different from 0, in which case we can subtract 0 from the observed value of r and divide by the standard error (in other words, we just divide z_r by its standard error):

$$z = \frac{z_r}{\mathrm{SE}z_r} \tag{6.6}$$

For our advert data this gives us 1.33/0.71 = 1.87. We can look up this value of z (1.87) in the table for the normal distribution in the Appendix and get the one-tailed probability from the column labelled 'Smaller Portion'. In this case the value is .0307. To get the two-tailed probability we simply multiply the one-tailed probability value by 2, which gives us .0614. As such the correlation is significant, $p < .05$ one-tailed, but not two-tailed.

In fact, the hypothesis that the correlation coefficient is different from 0 is usually (SAS, for example, does this) tested not using a z-score, but using a t-statistic with $N - 2$ degrees of freedom, which can be directly obtained from r:

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{6.7}$$

So, you might wonder then why I told you about z-scores. Partly it was to keep the discussion framed in concepts with which you are already familiar (we don't encounter the t-test properly for a few chapters), but also it is useful background information for the next section.

6.3.4. Confidence intervals for r ③

I was moaning on earlier about how SAS doesn't make tea for you. Another thing that it doesn't do is compute confidence intervals for r. This is a shame because as we have seen in Chapter 2 these intervals tell us something about the likely value (in this case of the correlation) in the population. However, we can calculate these manually (or if you search the web you will also find SAS macros that will do this). To do this we need to take advantage of what we learnt in the previous section about converting r to z_r (to make

the sampling distribution normal), and using the associated standard errors. We can then construct a confidence interval in the usual way. For a 95% confidence interval we have (see section 2.5.2.1):

$$\text{lower boundary of confidence interval} = \overline{X} - (1.96 \times \text{SE})$$
$$\text{upper boundary of confidence interval} = \overline{X} + (1.96 \times \text{SE})$$

In the case of our transformed correlation coefficients these equations become:

$$\text{lower boundary of confidence interval} = z_r - (1.96 \times \text{SE}_{z_r})$$
$$\text{upper boundary of confidence interval} = z_r + (1.96 \times \text{SE}_{z_r})$$

For our advert data this gives us $1.33 - (1.96 \times 0.71) = -0.062$, and $1.33 + (1.96 \times 0.71) = 2.72$. Remember that these values are in the z_r metric and so we have to convert back to correlation coefficients using:

$$r = \frac{e^{(2z_r)} - 1}{e^{(2z_r)} + 1} \tag{6.8}$$

This gives us an upper bound of $r = .991$ and a lower bound of $-.062$ (because this value is so close to zero the transformation to z has no impact).

CRAMMING SAM'S TIPS

- A crude measure of the relationship between variables is the *covariance*.
- If we standardize this value we get *Pearson's correlation coefficient, r.*
- The correlation coefficient has to lie between –1 and +1.
- A coefficient of +1 indicates a perfect positive relationship, a coefficient of –1 indicates a perfect negative relationship, a coefficient of 0 indicates no linear relationship at all.
- The correlation coefficient is a commonly used measure of the size of an effect: values of ± .1 represent a small effect, ± .3 is a medium effect and ± .5 is a large effect.

6.3.5. A word of warning about interpretation: causality ①

Considerable caution must be taken when interpreting correlation coefficients because they give no indication of the direction of *causality.* So, in our example, although we can

conclude that as the number of adverts watched increases, the number of packets of toffees bought increases also, we cannot say that watching adverts *causes* you to buy packets of toffees. This caution is for two reasons:

- **The third-variable problem**: We came across this problem in section 1.6.2. To recap, in any correlation, causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results. This is known as the *third-variable* problem or the *tertium quid* (see section 1.6.2 and Jane Superbrain Box 1.1).
- **Direction of causality**: Correlation coefficients say nothing about which variable causes the other to change. Even if we could ignore the third-variable problem described above, and we could assume that the two correlated variables were the only important ones, the correlation coefficient doesn't indicate in which direction causality operates. So, although it is intuitively appealing to conclude that watching adverts causes us to buy packets of toffees, there is no *statistical* reason why buying packets of toffees cannot cause us to watch more adverts. Although the latter conclusion makes less intuitive sense, the correlation coefficient does not tell us that it isn't true.

6.4. Data entry for correlation analysis using SAS ①

Data entry for correlation, regression and multiple regression is straightforward because each variable is entered in a separate column. So, for each variable you have measured, create a variable in the data editor with an appropriate name, and enter a participant's scores across one row of the data editor. There may be occasions on which you have one or more categorical variables (such as gender). As an example, if we wanted to calculate the correlation between the two variables in Table 6.1 we would enter each variable in a separate column, and each row represents a single individual's data (so the first consumer saw 5 adverts and bought 8 packets).

SELF-TEST Enter the advert data (you might want to look at a DATA step, in SAS Syntax 3-1 to remind yourself how to do this) and use the gplot to produce a scatterplot (number of packets bought on the *y*-axis, and adverts watched on the *x*-axis) of the data.

6.5. Bivariate correlation ①

6.5.1. General procedure for running correlations on SAS ①

There are two types of correlation: *bivariate* and *partial*. A **bivariate correlation** is a correlation between two variables (as described at the beginning of this chapter) whereas a

partial correlation looks at the relationship between two variables while 'controlling' the effect of one or more additional variables. Pearson's product-moment correlation coefficient (described earlier) and Spearman's rho (see section 6.5.3) are examples of bivariate correlation coefficients.

Let's return to the example from Chapter 4 about exam scores. Remember that a psychologist was interested in the effects of exam stress and revision on exam performance. She had devised and validated a questionnaire to assess state anxiety relating to exams (called the Exam Anxiety Questionnaire, or EAQ). This scale produced a measure of anxiety scored out of 100. Anxiety was measured before an exam, and the percentage mark of each student on the exam was used to assess the exam performance. She also measured the number of hours spent revising. These data are in **ExamAnxiety.sas7bdat** on the companion website. We have already created scatterplots for these data (section 4.7) so we don't need to do that again.

To conduct bivariate correlations we use PROC CORR. In the first line we list the data set, and then on the VAR line we list the variables that we want to correlate. SAS Syntax 6.1 shows how PROC CORR is used.

```
PROC CORR data=chapter6.examanxiety;
    VAR revise exam anxiety;
    RUN;
```

SAS Syntax 6.1

The default setting is Pearson's product-moment correlation, but you can also calculate Spearman's correlation and Kendall's correlation—we will see the differences between these correlation coefficients in due course.

6.5.2. Pearson's correlation coefficient ①

6.5.2.1. Assumptions of Pearson's *r* ③

Pearson's (Figure 6.3) correlation coefficient was described in full at the beginning of this chapter. Pearson's correlation requires only that data are interval (see section 1.5.1.2) for it to be an accurate measure of the linear relationship between two variables. However, if you want to establish whether the correlation coefficient is significant, then more assumptions are required: for the test statistic to be valid the sampling distribution has to be normally distributed and, as we saw in Chapter 5, we assume that it is if our sample data are normally distributed (or if we have a large sample). Although, typically, to assume that the sampling distribution is normal, we would want both variables to be normally distributed, there is one exception to this rule: one of the variables can be a categorical variable provided there are only two categories (this is the same as doing a *t*-test, but I'm jumping the gun a bit). In any case, if your data are non-normal (see Chapter 5) or are not measured at the interval level then you should not use a Pearson correlation.

OLIVER TWISTED

Please, Sir, can I have some more … options?

Oliver is so excited to get onto analysing his data that he doesn't want me to spend pages waffling on about options that you will probably never use. 'Stop writing, you waffling fool,' he says. 'I want to analyse my data.' Well, he's got a point. If you want to find out more about what other options are available in SAS PROC CORR, then the additional material for this chapter on the companion website will tell you.

SAS TIP 6.1 Pairwise or listwise? ①

As we run through the various analyses in this book, many of them have additional options. One common option is to choose whether you do 'pairwise', 'analysis by analysis' or 'listwise'. First, we can exclude cases listwise, which means that if a case has a missing value for any variable, then they are excluded from the whole analysis. So, for example, in our exam anxiety data if one of our students had reported their anxiety and we knew their exam performance but we didn't have data about their revision time, then their data would not be used to calculate any of the correlations: *they would be completely excluded from the analysis*. Another option is to excluded cases on a pairwise or analysis-by-analysis basis, which means that if a participant has a score missing for a particular variable or analysis, then their data are excluded only from calculations involving the variable for which they have no score. For our student about whom we don't have any revision data, this means that their data would be excluded when calculating the correlation between exam scores and revision time, and when calculating the correlation between exam anxiety and revision time; however, the student's scores would be *included* when calculating the correlation between exam anxiety and exam performance because for this pair of variables we have both of their scores.

SAS PROC CORR does pairwise deletion by default – that is, a person is included wherever possible. If you want to exclude people who have missing data for one variable, use the nomiss option on the PROC CORR line. You would write:

```
PROC CORR DATA=chapter6.examanxiety nomiss;
```

6.5.2.2. Running Pearson's *r* on SAS ①

We have already seen the syntax for PROC CORR (in SAS Syntax 6.1). SAS produces Pearson's correlation coefficient by default. Our researcher predicted that (1) as anxiety increases, exam performance will decrease, and (2) as the time spent revising increases, exam performance will increase.

SAS Output 6.1 first provides descriptive statistics – mean, standard deviation, sum, minimum and maximum – and then provides a matrix of the correlation coefficients for the three

FIGURE 6.3
Karl Pearson

variables. Underneath each correlation coefficient the significance value of the correlation is displayed (if there were missing data, it would also show the sample size for each correlation – as there are no missing data, the sample size is always the same and it tells us in the title to the table). Each variable is perfectly correlated with itself (obviously) and so $r = 1$ along the diagonal of the table. Exam performance is negatively related to exam anxiety with a Pearson correlation coefficient of $r = -.441$ and the significance value is a less than .0001. This significance value tells us that the probability of getting a correlation coefficient this big in a sample of 103 people if the null hypothesis were true (there was no relationship between these variables) is very low (close to zero in fact). Hence, we can gain confidence that there is a genuine relationship between exam performance and anxiety. (Remember that our criterion for statistical significance is usually less than 0.05.) The output also shows that exam performance is positively related to the amount of time spent revising, with a coefficient of $r = .397$, which is also significant at $p < .0001$. Finally, exam anxiety appears to be negatively related to the time spent revising, $r = -.709$, $p < .0001$.

In psychological terms, this all means that people who have higher anxiety about an exam obtain a lower percentage mark in that exam. Conversely, people who spend more time revising obtain higher marks in the exam. Finally, people who spend more time revising, have lower anxiety about the exam. So there is a complex interrelationship between the three variables.

6.5.2.3. Using R^2 for interpretation ③

Although we cannot make direct conclusions about causality from a correlation, there is still more that it can tell us. The square of the correlation coefficient (known as the

SAS OUTPUT 6.1

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
REVISE	103	19.85437	18.15910	2045	0	98.00000	Time Spent Revising
EXAM	103	56.57282	25.94058	5827	2.00000	100.00000	Exam Performance (%)
ANXIETY	103	74.34367	17.18186	7657	0.05600	97.58200	Exam Anxiety

Pearson Correlation Coefficients, N = 103 Prob > |r| under H0: Rho=0			
	REVISE	EXAM	ANXIETY
REVISE Time Spent Revising	1.00000	0.39672 <.0001	–0.70925 <.0001
EXAM Exam Performance (%)	0.39672 <.0001	1.00000	–0.44099 <.0001
ANXIETY Exam Anxiety	–0.70925 <.0001	–0.44099 <.0001	1.00000

coefficient of determination, R^2) is a measure of the amount of variability in one variable that is shared by the other. For example, we may look at the relationship between exam anxiety and exam performance. Exam performances vary from person to person because of any number of factors (different ability, different levels of preparation and so on). If we add up all of this variability (rather like when we calculated the sum of squares in section 2.4.1) then we would have an estimate of how much variability exists in exam performances. We can then use R^2 to tell us how much of this variability is shared by exam anxiety. These two variables had a correlation of –0.4410 and so the value of R^2 will be $(-0.4410)^2 = 0.194$. This value tells us how much of the variability in exam performance is shared by exam anxiety.

If we convert this value into a percentage (multiply by 100) we can say that exam anxiety shares 19.4% of the variability in exam performance. So, although exam anxiety was highly correlated with exam performance, it can account for only 19.4% of variation in exam scores. To put this value into perspective, this leaves 80.6% of the variability still to be accounted for by other variables. I should note at this point that although R^2 is an extremely useful measure of the substantive importance of an effect, it cannot be used to infer causal relationships. Although we usually talk in terms of 'the variance in *y* *accounted for* by *x*', or even the variation in one variable *explained* by the other, this still says nothing about which way causality runs. So, although exam anxiety can account for 19.4% of the variation in exam scores, it does not necessarily cause this variation – it may be that people who do well feel anxious because they are under pressure to do well. If Andy released a CD tomorrow he would be under less pressure to sell one million copies than if Metallica released a CD tomorrow.

6.5.3. Spearman's correlation coefficient ①

Spearman's correlation coefficient (Spearman, 1910; Figure 6.4), r_s, is a non-parametric statistic and so can be used when the data have violated parametric assumptions such as non-normally distributed data (see Chapter 5). You'll sometimes hear the test referred to as Spearman's rho (pronounced 'row', as in 'row your boat gently down the stream'), which does make it difficult for some people to distinguish from the London lap-dancing club Spearmint Rhino.[3] Spearman's test works by first ranking the data (see section 15.3.1), and then applying Pearson's equation (6.3) to those ranks.

I was born in England, which has some bizarre traditions. One such oddity is the World's Biggest Liar Competition held annually at the Santon Bridge Inn in Wasdale (in the Lake District). The contest honours a local publican, 'Auld Will Ritson', who in the nineteenth century was famous in the area for his far-fetched stories (one such tale being that Wasdale turnips were big enough to be hollowed out and used as garden sheds). Each year locals are encouraged to attempt to tell the biggest lie in the world (lawyers and politicians are apparently banned from the competition). Over the years there have been tales of mermaid farms, giant moles, and farting sheep blowing holes in the ozone layer. (I am thinking of entering next year and reading out some sections of this book.)

Imagine I wanted to test a theory that more creative people will be able to create taller tales. I gathered together 68 past contestants from this competition and asked them where they were placed in the competition (first, second, third, etc.) and also gave them a creativity questionnaire (maximum score 60). The position in the competition is an ordinal variable (see section 1.5.1.2) because the places are categories but have a meaningful order (first place is better than second place and so on). Therefore, Spearman's correlation coefficient should be used (Pearson's *r* requires interval or ratio data). The data for this study are in the file **TheBiggestLiar.sas7bdat.** The data are in two columns: one labelled **Creativity** and one labelled **Position** (there's actually a third variable in there but we will ignore it for the time being). For the **Position** variable, each of the categories described above has been coded with a numerical value. First place has been coded with the value 1, with positions being labelled 2, 3 and so on. Note that for each numeric code I have provided a value label (just like we did for coding variables).

The procedure for doing a Spearman correlation is the same as for a Pearson correlation except on the options line we add the word SPEARMAN. This is shown in SAS Syntax 6.2.

```
PROC CORR data=chapter6.thebiggestliar SPEARMAN;
    VAR creativity position;
    RUN;
```

SAS Syntax 6.2

SAS Output 6.2 shows the output for a Spearman correlation on the variables **Creativity** and **Position**. The output is very similar to that of the Pearson correlation: a matrix is

[3]Seriously, a colleague of mine asked a student what analysis she was thinking of doing and she responded 'a Spearman's Rhino'.

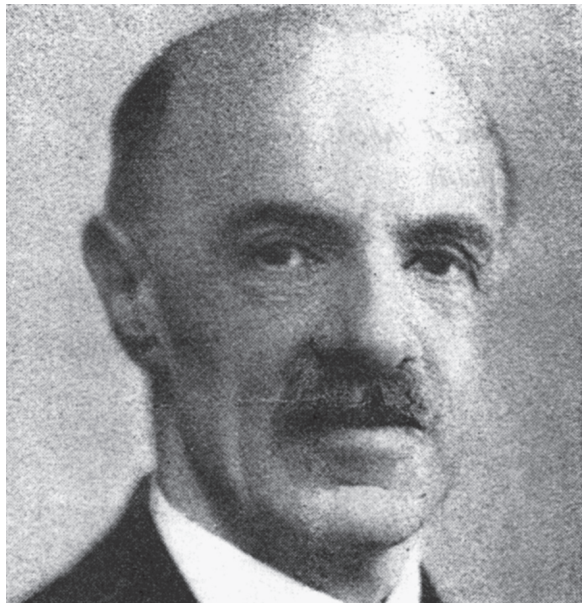


FIGURE 6.4
Charles Spearman, ranking furiously

SAS OUTPUT 6.2

Simple Statistics							
Variable	N	Mean	Std Dev	Median	Minimum	Maximum	Label
CREATIVITY	68	39.98529	8.11759	39.00000	21.00000	56.00000	Creativity
POSITION	68	2.22059	1.38052	2.00000	1.00000	6.00000	Position in Best Liar Competition

Spearman Correlation Coefficients, N = 68 Prob > |r| under H0: Rho=0		
	CREATIVITY	POSITION
CREATIVITY Creativity	1.00000	–0.30022 0.0017
POSITION Position in Best Liar Competition	–0.37322 0.0017	1.00000

displayed giving the correlation coefficient between the two variables (–.373), underneath is the significance value of this coefficient (.0017). The significance value for this correlation coefficient is less than .05; therefore, it can be concluded that there is a significant relationship between creativity scores and how well someone did in the World's Biggest Liar Competition. Note that the relationship is negative: as creativity increased, position decreased. This might seem contrary to what we predicted until you remember

that a low number means that you did well in the competition (a low number such as 1 means you came first, and a high number like 4 means you came fourth). Therefore, our hypothesis is supported: as creativity increased, so did success in the competition.

SELF-TEST Did creativity cause success in the World's Biggest Liar Competition?

6.5.4. Kendall's tau (non-parametric) ②

Kendall's tau, τ, is another non-parametric correlation and it should be used rather than Spearman's coefficient when you have a small data set with a large number of tied ranks. This means that if you rank all of the scores and many scores have the same rank, then Kendall's tau should be used. Although Spearman's statistic is the more popular of the two coefficients, there is much to suggest that Kendall's statistic is actually a better estimate of the correlation in the population (see Howell, 1997, p. 293). As such, we can draw more accurate generalizations from Kendall's statistic than from Spearman's. To carry out Kendall's correlation on the world's biggest liar data simply follow the same steps as for Pearson and Spearman correlations but write KENDALL on the PROC CORR line, instead of SPEARMAN. The output is much the same as for Spearman's correlation.

SAS OUTPUT 6.3

Kendall Tau b Correlation Coefficients, N = 68 Prob > |tau| under H0: Tau=0		
	CREATIVITY	**POSITION**
CREATIVITY Creativity	1.00000	–0.30024 0.0013
POSITION Position in Best Liar Competition	–0.30024 0.0013	1.00000

You'll notice from SAS Output 6.3 that the actual value of the correlation coefficient is closer to zero than the Spearman correlation (it has increased from –.373 to –.300). Despite the difference in the correlation coefficients we can still interpret this result as being a highly significant relationship (because the significance value of .001 is less than .05). However, Kendall's value is a more accurate gauge of what the correlation in the population would be. As with the Pearson correlation, we cannot assume that creativity caused success in the World's Best Liar Competition.

SELF-TEST Conduct a Pearson correlation analysis of the advert data from the beginning of the chapter.

CRAMMING SAM'S TIPS

- We can measure the relationship between two variables using *correlation coefficients*.
- These coefficients lie between –1 and +1.
- *Pearson's correlation coefficient, r,* is a parametric statistic and requires interval data for both variables. To test its significance we assume normality too.
- *Spearman's correlation coefficient,* r_s, is a non-parametric statistic and requires only ordinal data for both variables.
- *Kendall's correlation coefficient,* τ, is like Spearman's r_s but probably better for small samples.

6.6. Partial correlation ②

6.6.1. The theory behind part and partial correlation ②

I mentioned earlier that there is a type of correlation that can be done that allows you to look at the relationship between two variables when the effects of a third variable are held constant. For example, analyses of the exam anxiety data (in the file **ExamAnxiety.sas 7bdat**) showed that exam performance was negatively related to exam anxiety, but positively related to revision time, and revision time itself was negatively related to exam anxiety. This scenario is complex, but given that we know that revision time is related to both exam anxiety and exam performance, then if we want a pure measure of the relationship between exam anxiety and exam performance we need to take account of the influence of revision time. Using the values of R^2 for these relationships, we know that exam anxiety accounts for 19.4% of the variance in exam performance, that revision time accounts for 15.7% of the variance in exam performance, and that revision time accounts for 50.2% of the variance in exam anxiety. If revision time accounts for half of the variance in exam anxiety, then it seems feasible that at least some of the 19.4% of variance in exam performance that is accounted for by anxiety is the same variance that is accounted for by revision time. As such, some of the variance in exam performance explained by exam anxiety is not *unique* and can be accounted for by revision time. A correlation between two variables in which the effects of other variables are held constant is known as a *partial correlation*.

Figure 6.7 illustrates the principle behind partial correlation. In part 1 of the diagram there is a box for exam performance that represents the total variation in exam scores (this value would be the variance of exam performance). There is also a box that represents the variation in exam anxiety (again, this is the variance of that variable). We know already that exam anxiety and exam performance share 19.4% of their variation (this value is the correlation coefficient squared). Therefore, the variations of these two variables overlap (because they share variance) creating a third box (the one with diagonal lines). The overlap of the boxes representing exam performance and exam anxiety is the common variance. Likewise, in part 2 of the diagram the shared variation between exam performance and revision time is illustrated. Revision time shares 15.7% of the variation in exam scores. This shared variation is represented by the area of overlap (filled with diagonal lines). We know that revision time and exam anxiety also share 50% of their variation; therefore, it is very probable that some of

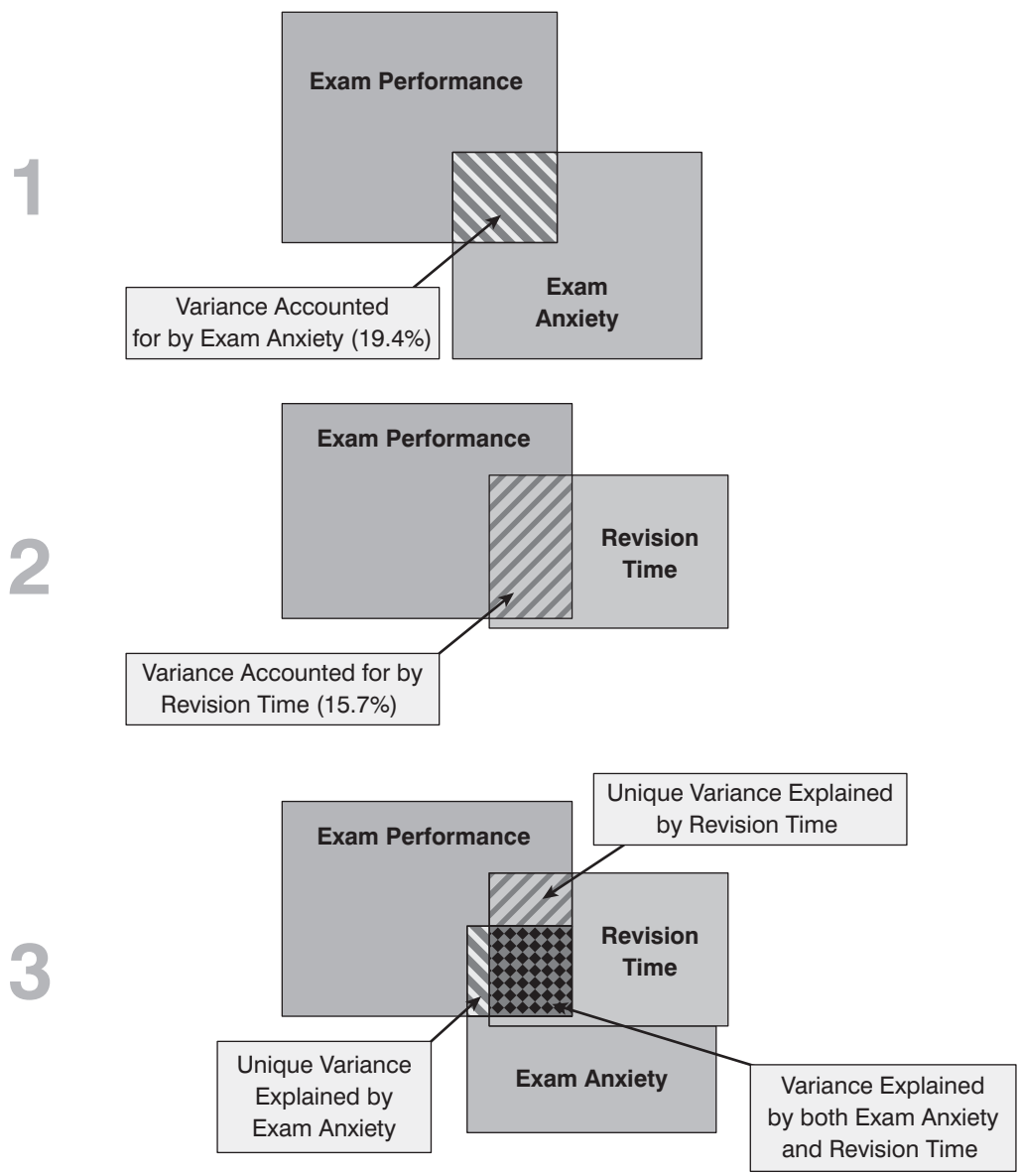


FIGURE 6.5
Diagram showing the principle of partial correlation

the variation in exam performance shared by exam anxiety is the same as the variance shared by revision time.

Part 3 of the diagram shows the complete picture. The first thing to note is that the boxes representing exam anxiety and revision time have a large overlap (this is because they share 50% of their variation). More important, when we look at how revision time and anxiety contribute to exam performance we see that there is a portion of exam performance that is shared by both anxiety and revision time (the dotted area). However, there are still small chunks of the variance in exam performance that are unique to the other two variables. So, although in part 1 exam anxiety shared a large chunk of variation in exam performance, some of this overlap is also shared by revision time. If we remove the portion of variation that is also shared by revision time, we get a measure of the unique relationship between exam performance and exam anxiety. We use partial correlations to find out the size of the unique portion of variance. Therefore, we could conduct a partial correlation between exam anxiety and exam performance while 'controlling' for the effect of revision time. Likewise, we could carry out a partial correlation between revision time and exam performance while 'controlling' for the effects of exam anxiety.

6.6.2. Partial correlation Using SAS ②

Let's use the **ExamAnxiety.sas7bdat** file so that, as I suggested above, we can conduct a partial correlation between exam anxiety and exam performance while 'controlling' for the effect of revision time.

We calculate partial correlations very similarly to correlations, using PROC CORR. Variables that we want to control for (partial out) are added in a line called (you might guess) PARTIAL. So to examine the relationship between test scores and test anxiety, controlling for revision time, we use the syntax shown in SAS Syntax 6.3.

```
PROC CORR data=chapter6.examanxiety;
  VAR exam anxiety;
  PARTIAL revise;
  RUN;
```

SAS Syntax 6.3

SAS OUTPUT 6.4

Pearson Partial Correlation Coefficients, N = 103 Prob > |r| under H0: Partial Rho=0		
	EXAM	**ANXIETY**
EXAM Exam Performance (%)	1.00000	–0.24667 0.0124
ANXIETY Exam Anxiety	–0.24667 0.0124	1.00000

SAS Output 6.4 shows the output for the partial correlation of exam anxiety and exam performance controlling for revision time. There is a matrix of correlations for the variables **anxiety** and **exam** but controlling for the effect of revision. In this instance we have controlled for one variable and so this is known as a first-order partial correlation. It is possible to

control for the effects of two variables at the same time (a second-order partial correlation) or control three variables (a third-order partial correlation) and so on. First, notice that the partial correlation between exam performance and exam anxiety is –.247, which is considerably less than the correlation when the effect of revision time is not controlled for ($r = -.441$). In fact, the correlation coefficient is only about half what it was before. Although this correlation is still statistically significant (its p-value is still below .05), the relationship is diminished. In terms of variance, the value of R^2 for the partial correlation is .06, which means that exam anxiety can now account for only 6% of the variance in exam performance. When the effects of revision time were not controlled for, exam anxiety shared 19.4% of the variation in exam scores and so the inclusion of revision time has severely diminished the amount of variation in exam scores shared by anxiety. As such, a truer measure of the role of exam anxiety has been obtained. Running this analysis has shown us that exam anxiety alone does explain some of the variation in exam scores, but there is a complex relationship between anxiety, revision and exam performance that might otherwise have been ignored. Although causality is still not certain, because relevant variables are being included, the third-variable problem is, at least, being addressed in some form.

These partial correlations can be done when variables are dichotomous (including the 'third' variable). So, for example, we could look at the relationship between bladder relaxation (did the person wet themselves or not?) and the number of large tarantulas crawling up your leg, controlling for fear of spiders (the first variable is dichotomous, but the second variable and 'controlled for' variables are continuous). Also, to use an earlier example, we could examine the relationship between creativity and success in the World's Greater Liar Contest controlling for whether someone had previous experience in the competition (and therefore had some idea of the type of tale that would win) or not. In this latter case the 'controlled for' variable is dichotomous.[4]

6.6.3. Semi-partial (or part) correlations ②

In the next chapter, we will come across another form of correlation known as a **semi-partial correlation** (also referred to as a part correlation). While I'm babbling on about partial correlations it is worth my explaining the difference between this type of correlation and a semi-partial correlation. When we do a partial correlation between two variables, we control for the effects of a third variable. Specifically, the effect that the third variable has on *both* variables in the correlation is controlled. In a semi-partial correlation we control for the effect that the third variable has on only one of the variables in the correlation. Figure 6.6 illustrates this principle for the exam performance data. The partial correlation that we calculated took account not only of the effect of revision on exam performance, but also of the effect of revision on anxiety. If we were to calculate the semi-partial correlation for the same data, then this would control for only the effect of revision on exam performance (the effect of revision on exam anxiety is ignored). Partial correlations are most useful for looking at the unique relationship between two variables when other variables are ruled out. Semi-partial correlations are, therefore, useful when trying to explain the variance in one particular

FIGURE 6.6
The difference between a partial and a semi-partial correlation

[4] Both these examples are, in fact, simple cases of hierarchical regression (see the next chapter) and the first example is also an example of analysis of covariance. This may be confusing now, but as we progress through the book I hope it'll become clearer that virtually all of the statistics that you use are actually the same things dressed up in different names.

CRAMMING SAM'S TIPS

- A *partial correlation* quantifies the relationship between two variables while controlling for the effects of a third variable on *both* variables in the original correlation.
- A *semi-partial correlation* quantifies the relationship between two variables while controlling for the effects of a third variable on only *one* of the variables in the original correlation.

variable (an outcome) from a set of predictor variables. (Bear this in mind when you read Chapter 7.)

6.7. Comparing correlations ③

6.7.1. Comparing independent *rs* ③

Sometimes we want to know whether one correlation coefficient is bigger than another. For example, when we looked at the effect of exam anxiety on exam performance, we might have been interested to know whether this correlation was different in men and women. We could compute the correlation in these two samples, but then how would we assess whether the difference was meaningful?

SELF-TEST Use the BY command to compute the correlation coefficient between exam anxiety and exam performance in men and women. (Remember to sort by gender with PROC SORT, and then add: BY gender; to the PROC CORR command.

If we did this, we would find that the correlations were $r_{Male} = -.506$ and $r_{Female} = -.381$. These two samples are independent; that is, they contain different entities. To compare these correlations we can again use what we discovered in section 6.3.3 to convert these coefficients to z_r (just to remind you, we do this because it makes the sampling distribution normal and we know the standard error). If you do the conversion, then we get z_r (males) = $-.557$ and z_r (females) = $-.401$. We can calculate a z-score of the differences between these correlations as:

$$z_{Difference} = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 + 3}}} \tag{6.9}$$

We had 52 men and 51 women so we would get:

$$z_{Difference} = \frac{-.557 - (-.401)}{\sqrt{\frac{1}{49} + \frac{1}{48}}} = \frac{-.156}{0.203} = -0.768$$

We can look up this value of z (0.768, we can ignore the minus sign) in the table for the normal distribution in the Appendix and get the one-tailed probability from the column labelled 'Smaller Portion'. In this case the value is .221. To get the two-tailed probability we simply multiply the one-tailed probability value by 2, which gives us .442. As such the correlation between exam anxiety and exam performance is not significantly different in men and women.

6.7.2. Comparing dependent *rs* ②

If you want to compare correlation coefficients that come from the same entities then things are a little more complicated. You can use a *t*-statistic to test whether a difference between two dependent correlations from the same sample is significant. For example, in our exam anxiety data we might want to see whether the relationship between exam anxiety (x) and exam performance (y) is stronger than the relationship between revision (z) and exam performance. To calculate this, all we need are the three rs that quantify the relationships between these variables: r_{xy}, the relationship between exam anxiety and exam performance (–.441); r_{zy}, the relationship between revision and exam performance (.397); and r_{xz}, the relationship between exam anxiety and revision (–.709). The *t*-statistic is computed as (Chen & Popovich, 2002):

$$t_{Difference} = (r_{xy} - r_{zy}) \sqrt{\frac{(n-3)(1+r_{xz})}{2\left(1 - r_{xy}^2 - r_{xz}^2 - r_{zy}^2 + 2r_{xy}r_{xz}r_{zy}\right)}} \tag{6.10}$$

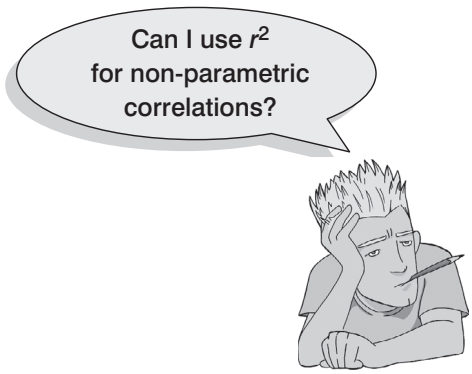
Admittedly that equation looks hideous, but really it's not too bad: it just uses the three correlation coefficients and the sample size N. Place the numbers from the exam anxiety example in it (N was 103) and you should end up with:

$$t_{Difference} = (-.838) \sqrt{\frac{29.1}{2(1 - .194 - .503 - .158 + 0.248}} = -5.09$$

This value can be checked against the appropriate critical value in the Appendix with $N - 3$ degrees of freedom (in this case 100). The critical values in the table are 1.98 ($p < .05$) and 2.63 ($p < .01$), two-tailed. As such we can say that the correlation between exam anxiety and exam performance was significantly higher (more towards the positive) than the correlation between revision time and exam performance (this isn't a massive surprise given that these relationships went in the opposite directions to each other).

6.8. Calculating the effect size ①

Calculating effect sizes for correlation coefficients couldn't be easier because, as we saw earlier in the book, correlation coefficients *are* effect sizes! So, no calculations (other than those you have already done) necessary! However, I do want to point out one caveat when using non-parametric correlation coefficients as effect sizes. Although the Spearman and Kendall correlations are comparable in many respects (their power, for example, is similar under parametric conditions), there are two important differences (Strahan, 1982).



First, we saw for Pearson's r that we can square this value to get the proportion of shared variance, R^2. For Spearman's r_s we can do this too because it uses the same equation as Pearson's r. However, the resulting R_s^2 needs to be interpreted slightly differently: it is the proportion of variance in the *ranks* that two variables share. Having said this, R_s^2 is usually a good approximation of R^2 (especially in conditions of near-normal distributions). Kendall's τ, however, is not numerically similar to either r or r_s and so τ^2 does not tell us about the proportion of variance shared by two variables (or the ranks of those two variables).

Second, Kendall's τ is 66–75% smaller than both Spearman's r_s and Pearson's r, but r and r_s are generally similar sizes (Strahan, 1982). As such, if τ is used as an effect size it should be borne in mind that it is not comparable to r and r_s and should not be squared. More generally, when using correlations as effect sizes you should remember (both when reporting your own analysis and when interpreting others) that the choice of correlation coefficient can make a substantial difference to the apparent size of the effect.

6.9. How to report correlation coefficents ①

Reporting correlation coefficients is pretty easy: you just have to say how big they are and what their significance value was (although the significance value isn't *that* important because the correlation coefficient is an effect size in its own right!). Four things to note are that: (1) coefficients are reported to 2 decimal places; (2) if you are quoting a one-tailed probability, you should say so; (3) each correlation coefficient is represented by a different letter (and some of them are Greek!); and (4) there are standard criteria of probabilities that we use (.05, .01 and .001). Let's take a few examples from this chapter:

- ✓ There was a significant relationship between the number of adverts watched and the number of packets of sweets purchased, $r = .87$, p (one-tailed) $< .05$.
- ✓ Exam performance was significantly correlated with exam anxiety, $r = -.44$, and time spent revising, $r = .40$; the time spent revising was also correlated with exam anxiety, $r = -.71$ (all ps $< .001$).
- ✓ Creativity was significantly related to how well people did in the World's Biggest Liar Competition, $r_s = -.37$, $p < .001$.

✓ Creativity was significantly related to how well people did in the World's Biggest Liar Competition, $\tau = -.30$, $p < .001$. (Note that I've quoted Kendall's τ here.)

Scientists, rightly or wrongly, tend to use several *standard* levels of statistical significance. Primarily, the most important criterion is that the significance value is less than .05; however, if the exact significance value is much lower then we can be much more confident about the strength of the experimental effect. In these circumstances we like to make a big song and dance about the fact that our result isn't just significant at .05, but is significant at a much lower level as well (hooray!). The values we use are .05, .01, and .001.

TABLE 6.2 An example of reporting a table of correlations

	Exam Performance	*Exam Anxiety*	*Revision Time*
Exam Performance	1	–.44***	.40***
Exam Anxiety	103	1	–.71***
Revision Time	103	103	1

Ns = not significant ($p > .05$), *$p < .05$, ** $p < .01$, *** $p < .001$

When we have lots of correlations we sometimes put them into a table. For example, our exam anxiety correlations could be reported as in Table 6.2. Note that above the diagonal I have reported the correlation coefficients and used symbols to represent different levels of significance. Under the table there is a legend to tell readers what symbols represent. (Actually, none of the correlations were non-significant or had p bigger than .001 so most of these are here simply to give you a reference point – you would normally include symbols that you had actually used in the table in your legend.) Finally, in the lower part of the table I have reported the sample sizes. These are all the same (103) but sometimes when you have missing data it is useful to report the sample sizes in this way because different values of the correlation will be based on different sample sizes. For some more ideas on how to report correlations have a look at Labcoat Leni's Real Research 6.1.

LABCOAT LENI'S REAL RESEARCH 6.1

Why do you like your lecturers? ①

As students you probably have to rate your lecturers at the end of the course. There will be some lecturers you like and others that you hate. As a lecturer I find this process horribly depressing (although this has a lot to do with the fact that I tend focus on negative feedback and ignore the good stuff). There is some evidence that students tend to pick courses of lecturers who they perceive to be enthusastic and good communicators. In a fascinating study, Tomas Chamorro-Premuzic and his colleagues (Chamorro-Premuzic, Furnham, Christopher, Garwood, & Martin, 2008) tested a slightly different hypothesis, which was that students tend to like lecturers who are like themselves. (This hypothesis will have the students on my course who like my lectures screaming in horror.)

First of all the authors measured students' own personalities using a very well-established measure (the NEO-FFI) which gives rise to scores on five fundamental personality traits: Neuroticism, Extroversion, Openness to experience, Agreeableness and Conscientiousness. They also gave students a questionnaire that asked them to rate how much they wanted their lecturer to have each of a list of characteristics. For example, they would be given the description 'warm: friendly, warm, sociable, cheerful, affectionate, outgoing' and asked to rate how much they wanted to see this in a lecturer from –5 (they don't want this characteristic at all) through 0 (the characteristic is not important) to +5 (I really want this characteristic in my lecturer). The characteristics on the questionnaire all related to personality characteristics measured by the NEO-FFI. As such, the authors had a measure of how much a student had each of the five core personality characteristics, but also a measure of how much they wanted to see those same characteristics in their lecturer.

In doing so, Tomas and his colleagues could test whether, for instance, extroverted students want extrovert lecturers. The data from this study (well, for the variables that I've mentioned) are in the file **Chamorro Premuzic.sas7bdat**. Run some Pearson correlations on these variables to see if students with certain personality characteristics want to see those characteristics in their lecturers. What conclusions can you draw?

Answers are in the additional material on the companion website (or look at Table 3 in the original article, which will also show you how to report a large number of correlations).

CHAMORRO-PREMUZIC, T., ET AL. (2008). *PERSONALITY AND INDIVIDUAL DIFFERENCES*, *44*, 965–976.

What have I discovered about statistics? ①

This chapter has looked at ways to study relationships between variables. We began by looking at how we might measure relationships statistically by developing what we already know about variance (from Chapter 1) to look at variance shared between variables. This shared variance is known as *covariance*. We then discovered that when data are parametric we can measure the strength of a relationship using Pearson's correlation coefficient, r. When data violate the assumptions of parametric tests we can use Spearman's r_s, or for small data sets Kendall's τ may be more accurate. We also saw that correlations can be calculated between two variables when one of those variables is a dichotomy (i.e. composed of two categories). Finally, we looked at the difference between *partial correlations*, in which the relationship between two variables is measured controlling for the effect that one or more variables has on both of those variables, and *semi-partial correlations*, in which the relationship between two variables is measured controlling for the effect that one or more variables has on only one of those variables. We also discovered that I had a guitar and, like my favourite record of the time, I was ready to 'Take on the world'. Well, Wales at any rate …

Key terms that I've discovered

Bivariate correlation
Coefficient of determination
Covariance
Cross-product deviations
Kendall's tau
Partial correlation
Pearson correlation coefficient
Semi-partial correlation
Spearman's correlation coefficient
Standardization

Smart Alex's tasks

- **Task 1**: A student was interested in whether there was a positive relationship between the time spent doing an essay and the mark received. He got 45 of his friends and timed how long they spent writing an essay (**hours**) and the percentage they got in the essay (**essay**). He also translated these grades into their degree classifications (**grade**): first, upper second, lower second and third class. Using the data in the file **EssayMarks.sas7bdat** find out what the relationship was between the time spent doing an essay and the eventual mark in terms of percentage and degree class (draw a scatterplot too!). ①

- **Task 2**: Using the **ChickFlick.sas7bdat.** data from Chapter 3, is there a relationship between gender and arousal? Using the same data, is there a relationship between the film watched and arousal? ①

- **Task 3**: As a statistics lecturer I am always interested in the factors that determine whether a student will do well on a statistics course. One potentially important factor is their previous expertise with mathematics. Imagine I took 25 students and looked at their degree grades for my statistics course at the end of their first year at university. In the UK, a student can get a first-class mark (the best), an upper-second-class mark, a lower second, a third, a pass or a fail (the worst). I also asked these students what grade they got in their GCSE maths exams. In the UK GCSEs are school exams taken at age 16 that are graded A, B, C, D, E or F (an A grade is better than all of the lower grades). The data for this study are in the file **grades.sas7bdat.** Carry out the appropriate analysis to see if GCSE maths grades correlate with first-year statistics grades. ①

Answers can be found on the companion website.

Further reading

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.

Howell, D. C. (2006). *Statistical methods for psychology* (6th ed.). Belmont, CA: Duxbury. (Or you might prefer his *Fundamental statistics for the behavioral sciences,* also in its 6th edition, 2007. Both are excellent texts that are a bit more technical than this book so they are a useful next step.)

Miles, J. N. V., & Banyard, P. (2007). *Understanding and using statistics in psychology: a practical introduction*. London: Sage. (A fantastic and amusing introduction to statistical theory.)

Wright, D. B., & London, K. (2009). *First steps in statistics* (2nd ed.). London: Sage. (This book is a very gentle introduction to statistical theory.)

Interesting real research

Chamorro-Premuzic, T., Furnham, A., Christopher, A. N., Garwood, J., & Martin, N. (2008). Birds of a feather: Students' preferences for lecturers' personalities as predicted by their own personality and learning approaches. *Personality and Individual Differences*, *44*, 965–976.

Regression

7

FIGURE 7.1 Me playing with my ding-a-ling in the Holimarine Talent Show. Note the groupies queuing up at the front

7.1. What will this chapter tell me? ①

Although none of us can know the future, predicting it is so important that organisms are hard wired to learn about predictable events in their environment. We saw in the previous chapter that I received a guitar for Christmas when I was 8. My first foray into public performance was a weekly talent show at a holiday camp called 'Holimarine' in Wales (it doesn't exist anymore because I am old and this was 1981). I sang a Chuck Berry song called 'My ding-a-ling'[1] and to my absolute amazement I won the competition.[2] Suddenly other 8 year olds across the land (well, a ballroom in Wales) worshipped me (I made lots of friends after the competition). I had tasted success, it tasted like praline chocolate, and so I wanted to enter the competition in the second week of our holiday. To ensure success, I needed to know why I had won in the first week. One way to do this would have been to collect data and to use these data to predict people's evaluations of children's performances in the contest

[1] It appears that even then I had a passion for lowering the tone of things that should be taken seriously.

[2] I have a very grainy video of this performance recorded by my dad's friend on a video camera the size of a medium-sized dog that had to be accompanied at all times by a 'battery pack' the size and weight of a tank. Maybe I'll put it up on the companion website …

from certain variables: the age of the performer, what type of performance they gave (singing, telling a joke, magic tricks), and maybe how cute they looked. A regression analysis on these data would enable us to predict future evaluations (success in next week's competition) based on values of the predictor variables. If, for example, singing was an important factor in getting a good audience evaluation, then I could sing again the following week; however, if jokers tended to do better then I could switch to a comedy routine. When I was 8 I wasn't the sad geek that I am today, so I didn't know about regression analysis (nor did I wish to know); however, my dad thought that success was due to the winning combination of a cherub-looking 8 year old singing songs that can be interpreted in a filthy way. He wrote me a song to sing in the competition about the keyboard player in the Holimarine Band 'messing about with his organ', and first place was mine again. There's no accounting for taste.

7.2. An introduction to regression ①

In the previous chapter we looked at how to measure relationships between two variables. These correlations can be very useful but we can take this process a step further and predict one variable from another. A simple example might be to try to predict levels of stress from the amount of time until you have to give a talk. You'd expect this to be a negative relationship (the smaller the amount of time until the talk, the larger the anxiety). We could then extend this basic relationship to answer a question such as 'if there's 10 minutes to go until someone has to give a talk, how anxious will they be?' This is the essence of regression analysis: we fit a model to our data and use it to predict values of the dependent variable from one or more independent variables.[3] Regression analysis is a way of predicting an **outcome variable** from one **predictor variable** (**simple regression**) or several predictor variables (**multiple regression**). This tool is incredibly useful because it allows us to go a step beyond the data that we collected.

In section 2.4.3 I introduced you to the idea that we can predict any data using the following general equation:

$$\text{outcome}_i = (\text{model}) + \text{error}_i \tag{7.1}$$

This just means that the outcome we're interested in for a particular person can be predicted by whatever model we fit to the data plus some kind of error. In regression, the model we fit is linear, which means that we summarize a data set with a straight line (think back to Jane Superbrain Box 2.1). As such, the word 'model' in the equation above simply gets replaced by 'things that define the line that we fit to the data' (see the next section).

With any data set there are several lines that could be used to summarize the general trend and so we need a way to decide which of many possible lines to chose. For the sake of making accurate predictions we want to fit a model that *best* describes the data. The simplest way to do this would be to use your eye to gauge a line that looks as though it summarizes the data well. You don't need to be a genius to realize that the 'eyeball' method is very subjective and so offers no assurance that the model is the best one that could have been chosen. Instead, we use a mathematical technique called the *method of least squares* to establish the line that best describes the data collected.

[3] I want to remind you here of something I discussed in Chapter 1: SAS refers to regression variables as dependent and independent variables (as in controlled experiments). However, correlational research by its nature seldom controls the independent variables to measure the effect on a dependent variable and so I will talk about 'independent variables' as *predictors*, and the 'dependent variable' as the *outcome*.

7.2.1. Some important information about straight lines ①

I mentioned above that in our general equation the word 'model' gets replaced by 'things that define the line that we fit to the data'. In fact, any straight line can be defined by two things: (1) the slope (or gradient) of the line (usually denoted by b_1); and (2) the point at which the line crosses the vertical axis of the graph (known as the *intercept* of the line, b_0). In fact, our general model becomes equation (7.2) below in which Y_i is the outcome that we want to predict and X_i is the ith participant's score on the predictor variable.[4] Here b_1 is the gradient of the straight line fitted to the data and b_0 is the intercept of that line. These parameters b_1 and b_0 are known as the *regression coefficients* and will crop up time and time again in this book, where you may see them referred to generally as b (without any subscript) or $\boldsymbol{b}_i$ (meaning the b associated with variable i). There is a residual term, ε_i, which represents the difference between the score predicted by the line for participant i and the score that participant i actually obtained. The equation is often conceptualized without this residual term (so ignore it if it's upsetting you); however, it is worth knowing that this term represents the fact that our model will not fit the data collected perfectly:

$$Y_i = (b_0 + b_1 X_i) + \varepsilon_i \tag{7.2}$$

A particular line has a specific intercept and gradient. Figure 7.2 shows a set of lines that have the same intercept but different gradients, and a set of lines that have the same gradient but different intercepts. Figure 7.2 also illustrates another useful point: the gradient of the line tells us something about the nature of the relationship being described. In Chapter 6 we saw how relationships can be either positive or negative (and I don't mean the difference between getting on well with your girlfriend and arguing all the time!). A line that has a gradient with a positive value describes a positive relationship, whereas a line with a negative gradient describes a negative relationship. So, if you look at the graph in Figure 7.2

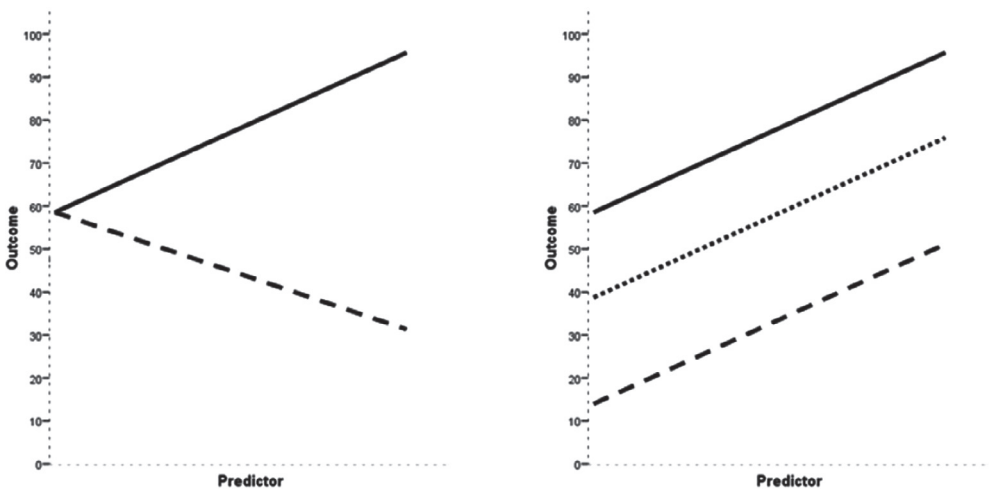


FIGURE 7.2 Lines with the same gradients but different intercepts, and lines that share the same intercept but have different gradients

[4] You'll sometimes see this equation written as:

$$Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_i$$

The only difference is that this equation has got βs in it instead of bs and in fact both versions are the same thing, they just use different letters to represent the coefficients.

in which the gradients differ but the intercepts are the same, then the dashed line describes a positive relationship whereas the solid line describes a negative relationship. Basically, then, the gradient (b_1) tells us what the model looks like (its shape) and the intercept (b_0) tells us where the model is (its location in geometric space).

If it is possible to describe a line knowing only the gradient and the intercept of that line, then we can use these values to describe our model (because in linear regression the model we use is a straight line). So, the model that we fit to our data in linear regression can be conceptualized as a straight line that can be described mathematically by equation (7.2). With regression we strive to find the line that best describes the data collected, then estimate the gradient and intercept of that line. Having defined these values, we can insert different values of our predictor variable into the model to estimate the value of the outcome variable.

7.2.2. The method of least squares ①

I have already mentioned that the method of least squares is a way of finding the line that best fits the data (i.e. finding a line that goes through, or as close to, as many of the data points as possible). This 'line of best fit' is found by ascertaining which line, of all of the possible lines that could be drawn, results in the least amount of difference between the observed data points and the line. Figure 7.3 shows that when any line is fitted to a set of data, there will be small differences between the values predicted by the line and the data that were actually observed.

Back in Chapter 2 we saw that we could assess the fit of a model (the example we used was the mean) by looking at the deviations between the model and the actual data collected. These deviations were the vertical distances between what the model predicted and each data point that was actually observed. We can do exactly the same to assess the fit of a regression line (which, like the mean, is a statistical model). So, again we are interested in the vertical differences between the line and the actual data because the line is our model: we use it to predict values of *Y* from values of the *X* variable. In regression these differences are usually called *residuals* rather than deviations, but they are the same thing. As with the mean, data points fall both above (the model underestimates their value) and

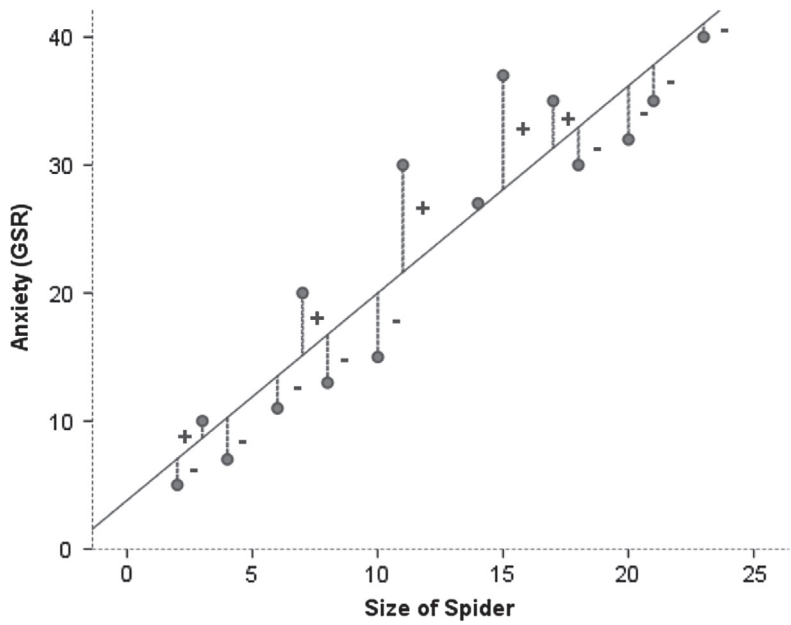


FIGURE 7.3 This graph shows a scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

below (the model overestimates their value) the line, yielding both positive and negative differences. In the discussion of variance in section 2.4.2 I explained that if we sum positive and negative differences then they cancel each other out and that to circumvent this problem we square the differences before adding them up. We do the same thing here. The resulting squared differences provide a gauge of how well a particular line fits the data: if the squared differences are large, the line is not representative of the data; if the squared differences are small, the line is representative.

You could, if you were particularly bored, calculate the sum of squared differences (or SS for short) for every possible line that is fitted to your data and then compare these 'goodness-of-fit' measures. The one with the lowest SS is the line of best fit. Fortunately we don't have to do this because the method of least squares does it for us: it selects the line that has the lowest sum of squared differences (i.e. the line that best represents the observed data). How exactly it does this is by using a mathematical technique for finding maxima and minima and this technique is used to find the line that minimizes the sum of squared differences. I don't really know much more about it than that, to be honest, so I tend to think of the process as a little bearded wizard called Nephwick the Line Finder who just magically finds lines of best fit. Yes, he lives inside your computer. The end result is that Nephwick estimates the value of the slope and intercept of the 'line of best fit' for you. We tend to call this line of best fit a *regression line*.

7.2.3. Assessing the goodness of fit: sums of squares, R and R^2 ①

Once Nephwick the Line Finder has found the line of best fit it is important that we assess how well this line fits the actual data (we assess the **goodness of fit** of the model). We do this because even though this line is the best one available, it can still be a lousy fit to the data! In section 2.4.2 we saw that one measure of the adequacy of a model is the sum of squared differences (or more generally we assess models using equation (7.3) below). If we want to assess the line of best fit, we need to compare it against something, and the thing we choose is the most basic model we can find. So we use equation (7.3) to calculate the fit of the most basic model, and then the fit of the best model (the line of best fit), and basically if the best model is any good then it should fit the data significantly better than our basic model:

$$\text{deviation} = \sum(\text{observed} - \text{model})^2 \tag{7.3}$$

This is all quite abstract so let's look at an example. Imagine that I was interested in predicting record sales (Y) from the amount of money spent advertising that record (X). One day my boss came into my office and said 'Andy, I know you wanted to be a rock star and you've ended up working as my stats-monkey, but how many records will we sell if we spend £100,000 on advertising?' If I didn't have an accurate model of the relationship between record sales and advertising, what would my best guess be? Well, probably the best answer I could give would be the mean number of record sales (say, 200,000) because on average that's how many records we expect to sell. This response might well satisfy a brainless record company executive (who didn't offer my band a record contract). However, what if he had asked 'How many records will we sell if we spend £1 on advertising?' Again, in the absence of any accurate information, my best guess would be to give the average number of sales (200,000). There is a problem: whatever amount of money is spent on advertising I

always predict the same level of sales. As such, the mean is a model of 'no relationship' at all between the variables. It should be pretty clear then that the mean is fairly useless as a model of a relationship between two variables – but it is the simplest model available.

So, as a basic strategy for predicting the outcome, we might choose to use the mean, because on average it will be a fairly good guess of an outcome, but that's all. Using the mean as a model, we can calculate the difference between the observed values, and the values predicted by the mean (equation (7.3)). We saw in section 2.4.1 that we square all of these differences to give us the sum of squared differences. This sum of squared differences is known as the **total sum of squares** (denoted SS_T) because it is the total amount of differences present when the most basic model is applied to the data. This value represents how good the mean is as a model of the observed data. Now, if we fit the more sophisticated model to the data, such as a line of best fit, we can again work out the differences between this new model and the observed data (again using equation (7.3)). In the previous section we saw that the method of least squares finds the best possible line to describe a set of data by minimizing the difference between the model fitted to the data and the data themselves. However, even with this optimal model there is still some inaccuracy, which is represented by the differences between each observed data point and the value predicted by the regression line. As before, these differences are squared before they are added up so that the directions of the differences do not cancel out. The result is known as the **sum of squared residuals** or **residual sum of squares** (SS_R). This value represents the degree of inaccuracy when the best model is fitted to the data. We can use these two values to calculate how much better the regression line (the line of best fit) is than just using the mean as a model (i.e. how much better is the best possible model than the worst model?). The improvement in prediction resulting from using the regression model rather than the mean is calculated by calculating the difference between SS_T and SS_R. This difference shows us the reduction in the inaccuracy of the model resulting from fitting the regression model to the data. This improvement is the *model sum of squares* (SS_M). Figure 7.4 shows each sum of squares graphically.

If the value of SS_M is large then the regression model is very different from using the mean to predict the outcome variable. This implies that the regression model has made a big improvement to how well the outcome variable can be predicted. However, if SS_M is small then using the regression model is little better than using the mean (i.e. the regression model is no better than taking our 'best guess'). A useful measure arising from these sums of squares is the proportion of improvement due to the model. This is easily calculated by dividing the sum of squares for the model by the total sum of squares. The resulting value is called R^2 and to express this value as a percentage you should multiply it by 100. R^2 represents the amount of variance in the outcome explained by the model (SS_M) relative to how much variation there was to explain in the first place (SS_T). Therefore, as a percentage, it represents the percentage of the variation in the outcome that can be explained by the model:

$$R^2 = \frac{SS_M}{SS_T} \tag{7.4}$$

This R^2 is the same as the one we met in Chapter 6 (section 6.5.2.3) and you might have noticed that it is interpreted in the same way. Therefore, in simple regression we can take the square root of this value to obtain Pearson's correlation coefficient. As such, the correlation coefficient provides us with a good estimate of the overall fit of the regression model, and R^2 provides us with a good gauge of the substantive size of the relationship.

A second use of the sums of squares in assessing the model is through the *F*-test. I mentioned way back in Chapter 2 that test statistics (like *F*) are usually the amount of systematic variance divided by the amount of unsystematic variance, or, put another way, the model compared against the error in the model. This is true here: *F* is based upon the ratio of the improvement due to the model (SS_M) and the difference between the model and the observed data (SS_R). Actually, because the sums of squares depend on the number

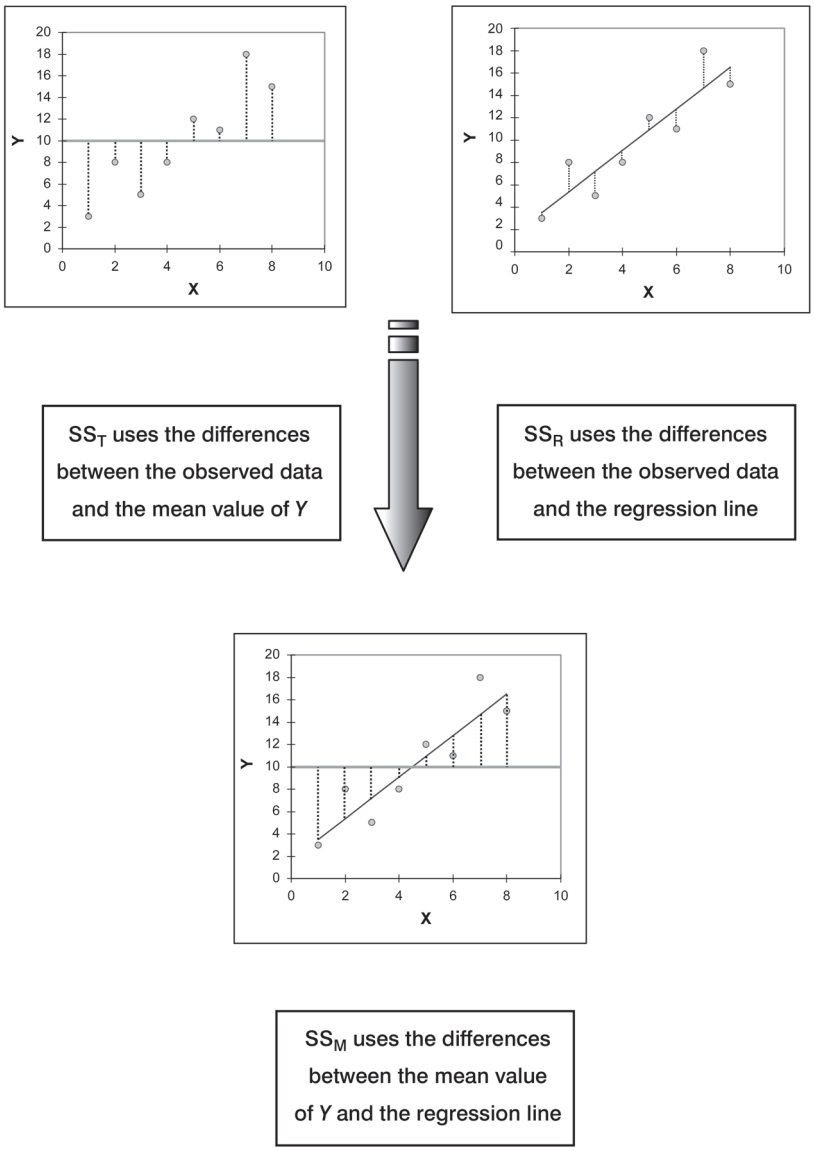


FIGURE 7.4
Diagram showing from where the regression sums of squares derive

of differences that we have added up, we use the average sums of squares (referred to as the **mean squares** or MS). To work out the mean sums of squares we divide by the degrees of freedom (this is comparable to calculating the variance from the sums of squares – see section 2.4.2). For SS_M the degrees of freedom are simply the number of variables in the model, and for SS_R they are the number of observations minus the number of parameters being estimated (i.e. the number of beta coefficients including the constant). The result is the mean squares for the model (MS_M) and the residual mean squares (MS_R). At this stage it isn't essential that you understand how the mean squares are derived (it is explained in Chapter 10). However, it is important that you understand that the ***F*-ratio** (equation (7.5)) is a measure of how much the model has improved the prediction of the outcome compared to the level of inaccuracy of the model:

$$F = \frac{MS_M}{MS_R} \tag{7.5}$$

If a model is good, then we expect the improvement in prediction due to the model to be large (so MS_M will be large) and the difference between the model and the observed data to be small (so MS_R will be small). In short, a good model should have a large *F*-ratio (greater than 1 at least) because the top of equation (7.5) will be bigger than the bottom. The exact magnitude of this *F*-ratio can be assessed using critical values for the corresponding degrees of freedom (as in the Appendix).

7.2.4. Assessing individual predictors ①

We've seen that the predictor in a regression model has a coefficient (b_1), which in simple regression represents the gradient of the regression line. The value of *b* represents the change in the outcome resulting from a unit change in the predictor. If the model was useless at predicting the outcome, then if the value of the predictor changes, what might we expect the change in the outcome to be? Well, if the model is very bad then we would expect the change in the outcome to be zero. Think back to Figure 7.4 (see the panel representing SS_T) in which we saw that using the mean was a very bad way of predicting the outcome. In fact, the line representing the mean is flat, which means that as the predictor variable changes, the value of the outcome does *not* change (because for each level of the predictor variable, we predict that the outcome will equal the mean value). The important point here is that a bad model (such as the mean) will have regression coefficients of 0 for the predictors. A regression coefficient of 0 means: (1) a unit change in the predictor variable results in no change in the predicted value of the outcome (the predicted value of the outcome does not change at all); and (2) the gradient of the regression line is 0, meaning that the regression line is flat. Hopefully, you'll see that it logically follows that if a variable significantly predicts an outcome, then it should have a *b*-value significantly different from zero. This hypothesis is tested using a *t*-test (see Chapter 9). The **t-statistic** tests the null hypothesis that the value of *b* is 0: therefore, if it is significant we gain confidence in the hypothesis that the *b*-value is significantly different from 0 and that the predictor variable contributes significantly to our ability to estimate values of the outcome.

Like *F*, the *t*-statistic is also based on the ratio of explained variance to unexplained variance or error. Well, actually, what we're interested in here is not so much variance but whether the *b* we have is big compared to the amount of error in that estimate. To estimate how much error we could expect to find in *b* we use the standard error. The standard error tells us something about how different *b*-values would be across different samples. We could take lots and lots of samples of data regarding record sales and advertising budgets and calculate the *b*-values for each sample. We could plot a frequency distribution of these samples to discover whether the *b*-values from all samples would be relatively similar, or whether they would be very different (think back to section 2.5.1). We can use the standard deviation of this distribution (known as the *standard error*) as a measure of the similarity of *b*-values across samples. If the standard error is very small, then it means that most samples are likely to have a *b*-value similar to the one in our sample (because there is little variation across samples). The *t*-test tells us whether the *b*-value is different from 0 relative to the variation in *b*-values across samples. When the standard error is small even a small deviation from zero can reflect a meaningful difference because *b* is representative of the majority of possible samples.

Equation (7.6) shows how the *t*-test is calculated and you'll find a general version of this equation in Chapter 9 (equation (9.1)). The $b_{expected}$ is simply the value of *b* that we would expect to obtain if the null hypothesis were true. I mentioned earlier that the null

hypothesis is that b is 0 and so this value can be replaced by 0. The equation simplifies to become the observed value of b divided by the standard error with which it is associated:

$$t = \frac{b_{\text{observed}} - b_{\text{expected}}}{\text{SE}_b} = \frac{b_{\text{observed}}}{\text{SE}_b} \tag{7.6}$$

The values of t have a special distribution that differs according to the degrees of freedom for the test. In regression, the degrees of freedom are $N - p - 1$, where N is the total sample size and p is the number of predictors. In simple regression when we have only one predictor, this reduces down to $N - 2$. Having established which t-distribution needs to be used, the observed value of t can then be compared to the values that we would expect to find if there was no effect (i.e. $b = 0$): if t is very large then it is unlikely to have occurred when there is no effect (these values can be found in the Appendix). SAS provides the exact probability that the observed value (or a larger one) of t would occur if the value of b was, in fact, 0. As a general rule, if this observed significance is less than .05, then scientists assume that b is significantly different from 0; put another way, the predictor makes a significant contribution to predicting the outcome.

7.3. Doing simple regression on SAS ①

So far, we have seen a little of the theory behind regression, albeit restricted to the situation in which there is only one predictor. To help clarify what we have learnt so far, we will go through an example of a simple regression on SAS. Earlier on I asked you to imagine that I worked for a record company and that my boss was interested in predicting record sales from advertising. There are some data for this example in the file **Record1.sas7bdat**. This data file has 200 rows, each one representing a different record. There are also two columns, one representing the sales of each record in the week after release and the other representing the amount (in pounds) spent promoting the record before release. This is the format for entering regression data: the outcome variable and any predictors should be entered in different columns, and each row should represent independent values of those variables.

The pattern of the data is shown in Figure 7.5 and it should be clear that a positive relationship exists: so, the more money spent advertising the record, the more it is likely to sell. Of course there are some records that sell well regardless of advertising (top left of scatterplot), but there are none that sell badly when advertising levels are high (bottom right of scatterplot). The scatterplot also shows the line of best fit for these data: bearing in mind that the mean would be represented by a flat line at around the 200,000 sales mark, the regression line is noticeably different.

To find out the parameters that describe the regression line, and to see whether this line is a useful model, we need to run a regression analysis. To do the analysis you need to use PROC REG.

PROC REG syntax is very straightforward, and is shown in SAS Syntax 7.1. The MODEL statement is written in the form of the equation: we think that sales are a function of adverts, so we write sales = adverts. Notice that PROC REG is slightly different to other procedures, because we need to write QUIT; after the RUN statement.

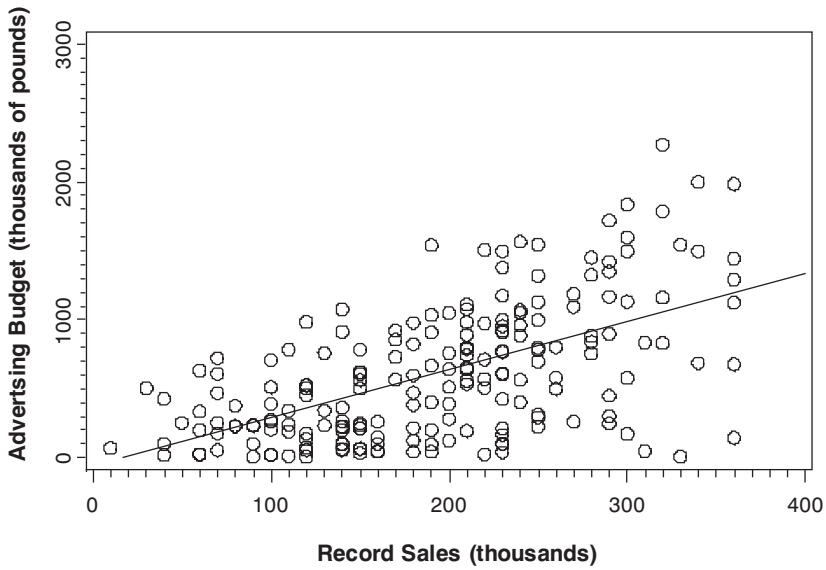


FIGURE 7.5 Scatterplot showing the relationship between record sales and the amount spent promoting the record

```
PROC REG DATA=chapter6.record1;
    MODEL sales = adverts;
    RUN;
QUIT;
```

SAS Syntax 7.1

7.4. Interpreting a simple regression ①

7.4.1. Overall fit of the model ①

The output from the regression is shown in SAS Output 7.1. The first part of the output reports an analysis of variance (ANOVA – see Chapter 10). The summary table shows the various sums of squares described in Figure 7.4 and the degrees of freedom associated with each. From these two values, the average sums of squares (the mean squares) can be calculated by dividing the sums of squares by the associated degrees of freedom. The most important part of the table is the *F*-ratio, which is calculated using equation (7.5), and the associated significance value of that *F*-ratio. For these data, *F* is 99.59, which is significant at $p < .001$ (because the value in the column labelled Sig. is less than .001). Researchers usually dont report p-values below 0.001 though. This result tells us that there is less than a 0.1% chance that an *F*-ratio at least this large would happen if the null hypothesis were true. Therefore, we can conclude that our regression model results in significantly better prediction of record sales than if we used the mean value of record sales. In short, the regression model overall predicts record sales significantly well.

The second part of the output is a summary of the model. This summary table provides the value of *R* and R^2 for the model that has been derived (as well as some other things we are not going to worry about for now). For these data, *R* has a value of .578 and because there is only one predictor, this value represents the simple correlation between advertising

SAS OUTPUT 7.1

The REG Procedure
Model: MODEL1
Dependent Variable: SALES Record Sales (thousands)

Number of Observations Read	200
Number of Observations Used	200

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	433688	433688	99.59	<.0001
Error	198	862264	4354.86953		
Corrected Total	199	1295952			

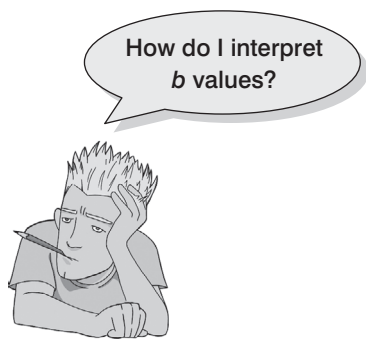
Root MSE	65.99144	R-Square	0.3346
Dependent Mean	193.20000	Adj R-Sq	0.3313
Coeff Var	34.15706		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	134.13994	7.53657	17.80	<.0001
ADVERTS	Advertsing Budget (thousands of pounds)	1	0.09612	0.00963	9.98	<.0001

and record sales (you can confirm this by running a correlation using what you were taught in Chapter 6). The value of R^2 is .335, which tells us that advertising expenditure can account for 33.5% of the variation in record sales. In other words, if we are trying to explain why some records sell more than others, we can look at the variation in sales of different records. There might be many factors that can explain this variation, but our model, which includes only advertising expenditure, can explain approximately 33% of it. This means that 67% of the variation in record sales cannot be explained by advertising alone. Therefore, there must be other variables that have an influence also.

7.4.2. Model parameters ①

The ANOVA tells us whether the model, overall, results in a significantly good degree of prediction of the outcome variable. However, the ANOVA doesn't tell us about the individual contribution of variables in the model (although in this simple case there is only one variable in the model and so we can infer that this variable is a good predictor. The third part of the output provides details of the model parameter estimates (the beta values) and the significance of these values. We saw in equation (7.2) that b_0 was the *Y* intercept and this value is the value labelled *Parameter Estimate* (in the SAS output) for the constant. So, from the table, we



can say that b_0 is 134.14, and this can be interpreted as meaning that when no money is spent on advertising (when $X = 0$), the model predicts that 134,140 records will be sold (remember that our unit of measurement was thousands of records). We can also read off the value of b_1 from the table and this value represents the gradient of the regression line. It is 0.096. Although this value is the slope of the regression line, it is more useful to think of this value as representing *the change in the outcome associated with a unit change in the predictor*. Therefore, if our predictor variable is increased by one unit (if the advertising budget is increased by 1), then our model predicts that 0.096 extra records will be sold. Our units of measurement were thousands of pounds and thousands of records sold, so we can say that for an increase in advertising of £1000 the model predicts 96 ($0.096 \times 1000 = 96$) extra record sales. As you might imagine, this investment is pretty bad for the record company: it invests £1000 and gets only 96 extra sales! Fortunately, as we already know, advertising accounts for only one-third of record sales.

We saw earlier that, in general, values of the regression coefficient b represent the change in the outcome resulting from a unit change in the predictor and that if a predictor is having a significant impact on our ability to predict the outcome then this b should be different from 0 (and big relative to its standard error). We also saw that the t-test tells us whether the b-value is different from 0. SAS provides the exact probability that the observed value of t would occur if the value of b in the population were 0. If this observed significance is less than .05, then scientists agree that the result reflects a genuine effect (see Chapter 2). For these two values, the probabilities are <.0001 and so we can say that the probability of these t values (or larger) occurring if the values of b in the population were 0 is less than .0001. Therefore, the bs are different from 0 and we can conclude that the advertising budget makes a significant contribution ($p < .0001$) to predicting record sales.

SELF-TEST How is the t in SAS Output 7.1 calculated? Use the values in the table to see if you can get the same value as SAS.

7.4.3. Using the model ①

So far, we have discovered that we have a useful model, one that significantly improves our ability to predict record sales. However, the next stage is often to use that model to make some predictions. The first stage is to define the model by replacing the b-values in equation (7.2) with the values from SAS Output 7.1. In addition, we can replace the X and Y with the variable names so that the model becomes:

$$\begin{aligned}\text{record sales}_i &= b_0 + b_1\text{advertising budget}_i \\ &= 134.14 + (0.096 \times \text{advertising budget}_i)\end{aligned} \tag{7.7}$$

It is now possible to make a prediction about record sales, by replacing the advertising budget with a value of interest. For example, imagine a record executive wanted to spend £100,000 on advertising a new record. Remembering that our units are already in thousands of pounds; we can simply replace the advertising budget with 100. He would discover that record sales should be around 144,000 for the first week of sales:

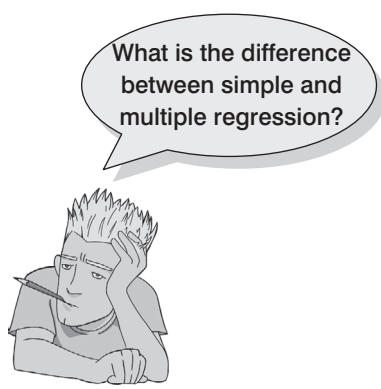
$$\begin{aligned}\text{record sales}_i &= 134.14 + (0.096 \times \text{advertising budget}_i)\\ &= 134.14 + (0.096 \times 100)\\ &= 143.74\end{aligned} \tag{7.8}$$

SELF-TEST How many records would be sold if we spent £666,000 on advertising the latest CD by black metal band Abgott?

CRAMMING SAM'S TIPS Simple regression

- Simple regression is a way of predicting values of one variable from another.
- We do this by fitting a statistical model to the data in the form of a straight line.
- This line is the line that best summarizes the pattern of the data.
- We have to assess how well the line fits the data using:
  - R^2 which tells us how much variance is explained by the model compared to how much variance there is to explain in the first place. It is the proportion of variance in the outcome variable that is shared by the predictor variable.
  - F, which tells us how much variability the model can explain relative to how much it can't explain (i.e. it's the ratio of how good the model is compared to how bad it is).
- The b-value tells us the gradient of the regression line and the strength of the relationship between a predictor and the outcome variable. If it is significant (*Sig.* < .05 in the SAS table) then the predictor variable significantly predicts the outcome variable.

7.5. Multiple regression: the basics ①



To summarize what we have learnt so far, in simple linear regression the outcome variable Y is predicted using the equation of a straight line (equation (7.2)). Given that we have collected several values of Y and X, the unknown parameters in the equation can be calculated. They are calculated by fitting a model to the data (in this case a straight line) for which the sum of the squared differences between the line and the actual data points is minimized. This method is called the method of least squares. Multiple regression is a logical extension of these principles to situations in which there are several predictors. Again, we still use our basic equation of:

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

but this time the model is slightly more complex. It is basically the same as for simple regression except that for every extra predictor you include, you have to add a coefficient;