# Regression

# 8

**FIGURE 8.1**
Me playing with my ding-a-ling in the Holimarine Talent Show. Note the groupies queuing up at the front

## 8.1. 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 any more 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 from certain variables: the age of the performer, what type of performance they gave (singing, telling a joke, magic tricks), and perhaps how cute they looked. A regression analysis on these data would enable us to predict the future (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 a song for me to sing about the keyboard player in the Holimarine Band ‘messing about with his organ’. He said ‘take this song, son, and steal the show’ … and that’s what I did: I came first again. There’s no accounting for taste.

ODITI’S LANTERN

*Words that go unspoken, deeds that go undone*

‘I, Oditi, do not want my followers to get distracted by playing with their ding-a-lings. To warn you all of the dangers of such frivolity, I have uncovered a song, sung by an innocent child, that explains the risks. Stare into my lantern and shake your booty to the funky tune.’

## 8.2. An introduction to regression ①

### 8.2.1. The simple linear model ①

In the previous chapter we started getting down to the nitty-gritty of the linear model that we’ve been discussing since way back in Chapter 2. We saw that if we wanted to look at the relationship between two variables we could use the model in equation (2.3):

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

In this model, $b$ is the correlation coefficient (more often denoted as $r$) and it is a standardized measure. However, we can also work with an unstandardized version of $b$, but in doing so we need to add something to the model:

$$\begin{aligned}\text{outcome}_i &= (b_0 + b_1X_i) + \text{error}_i\\ y_i &= (b_0 + b_1X_i) + \varepsilon_i \qquad (8.1)\end{aligned}$$

The important thing to note is that this equation keeps the fundamental idea that an outcome for a person can be predicted from a model (the stuff in brackets) and some error associated with that prediction ($\varepsilon_i$). We are still predicting an outcome variable ($y_i$) from a predictor variable ($X_i$) and a parameter, $b_1$, associated with the predictor variable that quantifies the relationship it has with the outcome variable. This model differs from that of a correlation only in that it uses an *unstandardized* measure of the relationship ($b$) and consequently we need to include a parameter that tells us the value of the outcome when the predictor is zero.[3] This parameter is $b_0$.

Focus on the model itself for a minute. Does it seem familiar? Let's imagine that instead of $b_0$ we use the letter $c$, and instead of $b_1$ we use the letter $m$. Let's also ignore the error term for the moment. We could predict our outcome as follows:

$$\text{outcome}_i = mx + c$$

Or if you're American, Canadian or Australian let's use the letter $b$ instead of $c$:

$$\text{outcome}_i = mx + b$$

Perhaps you're French, Dutch or Brazilian, in which case let's use $a$ instead of $m$:

$$\text{outcome}_i = ax + b$$

Do any of these look familiar to you? If not, there are two explanations: (1) you didn't pay enough attention at school, or (2) you're Latvian, Greek, Italian, Swedish, Romanian, Finnish or Russian – to avoid this section being even more tedious, I used only the three main international differences in the equation above. The different forms of the equation make an important point: the symbols or letters we use in an equation don't necessarily change it.[4] Whether we write $mx + c$ or $b_1X + b_0$ doesn't really matter, what matters is what the symbols represent. So, what do the symbols represent?

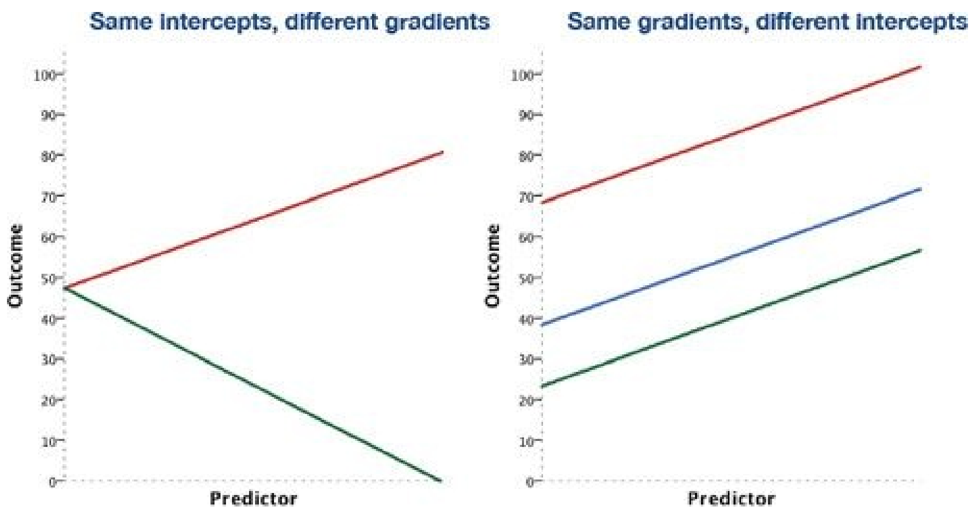
Hopefully, some of you recognized this model as 'the equation of a straight line'. I have talked throughout this book about fitting 'linear models', and linear simply means 'straight line'. So, it should come as no surprise that the equation we use is the one that describes a straight line. Any straight 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, $b_0$). 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 ***bn*** (meaning the $b$ associated with variable $n$). A particular line (i.e., model) will have a specific intercept and gradient.

Figure 8.2 shows a set of lines that have the same intercept but different gradients. For these three models, $b_0$ will be the same in each but the values of $b_1$ will differ in each model.

Figure 8.2 also shows models that have the same gradients ($b_1$ is the same in each model) but different intercepts (the $b_0$ is different in each model). I've mentioned already that $b_1$ quantifies the relationship between the predictor variable and the outcome, and Figure 8.2 illustrates this point. In Chapter 6 we saw how relationships can be either positive or negative (and I don't mean whether or not you and your partner argue all the time). A model with a positive $b_1$ describes a positive relationship, whereas a line with a negative $b_1$ describes a negative relationship. Looking at Figure 8.2 (left), the red line describes a positive relationship whereas the green line describes a negative relationship. As such, we can use a linear model (i.e., a straight line) to summarize the relationship between two variables: 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).

**FIGURE 8.2**

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



This is all quite abstract, so let’s look at an example. Imagine that I was interested in predicting physical and downloaded album sales (outcome) from the amount of money spent advertising that album (predictor). We could summarize this relationship using a linear model by replacing the names of our variables into equation (8.1):

$$\begin{aligned} y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \text{album sales}_i &= b_0 + b_1 \text{advertising budget}_i + \varepsilon_i \end{aligned} \tag{8.2}$$

Once we have estimated the values of the *b*s we would be able to make a prediction about album sales by replacing ‘advertising’ with a number representing how much we wanted to spend advertising an album. For example, imagine that $b_0$ turned out to be 50 and $b_1$ turned out to be 100. Our model would be:

$$\text{album sales}_i = 50 + (100 \times \text{advertising budget}_i) + \varepsilon_i \tag{8.3}$$

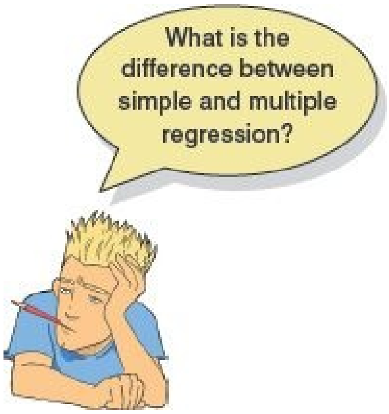
Note that I have replaced the betas with their numeric values. Now, we can make a prediction. Imagine we wanted to spend £5 on advertising, we can replace the variable ‘advertising budget’ with this value and solve the equation to discover how many album sales we will get:

$$\begin{aligned} \text{album sales}_i &= 50 + (100 \times 5) + \varepsilon_i \\ &= 550 + \varepsilon_i \end{aligned}$$

So, based on our model we can predict that if we spend £5 on advertising, we’ll sell 550 albums. I’ve left the error term in there to remind you that this prediction will probably not be perfectly accurate. This value of 550 album sales is known as a **predicted value**.

### 8.2.2. The linear model with several predictors ②

We have seen that we can use a straight line to ‘model’ the relationship between two variables. However, life is usually more complicated than that: there are often numerous variables that might be related to the outcome of interest. To take our album sales example, we might expect variables other than simply advertising to have an effect. For example, how much someone hears songs from the album on the radio, or the 'look' of the band might have an influence. One of the beautiful things about the linear model is that it can be expanded to include as many predictors as you like. We hinted at this back in Chapter 2 (equation (2.4)). To add a predictor all we need to do is place it into the model and give it a *b* that estimates the relationship between that predictor and the outcome. For example, if we wanted to add the number of plays of the band on the radio per week (airplay), we could add this second predictor in general as:


What is the difference between simple and multiple regression?

$$Y_i = (b_0 + b_1X_{1i} + b_2X_{2i}) + \varepsilon_i \tag{8.4}$$

Note that all that has changed is the addition of a second predictor ($X_2$) and an associated parameter ($b_2$). To make things more concrete, let's use the variable names instead:

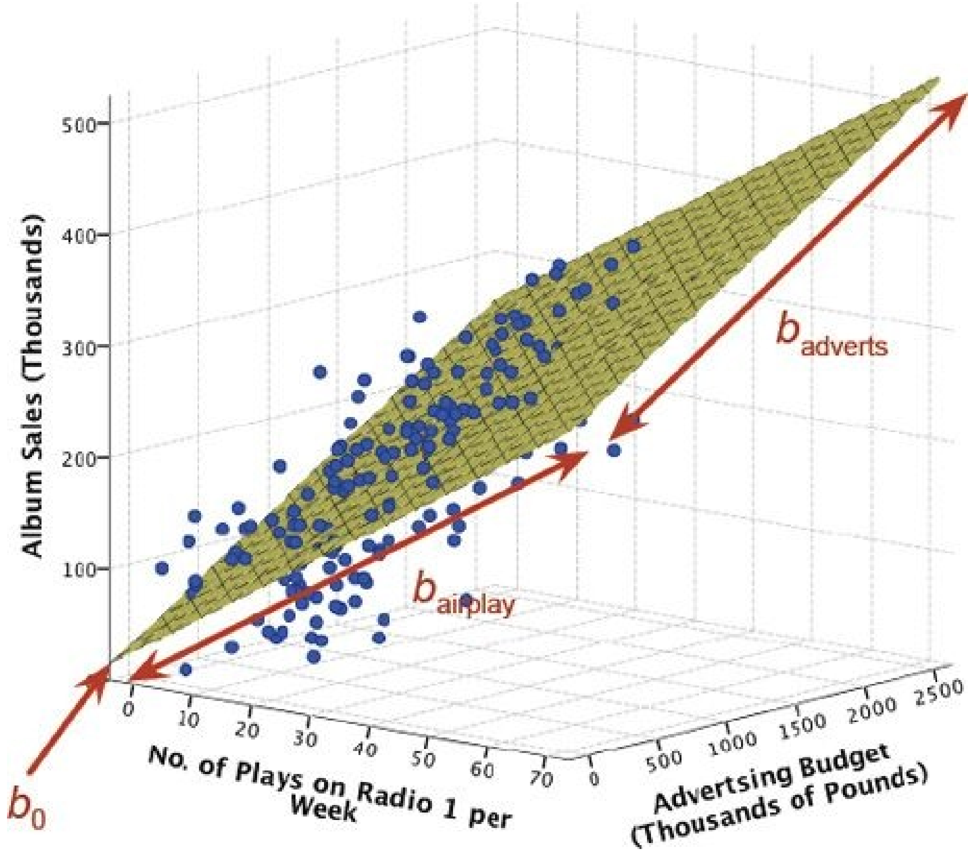
$$\text{album sales}_i = b_0 + b_1\text{advertising budget}_i + b_2\text{airplay}_i + \varepsilon_i \tag{8.5}$$

The new model includes a *b*-value for both predictors (and, of course, the constant, $b_0$). If we estimate the *b*-values, we could make predictions about album sales based not only on the amount spent on advertising but also in terms of radio play. There are only two predictors in this model and so we could display this model graphically in three dimensions (Figure 8.3).

The tinted trapezium in the diagram (known as the regression *plane*) is described by equation (8.5) and the dots represent the observed data points. Like a regression line, a regression plane aims to give the best prediction for the observed data. However, there are invariably some differences between the model and the real-life data (this fact is evident because some of the dots do not lie exactly on the tinted area of the graph). The vertical distances between the regression plane and each data point are the errors or *residuals* in the model. The *b*-value for advertising describes the slope of the left and right sides of the regression plane, whereas the *b*-value for airplay describes the slope of the top and bottom of the regression plane. Just like simple regression, knowledge of these two slopes tells us about the shape of the model (what it looks like) and the intercept locates the regression plane in space.

**FIGURE 8.3**

Scatterplot of the relationship between album sales, advertising budget and radio play



It is fairly easy to visualize a regression model with two predictors, because it is possible to plot the regression plane using a 3-D scatterplot. However, multiple regression can be used with three, four or even ten or more predictors. Although you can't immediately visualize what such complex models look like, or visualize what the *b*-values represent, you should be able to apply the principles of these basic models to more complex scenarios. In fact, in general we can add as many predictors as we like, and the linear model will expand accordingly:

$$Y_i = (b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni}) + \varepsilon_i \qquad (8.6)$$

in which *Y* is the outcome variable, $b_1$ is the coefficient of the first predictor ($X_1$), $b_2$ is the coefficient of the second predictor ($X_2$), $b_n$ is the coefficient of the *n*th predictor ($X_{ni}$), and $\varepsilon_i$ is the error for the *i*th participant. (The parentheses aren't necessary, they're just there to make the connection to equation (8.1)). This equation illustrates that we can add in as many predictors as we like until we reach the final one ($X_n$), but each time we do, we assign it a regression coefficient (*b*).

To sum up, regression analysis is when we fit a linear model to our data and use it to predict values of an **outcome variable** (a.k.a. dependent variable) from one or more **predictor variables** (a.k.a. independent variables). With one predictor variable, the technique is sometimes referred to as **simple regression**, but when there are several predictors in the model we call it **multiple regression**. This tool is incredibly useful because it enables us to go a step beyond the data that we collected.

### 8.2.3. Estimating the model ②

We have seen that the linear model is a versatile model for summarizing the relationship between one or more predictor variables and an outcome variable. No matter how many predictors we have, the model can be described entirely by a constant ($b_0$) and by parameters associated with each predictor (*b*s). You might wonder how we estimate these parameters, and the quick answer is that we typically use the method of least squares that was described in Section 2.4.3. We saw then 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).

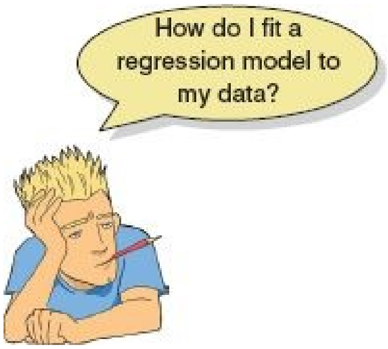
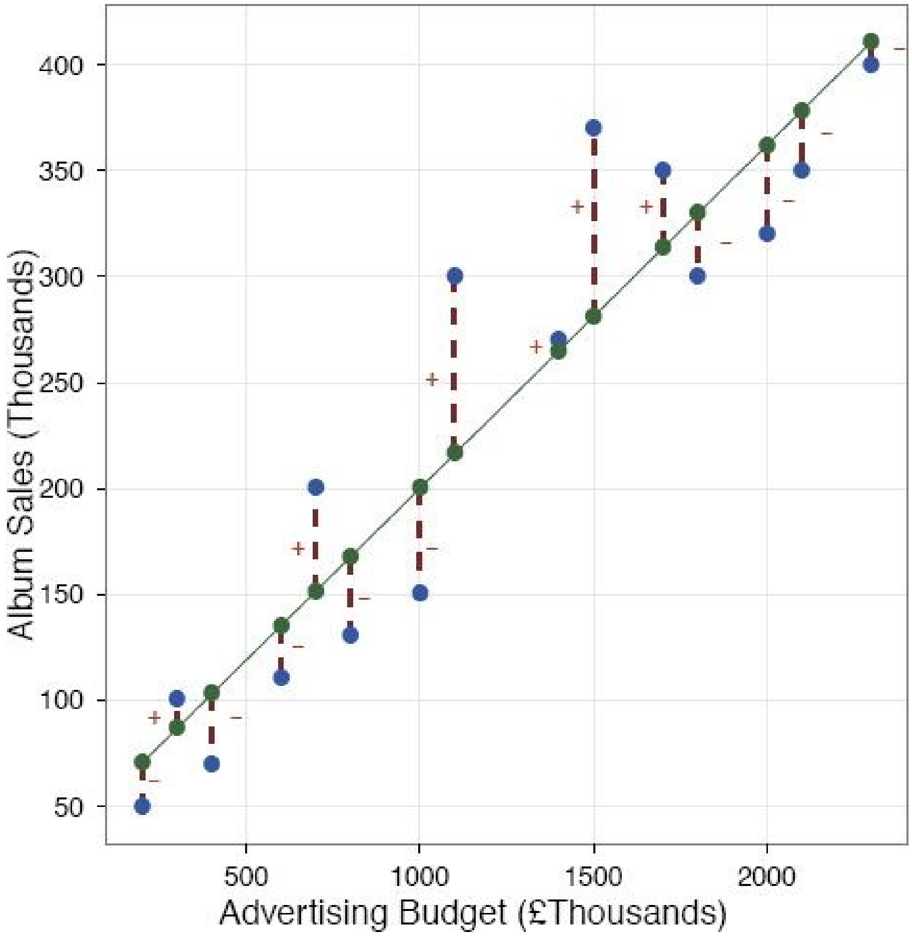


Figure 8.4 shows some data about advertising budget and album sales. A model has been fitted to these data (the straight line). The blue circles are the observed data. The line is the model. The green dots on the line are the predicted values. We saw earlier that predicted values are the values of the outcome variable calculated from the model. In other words, if we estimated the values of *b* that define the model and put these values into the linear model (as we did in equation (8.3)), then we insert different values for advertising budget, the predicted values are the resulting estimates of album sales. The question is what values of advertising budget to use to get these predicted values. One very useful thing to do is to use the values of the predictor that actually occurred in the data from which the model was estimated. If you think about it, this makes sense because if the model is a perfect fit of the data then for a given value of the predictor(s) the model should predict the same value of the outcome as was actually observed. In terms of Figure 8.4 this would mean that the green dots fall in exactly the same locations as the blue dots. As you can see, they don't, which shows that the model is not perfect (and it never will be): there is error in the predicted values – sometimes they overestimate the observed value of the outcome and sometimes they underestimate it. In regression, the differences between what the model predicts and the observed data are usually called **residuals** (they are the same as *deviations* when we looked at the mean) and they are the vertical dashed lines in Figure 8.4.

**FIGURE 8.4**

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



We saw in Chapter 2, equation (2.6), that if we want to calculate the total error in a model we do so by looking at the squared differences between the observed values of the outcome, and the predicted values that come from the model:

$$\text{Total error} = \sum_{i=1}^{n} \left(\text{observed}_i - \text{model}_i\right)^2 \qquad (8.7)$$

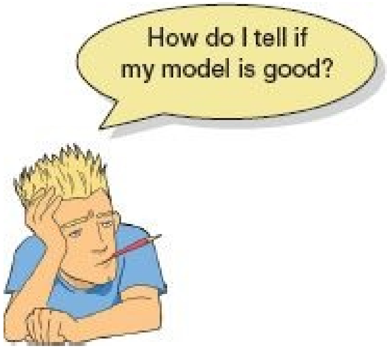
Sometimes the predicted value of the outcome is less than the actual value and sometimes it is greater, meaning that sometimes the residuals are positive and sometimes they are negative. If we add the residuals, the positive ones will cancel out the negative ones, so we square them before we add them up (this idea should be familiar from Section 2.4.2). Therefore, to assess the error in a regression model, just like when we assessed the fit of the mean using the variance, we use a sum of squared errors, and because in regression we call these errors residuals, we refer to this total as the *sum of squared residuals* or **residual sum of squares** ($SS_R$). The residual sum of squares is 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.

How do we find the optimal model to summarize our data? You could, if you were particularly bored, calculate the residual sum of squares for every possible line that could be fitted to your data and then compare these ‘goodness-of-fit’ measures. The one with the lowest $SS_R$ would be the best fitting model. However, we have better things to do, so just like when we estimate the mean, we can use the method of least squares to estimate the parameters (*b*) that define the model for which the sum of squared errors is the minimum it can be (given the data). This method is known as **ordinary least squares (OLS)** regression. How exactly the method of least squares does this is beyond me: it uses a mathematical technique for finding maxima and minima to find the *b*-values that describe the model that minimizes the sum of squared differences.

I don't really know much more about it than that, to be honest, so with one predictor 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. For more complex models, Nephwick invites his brother Clungglewad the Beta Seeker for tea and cake inside your computer and together they stare into the tea leaves in their cups until the optimal beta-values are revealed to them. Then they compare beard growth since their last meeting. In short, they use the method of least squares to estimate the values of *b* that describe the **regression model** that best fits the data.

## 8.2.4. Assessing the goodness of fit, sums of squares, R and $R^2$ ①

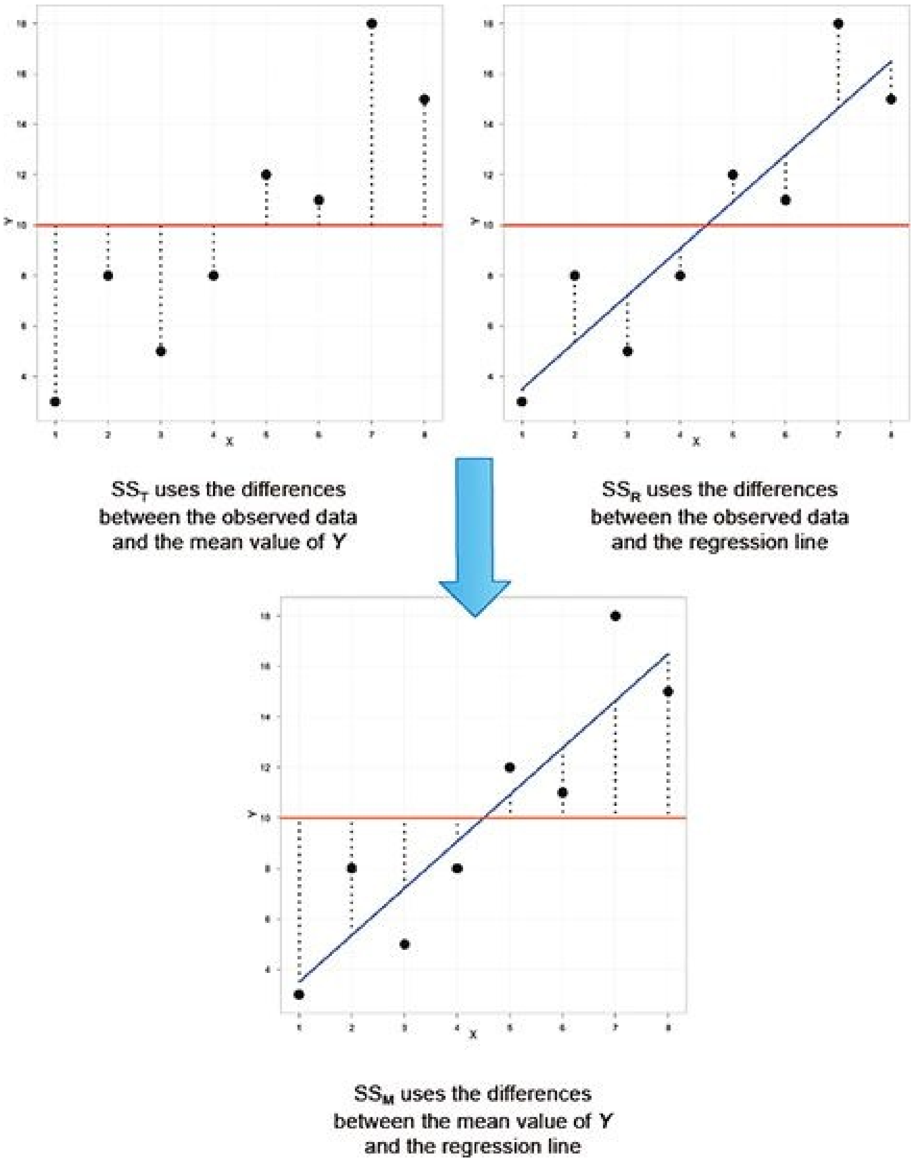
Once Nephwick and Clungglewad have found the model of best fit, it is important that we assess how well this model fits the actual data (we assess the goodness of fit of the model). We do this because even though the model is the best one available, it can still be a lousy fit to the data. We saw above that the residual sum of squares, $SS_R$, is a measure of how much error there is in the model: it gives us an idea of how much error there is in prediction, but it doesn't tell us whether using the model is better than nothing. It is not enough to simply assess the error within the model, we need to compare it against a baseline to see whether it 'improves' how well we can predict the outcome. So, we fit the most basic model we can, we use equation (8.7) to calculate the fit of this baseline model. Then we fit the best model, and also calculate the error, $SS_R$, within it using equation (8.7). Basically if the best model is any good then it should have significantly less error within it than our basic model.



This is all quite abstract, so let's go back to our example of predicting album sales (*Y*) from the amount of money spent advertising that album (*X*). One day my boss came in to 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 albums will we sell if we spend £100,000 on advertising?' If I didn't have an accurate model of the relationship between album sales and advertising, what would my best guess be? Probably the best answer I could give would be the mean number of album sales (say, 200,000) because on average that's how many albums we expect to sell. This response might well satisfy a brainless record company executive (who didn't offer my band a recording contract). However, what if he had asked: 'How many albums 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 levels 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. Using the mean as a model, we can calculate the difference between the observed values, and the values predicted by the mean (equation (8.7)). 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 a more sophisticated model to the data, such as a regression model, we can again work out the differences between this new model and the observed data (again using equation (8.7)). This value is the residual sum of squares ($SS_R$) discussed in the previous section. 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 model is than using a baseline model such as the mean (i.e., how much better the best possible model is 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 8.5 shows each sum of squares graphically for the example where the regression model is a line (i.e., one predictor) but the same principles apply with more than one predictor.

**FIGURE 8.5**
Diagram showing from where the regression sums of squares derive



$SS_T$ uses the differences between the observed data and the mean value of $Y$

$SS_R$ uses the differences between the observed data and the regression line

$SS_M$ uses the differences between the mean value of $Y$ and the regression line

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 to give a quantity called $R^2$:

$$R^2 = \frac{SS_M}{SS_T} \tag{8.8}$$

To express this value as a percentage you should multiply it by 100. This $R^2$ 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 ($SS_T$); it is the same as the $R^2$ we met in Chapter 7 (Section 7.4.2.2) and it is interpreted in the same way: as a percentage, it represents the percentage of the variation in the outcome that can be explained by the model. We can take the square root of this value to obtain Pearson's correlation coefficient for the relationship between the values of the outcome predicted by the model and the values of the outcome we actually observed.[5] As such, the correlation coefficient provides us with a good estimate of the overall fit of the regression model (i.e., the correspondence between predicted values of the outcome and the actual values), and $R^2$ provides us with a gauge of the substantive size of the model fit.[6]

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 to 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 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 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 11). However, it is important that you understand that the **F-ratio**,

$$F = \frac{MS_M}{MS_R} \qquad (8.9$$

is a measure of how much the model has improved the prediction of the outcome compared to the level of inaccuracy of the model. 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 (8.9) 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). The *F*-statistic can also be used to calculate the significance of $R^2$ using the following equation:

$$F = \frac{(N-k-1)R^2}{k\left(1-R^2\right)} \qquad (8.10)$$

in which $N$ is the number of cases or participants, and $k$ is the number of predictors in the model. This $F$ tests the null hypothesis that $R^2$ is zero (i.e., there is no improvement in the sum of squared error due to fitting the model).

## 8.2.5. Assessing individual predictors ①

We've seen that any 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 changed, what might we expect the change in the outcome to be? Well, if the model was very bad then we would expect the change in the outcome to be zero. Think back to Figure 8.5 (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 with only one predictor in the model (2) the gradient of the regression line is 0, meaning that the regression line is flat. Hopefully, you'll see that logically if a variable significantly predicts an outcome, then it should have a *b*-value that is 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 (think back to Section 2.5.1). 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.

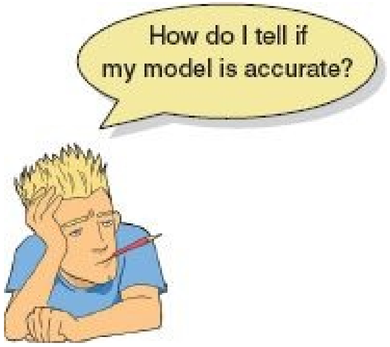
Equation (8.11) shows how the *t*-test is calculated and you'll find a general version of this equation in Chapter 9 (equation (9.2)). The $b_{\text{expected}}$ is simply the value of $b$ that we would expect to obtain if the null hypothesis were true. I mentioned earlier that the null hypothesis is 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:

$$
\begin{aligned}
t &= \frac{b_{\text{observed}} - b_{\text{expected}}}{SE_b} \\
&= \frac{b_{\text{observed}}}{SE_b}
\end{aligned}
\tag{8.11}
$$

The values of *t* have a special distribution that differs according to the degrees of freedom for the test. In this context, 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). SPSS 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.

## 8.3. Bias in regression models? ②

In Chapter 5 we saw that statistical models can be biased by unusual cases or by failing to meet certain assumptions. Therefore, when we have produced a model based on a sample of data, and assessed the fit, there are two important questions to ask: (1) is the model influenced by a small number of cases; and (2) can the model generalize to other samples? These questions are, in some sense, hierarchical because we wouldn’t want to generalize a bad model. However, it is a mistake to think that because a model fits the observed data well we can draw conclusions beyond our sample. **Generalization** is a critical additional step, and if we find that our model is not generalizable, then we must restrict any conclusions based on the model to the sample used. In Section 8.3.1 we will look at how we establish whether a model has been biased by unusual cases, and in Section 8.3.2 we move on to look at how we assess whether a model can be used to make inferences beyond the sample of data that has been collected.



### 8.3.1. Is the model biased by unusual cases? ②

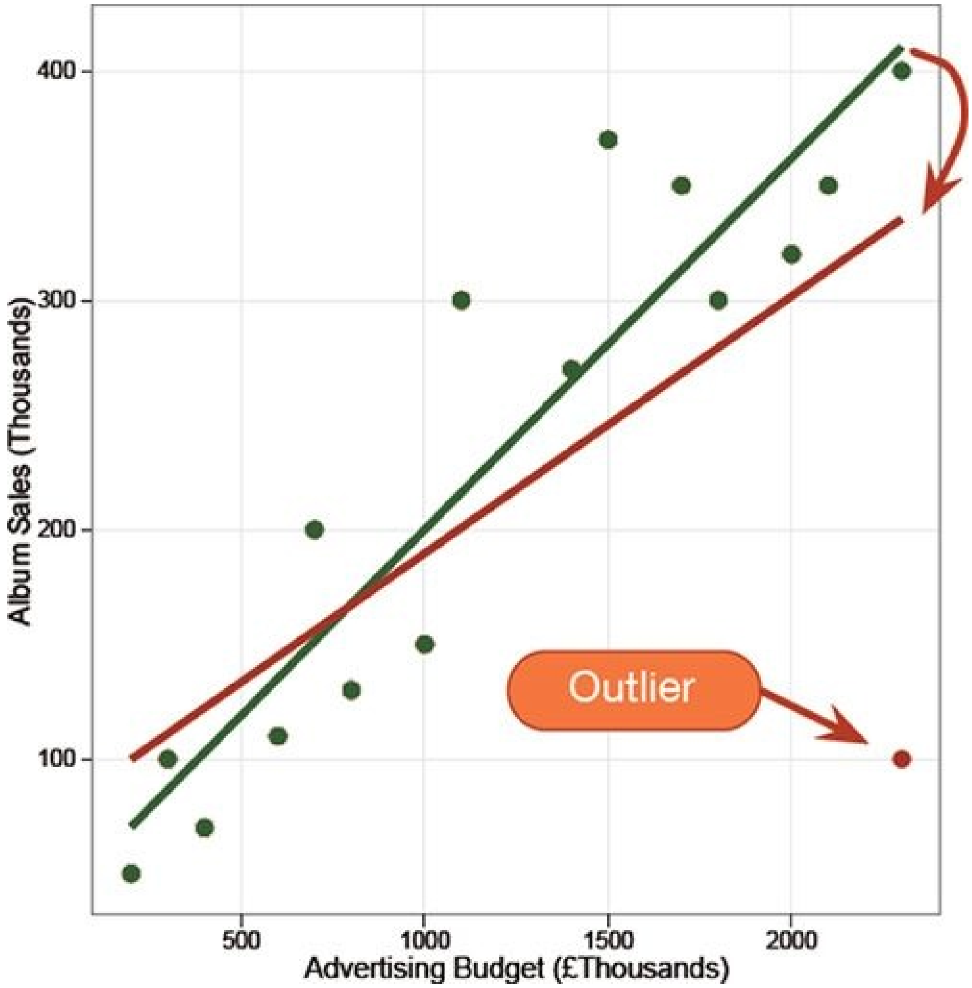
To answer the question of whether the model is influenced by a small number of cases, we can look for outliers and influential cases (the difference is explained in Jane Superbrain Box 8.1). We will look at these in turn.

### 8.3.1.1. Outliers and residuals ②

An outlier is a case that differs substantially from the main trend of the data (see Section 5.2.2). Outliers can affect the estimates of the regression coefficients. For example, Figure 8.6 uses the same data as Figure 8.4 except that the score of one album has been changed to be an outlier (in this case an album that sold relatively few despite a very large advertising budget). The green line shows the original model, and the red line shows the model with the outlier included. The outlier has a dramatic effect on the regression model: the line becomes flatter (i.e., $b_1$ is smaller) and the intercept increases (i.e., $b_0$ is larger). If outliers affect the estimates of the *b*s that define the model then it is important to detect these cases.

**FIGURE 8.6**

Graph demonstrating the effect of an outlier. The green line represents the original regression line for these data, whereas the red line represents the regression line when an outlier is present



How do you think that you might detect an outlier? Well, we know that an outlier, by its nature, is very different from all of the other scores. This being true, do you think that the model will predict that person's score very accuratelly? The answer is *no*: looking at Figure 8.6, it is evident that even though the outlier has biased the model, the model still predicts that one value very badly (the regression line is a long way from the outlier). Therefore, if we were to work out the differences between the data values that were collected, and the values predicted by the model, we could detect an outlier by looking for large differences. This process is the same as looking for cases that the model predicts inaccurately. We saw earlier that the differences between the values of the outcome predicted by the model and the values of the outcome observed in the sample are called *residuals*. These residuals represent the error present in the model. If a model fits the sample data well then all residuals will be small (if the model was a perfect fit of the sample data – all data points fall on the regression line – then all residuals would be zero). If a model is a poor fit of the sample data then the residuals will be large. Also, if any cases stand out as having a large residual, then they could be outliers.

SELF-TEST Residuals are used to compute which of the three sums of squares?

The *normal* or unstandardized residuals described above are measured in the same units as the outcome variable and so are difficult to interpret across different models. All we can do is to look for residuals that stand out as being particularly large: we cannot define a universal cut-off point for what constitutes a large residual. To overcome this problem, we use **standardized residuals**, which are the residuals converted to *z*-scores (see Section 1.6.4), which means they are converted into standard deviation units (i.e., they are distributed around a mean of 0 with a standard deviation of 1). By converting residuals into *z*-scores (standardized residuals) we can compare residuals from different models and use what we know about the properties of *z*-scores to devise universal guidelines for what constitutes an acceptable (or unacceptable) value. For example, we know from Chapter 1 that in a normally distributed sample, 95% of *z*-scores should lie between −1.96 and +1.96, 99% should lie between −2.58 and +2.58, and 99.9% (i.e., nearly all of them) should lie between −3.29 and +3.29. Some general rules for standardized residuals are derived from these facts: (1) standardized residuals with an absolute value greater than 3.29 (we can use 3 as an approximation) are cause for concern because in an average sample a value this high is unlikely to occur; (2) if more than 1% of our sample cases have standardized residuals with an absolute value greater than 2.58 (we usually just say 2.5) there is evidence that the level of error within our model is unacceptable (the model is a fairly poor fit of the sample data); and (3) if more than 5% of cases have standardized residuals with an absolute value greater than 1.96 (we can use 2 for convenience) then there is also evidence that the model is a poor representation of the actual data.

A third form of residual is the **Studentized residual**, which is the unstandardized residual divided by an estimate of its standard deviation that varies point by point. These residuals have the same properties as the standardized residuals but usually provide a more precise estimate of the error variance of a specific case.

#### 8.3.1.2. Influential cases ③

As well as testing for outliers by looking at the error in the model, it is also possible to look at whether certain cases exert undue influence over the parameters of the model. So, if we were to delete a certain case, would we obtain different regression coefficients? This type of analysis can help to determine whether the regression model is stable across the sample, or whether it is biased by a few influential cases. Again, this process will unveil outliers.

There are several residual statistics that can be used to assess the influence of a particular case. One statistic is the **adjusted predicted value** for a case when that case is excluded from the analysis. In effect, the computer calculates a new model without a particular case and then uses this new model to predict the value of the outcome variable for the case that was excluded. If a case does not exert a large influence over the model then we would expect the adjusted predicted value to be very similar to the predicted value when the case is included. Put simply, if the model is stable then the predicted value of a case should be the same regardless of whether or not that case was used to estimate the model. We can also look at the residual based on the adjusted predicted value: that is, the difference between the adjusted predicted value and the original observed value. This is the **deleted residual**. The deleted residual can be divided by the standard error to give a standardized value known as the **Studentized deleted residual**. This residual can be compared across different regression analyses because it is measured in standard units.

SMART ALEX ONLY

The deleted residuals are very useful to assess the influence of a case on the ability of the model to predict that case. However, they do not provide any information about how a case influences the model as a whole (i.e., the impact that a case has on the model's ability to predict *all* cases). One statistic that does consider the effect of a single case on the model as a whole is **Cook's distance**. Cook's distance is a measure of the overall influence of a case on the model, and Cook and Weisberg (1982) have suggested that values greater than 1 may be cause for concern.

A second measure of influence is **leverage** (sometimes called **hat values**), which gauges the influence of the observed value of the outcome variable over the predicted values. The average leverage value is defined as $(k + 1)/n$, in which $k$ is the number of predictors in the model and $n$ is the number of participants.[7] The maximum value for leverage is $(N – 1)/N$; however, SPSS calculates a version of the leverage that takes a maximum value of 1 (indicating that the case has complete influence over prediction).

- If no cases exert undue influence over the model then we would expect all of the leverage values to be close to the average value ($(k + 1)/n$).
- Hoaglin and Welsch (1978) recommend investigating cases with values greater than twice the average ($2(k + 1)/n$).
- Stevens (2002) recommends using three times the average ($3(k + 1)/n$) as a cut-off point for identifying cases having undue influence.

We will see how to use these cut-off points later. However, cases with large leverage values will not necessarily have a large influence on the regression coefficients because they are measured on the outcome variables rather than the predictors.

Related to the leverage values are the **Mahalanobis distances**, which measure the distance of cases from the mean(s) of the predictor variable(s). Look for the cases with the highest values. These distances have a chi-square distribution, with degrees of freedom equal to the number of predictors (Tabachnick & Fidell, 2012). One way to establish a cut-off point is to find the critical value of chi-square for the desired alpha level (values for $p$ = .05 and .01 are in the Appendix). For example, with three predictors, a distance greater than 7.81 ($p$ = .05) or 11.34 ($p$ = .01) would be cause for concern.

Barnett and Lewis (1978) have also produced a table of critical values dependent on the number of predictors and the sample size. From their work it is clear that even with large samples ($N$ = 500) and five predictors, values above 25 are cause for concern. In smaller samples ($N$ = 100) and with fewer predictors (namely, three), values greater than 15 are problematic, and in very small samples ($N$ = 30) with only two predictors, values, greater than 11 should be examined.

It is possible to run the regression analysis with a case included and then rerun the analysis with that same case excluded. If we did this, undoubtedly there would be some difference between the $b$ coefficients in the two regression equations. This difference would tell us how much influence a particular case has on the parameters of the regression model. To take a hypothetical example, imagine two variables that had a perfect negative relationship except for a single case (case 30). If a regression analysis was done on the 29 cases that were perfectly linearly related then we would get a model in which the predictor variable $X$ perfectly predicts the outcome variable $Y$, and there are no errors. If we then ran the analysis but this time include the case that didn't conform (case 30), then the resulting model would have different parameters. Some data are stored in the file **DFBeta.sav** which illustrate such a situation.

**FIGURE 8.7**
Prasanta Chandra Mahalanobis staring into his distances

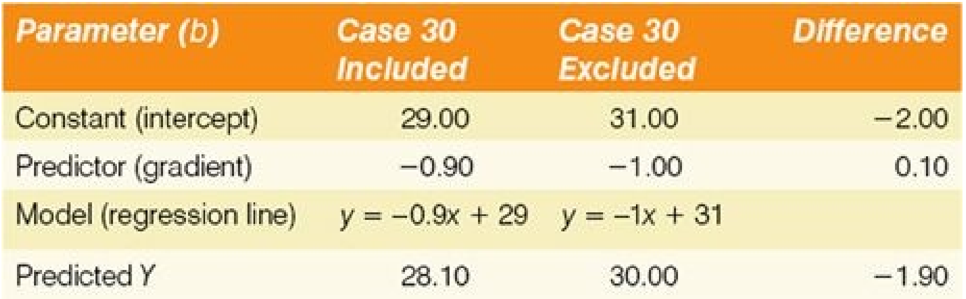
SELF-TEST Once you have read Section 8.4, run a regression first with all the cases included and then with case 30 deleted.

The results are summarized in Table 8.1, which shows: (1) the parameters for the regression model when the extreme case is included or excluded; (2) the resulting regression equations; and (3) the value of $Y$ When case 30's score on the $X$ variable (which is obtained by replacing the X in the regression equation with participant 30's score for $X$, which was 1).

When case 30 is excluded, these data have a perfect negative relationship; hence the coefficient for the predictor ($b_1$) is −1 (remember that in simple regression this term is the same as Pearson's correlation coefficient), and the coefficient for the constant (the intercept, $b_0$) is 31. However, when case

30 is included, both parameters are reduced[8] and the difference between the parameters is also displayed. The difference between a parameter estimated using all cases and estimated when one case is excluded is known as the **DFBeta**. DFBeta is calculated for every case and for each of the parameters in the model. So, in our hypothetical example, the DFBeta for the constant is −2, and the DFBeta for the predictor variable is 0.1. By looking at the values of DFBeta, it is possible to identify cases that have a large influence on the parameters of the regression model. Again, the units of measurement used will affect these values and so SPSS produces a **standardized DFBeta**. These standardized values are easier to use because universal cut-off points can be applied. In this case absolute values above 1 indicate cases that substantially influence the model parameters (although Stevens (2002) suggests looking at cases with absolute values greater than 2).

**TABLE 8.1** The difference in the parameters of the regression model when one case is excluded



| *Parameter (b)* | *Case 30 Included* | *Case 30 Excluded* | *Difference* |
|---|---|---|---|
| Constant (intercept) | 29.00 | 31.00 | −2.00 |
| Predictor (gradient) | −0.90 | −1.00 | 0.10 |
| Model (regression line) | $y = -0.9x + 29$ | $y = -1x + 31$ | |
| Predicted *Y* | 28.10 | 30.00 | −1.90 |

A related statistic is the **DFFit**, which is the difference between the predicted value for a case when the model is calculated including that case and when the model is calculated excluding that case: in this example the value is −1.90 (see Table 8.1). If a case is not influential then its DFFit should be zero – hence, we expect non-influential cases to have small DFFit values. However, we have the problem that this statistic depends on the units of measurement of the outcome and so a DFFit of 0.5 will be very small if the outcome ranges from 1 to 100, but very large if the outcome varies from 0 to 1. Therefore, SPSS also produces standardized versions of the DFFit values (**Standardized DFFit**).

A final measure is the **covariance ratio (CVR)**, which is a measure of whether a case influences the variance of the regression parameters. A description of the computation of this statistic leaves me dazed and confused, so suffice to say that when this ratio is close to 1 the case has very little influence on the variances of the model parameters. Belsey, Kuh, and Welsch (1980) recommend the following:

- If $\text{CVR}_i > 1 + [3(k + 1)/n]$ then deleting the *i*th case will damage the precision of some of the model's parameters.
- If $\text{CVR}_i < 1 - [3(k + 1)/n]$ then deleting the *i*th case will improve the precision of some of the model's parameters.

In both equations, $k$ is the number of predictors, $\text{CVR}_i$ is the covariance ratio for the *i*th participant, and $n$ is the sample size.

### 8.3.1.3. A final comment on diagnostic statistics ②

There are a lot of diagnostic statistics that should be examined after a regression analysis, and it is difficult to summarize this wealth of material into a concise conclusion. However, one thing I would like to stress is a point made by Belsey et al. (1980) who noted the dangers inherent in these procedures.

The point is that diagnostics are tools that enable you to see how good or bad your model is in terms of fitting the sampled data. They are a way of assessing your model. They are *not*, however, a way of justifying the removal of data points to effect some desirable change in the regression parameters (e.g., deleting a case that changes a non-significant *b*-value into a significant one). Stevens (2002), as ever, offers excellent advice:

EVERYBODY

> If a point is a significant outlier on *Y*, but its Cook's distance is < 1, there is no real need to delete that point since it does not have a large effect on the regression analysis. However, one should still be interested in studying such points further to understand why they did not fit the model. (p. 135)

## 8.3.2. Generalizing the model ②

When a regression analysis is done, an equation can be produced that is correct for the sample of observed values. However, we are usually interested in generalizing our findings outside of the sample. For a regression model to generalize we must be sure that underlying assumptions have been met, and to test whether the model does generalize we can look at cross-validating it.

### 8.3.2.1. Assumptions of the linear model ②

We have already looked at the main assumptions of the linear model and how to assess them in Chapter 5. I will recap the main ones in order of importance (Gelman & Hill, 2007):

- *Additivity and linearity*: The outcome variable should, in reality, be linearly related to any predictors and, with several predictors, their combined effect is best described by adding their effects together. In other words, the process we're trying to model can be described by the linear model. If this assumption isn't met then the model is invalid. You can sometimes transform variables to make their relationships linear (see Chapter 5).

## JANE SUPERBRAIN 8.1

### *The difference between residuals and influence statistics* ③

To illustrate how residuals and influence statistics differ, imagine that the Mayor of London at the turn of the last century was interested in how drinking affected mortality. London is divided up into different regions called boroughs, and so he might measure the number of pubs and the number of deaths over a period of time in eight of his boroughs. The data are in a file called **pubs.sav**.

The scatterplot of these data (Figure 8.8) reveals that without the last case there is a perfect linear relationship (the dashed straight line). However, the presence of the last case (case 8) changes the line of best fit dramatically (although this line is still a significant fit to the data – do the regression analysis and see for yourself).

What's interesting about these data is when we look at the residuals and influence statistics. The standardized residual for case 8 is the second *smallest:* this outlier produces a very small residual (most of the non-outliers have larger residuals) because it sits very close to the line that has been fitted to the data. How can this be? Look at the influence statistics below and you'll see that they're massive for case 8: it exerts a huge influence over the model.

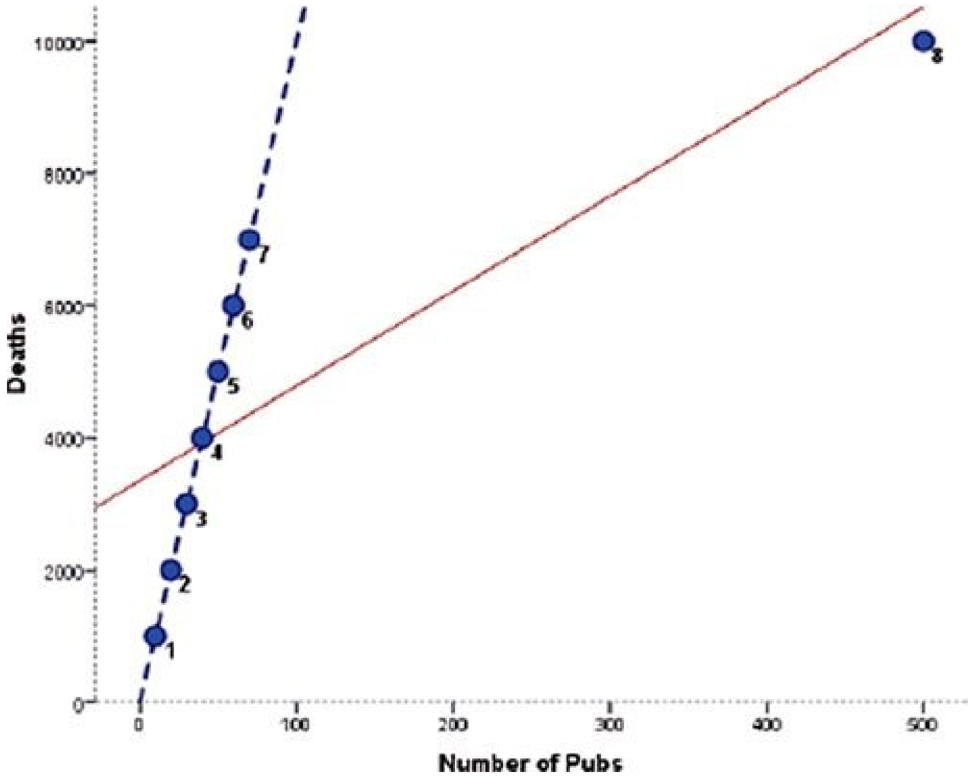
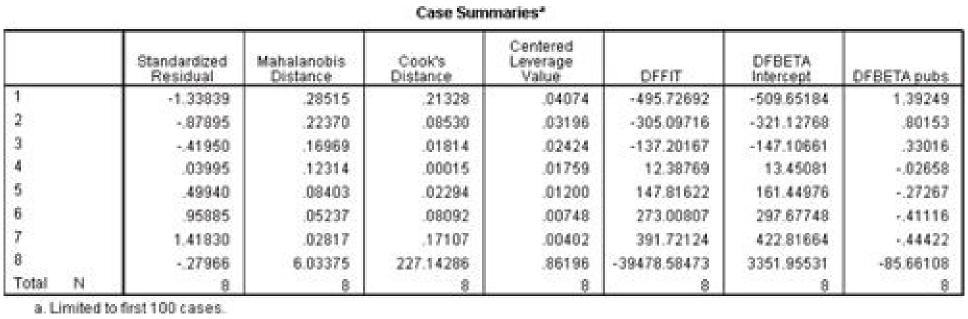


**FIGURE 8.8** With non-parametric tests you must double-click the summary table within the viewer window to open up the model viewer window

As always, when you see a statistical oddity you should ask what was happening in the real world. The last data point represents the City of London, a tiny area of only 1 square mile in the centre of London where very few people lived but where thousands of commuters (even then) came to work and had lunch in the pubs. Hence the pubs didn't rely on the resident population for their business and the residents didn't consume all of their beer. Therefore, there was a massive number of pubs. This illustrates that a case exerting a massive influence can produce a small residual – so look at both (I'm very grateful to David Hitchin for this example, and he in turn got it from Dr Richard Roberts.)



Case Summaries[a]

| | | Standardized Residual | Mahalanobis Distance | Cook's Distance | Centered Leverage Value | DFFIT | DFBETA Intercept | DFBETA pubs |
|---|---|---|---|---|---|---|---|---|
| 1 | | -1.33839 | .28515 | .21328 | .04074 | -495.72692 | -509.65184 | 1.39249 |
| 2 | | -.87895 | .22370 | .08530 | .03196 | -305.09716 | -321.12768 | .80153 |
| 3 | | -.41950 | .16969 | .01814 | .02424 | -137.20167 | -147.10661 | .33016 |
| 4 | | .03995 | .12314 | .00015 | .01759 | 12.38769 | 13.45081 | -.02658 |
| 5 | | .49940 | .08403 | .02294 | .01200 | 147.81622 | 161.44976 | -.27267 |
| 6 | | .95885 | .05237 | .08092 | .00748 | 273.00807 | 297.67748 | -.41116 |
| 7 | | 1.41830 | .02817 | .17107 | .00402 | 391.72124 | 422.81664 | -.44422 |
| 8 | | -.27966 | 6.03375 | 227.14286 | .86196 | -39478.58473 | 3351.95531 | -85.66108 |
| Total | N | 8 | 8 | 8 | 8 | 8 | 8 | 8 |

a. Limited to first 100 cases.

- **Independent errors**: For any two observations the residual terms should be uncorrelated (i.e., independent). This eventuality is sometimes described as a lack of **autocorrelation**. If we violate the assumption of independence then our confidence intervals and significance tests will be invalid. However, in terms of the model parameters themselves, the estimates using the method of least squares will still be valid but not optimal (see Section 5.2.6). This assumption can be tested with the **Durbin–Watson test**, which tests for serial correlations between errors. Specifically, it tests whether adjacent residuals are correlated. The test statistic can vary between 0 and 4, with a value of 2 meaning that the residuals are uncorrelated. A value greater than 2 indicates a negative correlation between adjacent residuals, whereas a value below 2 indicates a positive correlation. The size of the Durbin–Watson statistic depends upon the number of predictors in the model and the number of observations. For accuracy, you should look up the exact acceptable values in Durbin and Watson's (1951) original paper. As a very conservative rule of thumb, values less than 1 or greater than 3 are definitely cause for concern; however, values closer to 2 may still be problematic, depending on your sample and model.

- *Homoscedasticity* (see Section 5.2.5): At each level of the predictor variable(s), the variance of the residual terms should be constant. This just means that the residuals at each level of the predictor(s) should have the same variance (homoscedasticity); when the variances are very unequal there is said to be heteroscedasticity. Violating this assumption invalidates our confidence intervals and significance tests. However, estimates of the model parameters (*b*) using the method of least squares are still valid but not optimal. This problem can be overcome using weighted least squares regression in which each case is weighted by a function of its variance.
- *Normally distributed errors* (see Section 5.2.4): It is assumed that the residuals in the model are random, normally distributed variables with a mean of 0. This assumption simply means that the differences between the model and the observed data are most frequently zero or very close to zero, and that differences much greater than zero happen only occasionally. Some people confuse this assumption with the idea that predictors have to be normally distributed. In fact, predictors do not need to be normally distributed. In small samples a lack of normality will invalidate confidence intervals and significance tests; in large samples it will not, because of the central limit theorem. If you are concerned only with estimating the model parameters (and not significance tests and confidence intervals) then this assumption barely matters. If you bootstrap confidence intervals then you really can ignore this assumption.

There are some other considerations that we have not yet discussed (see Berry, 1993):

- *Predictors are uncorrelated with 'external variables'*: *External variables* are variables that haven't been included in the regression model and that influence the outcome variable.[9] These variables can be thought of as similar to the 'third variable' that was discussed with reference to correlation. This assumption means that there should be no external variables that correlate with any of the variables included in the regression model. Obviously, if external variables do correlate with the predictors, then the conclusions we draw from the model become unreliable (because other variables exist that can predict the outcome just as well).
- *Variable types*: All predictor variables must be quantitative or categorical (with two categories), and the outcome variable must be quantitative, continuous and unbounded. By 'quantitative' I mean that they should be measured at the interval level and by 'unbounded' I mean that there should be no constraints on the variability of the outcome. If the outcome is a measure ranging from 1 to 10 yet the data collected vary between 3 and 7, then these data are constrained.
- *No perfect* **multicollinearity**: If your model has more than one predictor then there should be no perfect linear relationship between two or more of the predictors. So, the predictor variables should

not correlate too highly (see Section 8.5.3).

- *Non-zero variance*: The predictors should have some variation in value (i.e., they do not have variances of 0). This is self-evident really.

As we saw in Chapter 5, violating these assumptions has implications mainly for significance tests and confidence intervals; the estimates of *b*s are not dependent on these assumptions (although least squares methods will be optimal when the assumptions are met). However, the confidence interval for a *b* tells us the boundaries within which the population values of that *b* are likely to fall. Therefore, if confidence intervals are inaccurate (as they are when these assumptions are broken) then we cannot accurately estimate the likely population value. This means we can't generalize our model to the population. When the assumptions are met, then, *on average* the regression model from the sample is the same as the population model. However, you should be clear that even when the assumptions are met, it is possible that a model obtained from a sample may not be the same as the population model – but the likelihood of them being the same is increased.

### 8.3.2.2. Cross-validation of the model ③

Even if we can't be confident that the model derived from our sample accurately represents the entire population, we can assess how well our model can predict the outcome in a different sample. Assessing the accuracy of a model across different samples is known as **cross-validation**. If a model can be generalized, then it must be capable of accurately predicting the same outcome variable from the same set of predictors in a different group of people. If the model is applied to a different sample and there is a severe drop in its predictive power, then the model clearly does *not* generalize. As a first rule of thumb, we should aim to collect enough data to obtain a reliable regression model (see the next section). Once we have a regression model there are two main methods of cross-validation:

- *Adjusted* $R^2$: SPSS computes an **adjusted** $R^2$. Whereas $R^2$ tells us how much of the variance in $Y$ is accounted for by the regression model from our sample, the adjusted value tells us how much variance in $Y$ would be accounted for if the model had been derived from the population from which the sample was taken. Therefore, the adjusted value indicates the loss of predictive power or **shrinkage**. SPSS derives the adjusted $R^2$ using Wherry's equation. This equation has been criticized because it tells us nothing about how well the regression model would predict scores of a different sample of data from the same population. One version of $R^2$ that does tell us how well the model cross-validates uses Stein's formula (see Stevens, 2002).

  $$\text{adjusted } R^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\right](1-R^2) \qquad (8.12)$$

  In Stein's equation, $R^2$ is the unadjusted value, $n$ is the number of participants and $k$ is the number of predictors in the model. For the more mathematically minded of you, it is worth using this equation to cross-validate a regression model.
- *Data splitting*: This approach involves randomly splitting your sample data, computing a regression equation on both halves of the data and then comparing the resulting models. When using stepwise methods (see Section 8.5.1.3), cross-validation is particularly important; you should run the stepwise regression on a random selection of about 80% of your cases. Then force this model on the remaining 20% of the data. By comparing values of the $R^2$ and $b$-values in the two samples you can tell how well the original model generalizes (see Tabachnick & Fidell, 2012, for more detail).

### 8.3.3. Sample size in regression ③

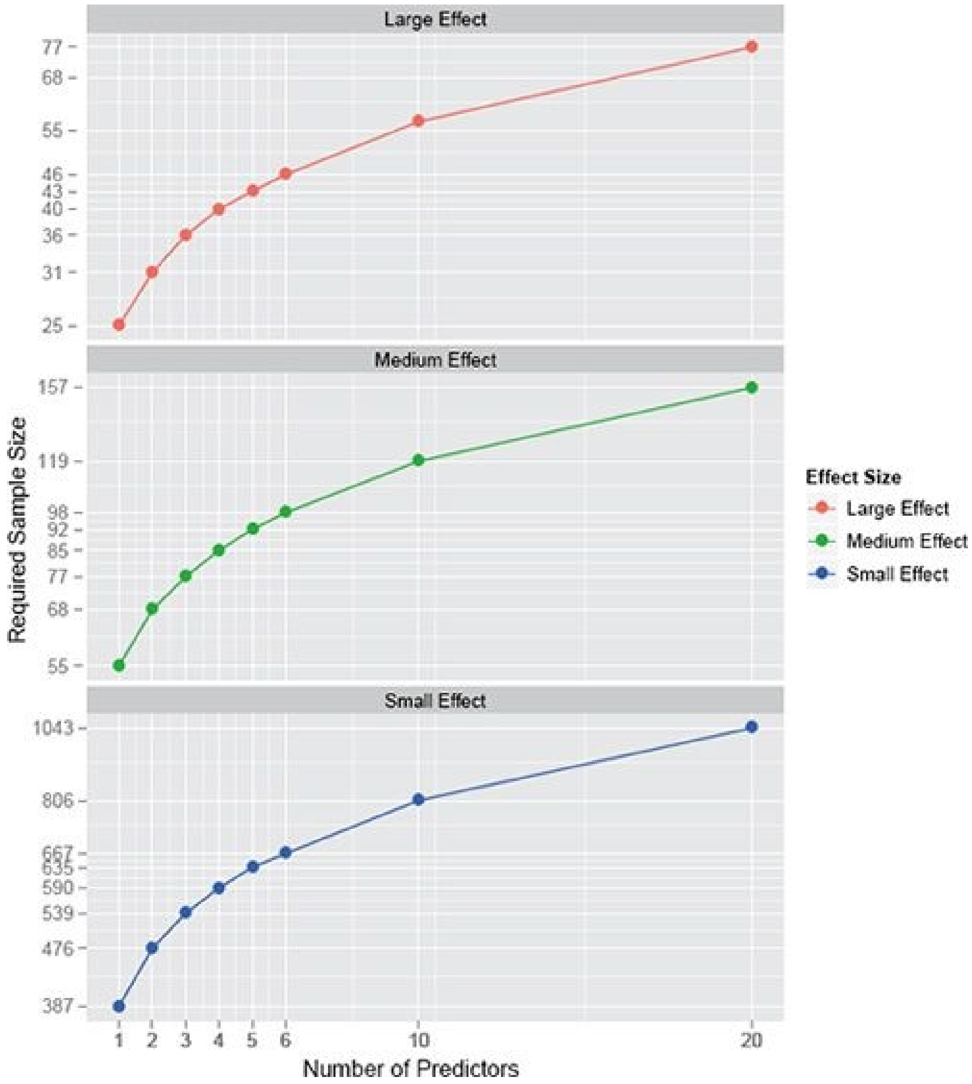
In the previous section I said that it’s important to collect enough data to obtain a reliable regression model. Also, larger samples enable us to assume that our *b*s are from a normally distributed sampling distribution because of the central limit theorem (Section 5.2.4.2). Well, how much is enough?

You’ll find a lot of rules of thumb floating about, the two most common being that you should have 10 cases of data for each predictor in the model, or 15 cases of data per predictor. So, with five predictors, you’d need 50 or 75 cases respectively (depending on the rule you use). These rules are very pervasive but they oversimplify the issue. In fact, the sample size required will depend on the size of effect that we’re trying to detect (i.e., how strong the relationship is that we’re trying to measure) and how much power we want to detect these effects. The simplest rule of thumb is that the bigger the sample size, the better: the estimate of $R$ that we get from regression is dependent on the number of predictors, $k$, and the sample size, $N$. In fact, the expected $R$ for random data is $k/(N - 1)$ and so with small sample sizes random data can appear to show a strong effect: for example, with six predictors and 21 cases of data, $R = 6/(21 - 1) =.3$ (a medium effect size by Cohen’s criteria described in Section 7.2.2). Obviously for random data we’d want the expected $R$ to be 0 (no effect) and for this to be true we need large samples (to take the previous example, if we had 100 cases rather than 21, then the expected $R$ would be a more acceptable .06).

It’s all very well knowing that larger is better, but researchers usually need some more concrete guidelines (much as we’d all love to collect 1000 cases of data, it isn’t always practical). As I’ve mentioned before, the sample size required depends on the size of the effect (i.e., how well our predictors predict the outcome), how much statistical power we want to detect these effects, and what we’re testing (the significance of the *b*-values, or the significance of the model overall). Figure 8.9 shows the sample size required[10] to achieve a high level of power (I’ve taken Cohen’s (1988) benchmark of .8) to test that the model is significant overall (i.e., R2 is not equal to zero). I’ve varied the number of predictors and the size of expected effect: I used $R^2 = .02$ (small), .13 (medium) and .26 (large), which correspond to benchmarks in Cohen (1988). Broadly speaking, if your aim is to test the overall fit of the model: (1) if you expect to find a large effect then a sample size of 77 will always suffice (with up to 20 predictors) and if there are fewer predictors then you can afford to have a smaller sample; (2) if you’re expecting a medium effect, then a sample size of 160 will always suffice (with up to 20 predictors), you should always have a sample size above 55, and with six or fewer predictors you’ll be fine with a sample of 100; and (3) if you’re expecting a small effect size then just don’t bother unless you have the time and resources to collect hundreds of cases of data. Miles and Shevlin (2001) produce some more detailed graphs that are worth a look, but the take-home message is that if you’re looking for medium to large effects, sample sizes don’t need to be massive, regardless of how many predictors you have.

**FIGURE 8.9**

The sample size required to test the overall regression model depending on the number of predictors and the size of expected effect, $R^2$ = .02 (small), .13 (medium) and .26 (large)

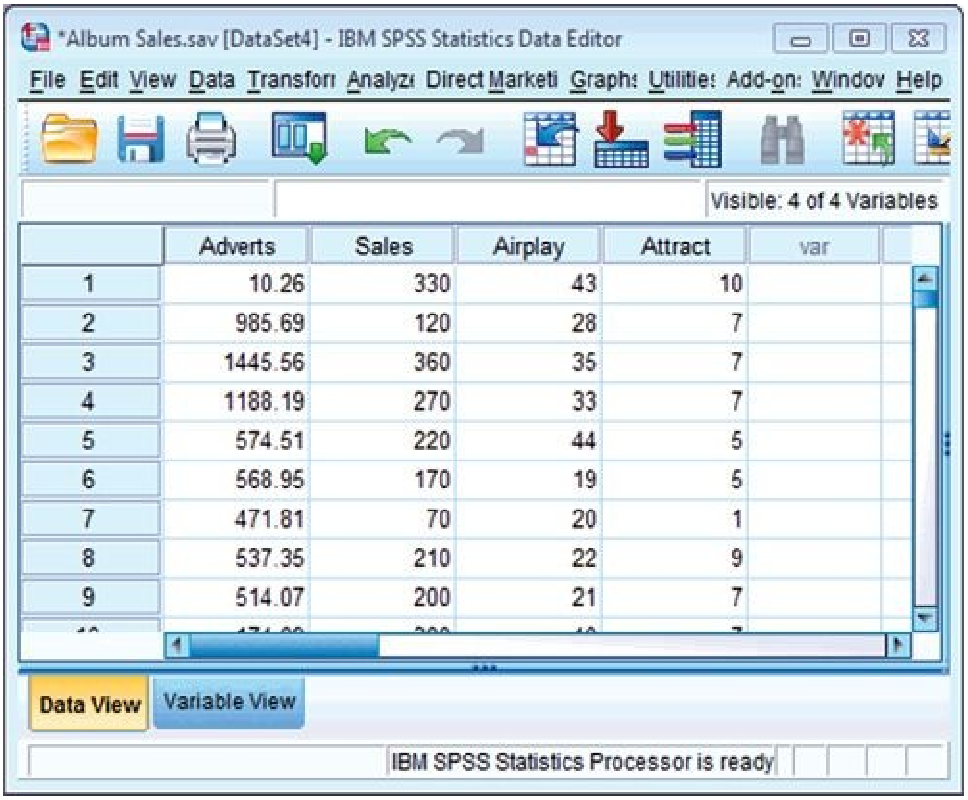


## 8.4. Regression using SPSS: One Predictor ①

To help clarify what we have learnt so far, we will go through an example of a regression with one predictor before looking in a bit more detail at models with several predictors. Earlier on I asked you to imagine that I worked for a record company and that my boss was interested in predicting album sales from advertising. There are some data for this example in the file **Album Sales.sav**. This data file has 200 rows, each one representing a different album. There are also several columns, one of which contains the sales (in thousands) of each album in the week after release (**Sales**) and one containing the amount (in thousands of pounds) spent promoting the album before release (**Adverts**). The other columns represent how many times songs from the album were played on radio in the week before release (**Airplay**), and how attractive people found the band out of 10 (**Attract**). Ignore these last two variables for now; we'll use them later. Note how the data are laid out (Figure 8.10): each variable is in a column and each row represents a different album. So, the first album had £10,260 spent advertising it, sold 330,000 copies, received 43 plays on Radio 1 the week before release, and was made by a band that the majority of people rated as gorgeous sex objects.

FIGURE 8.10

Data layout for regression


*Album Sales.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transforı Analyzı Direct Marketi Graphı Utilitieı Add-onı Windov Help
Visible: 4 of 4 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

| | Adverts | Sales | Airplay | Attract | var |
|---|---|---|---|---|---|
| 1 | 10.26 | 330 | 43 | 10 | |
| 2 | 985.69 | 120 | 28 | 7 | |
| 3 | 1445.56 | 360 | 35 | 7 | |
| 4 | 1188.19 | 270 | 33 | 7 | |
| 5 | 574.51 | 220 | 44 | 5 | |
| 6 | 568.95 | 170 | 19 | 5 | |
| 7 | 471.81 | 70 | 20 | 1 | |
| 8 | 537.35 | 210 | 22 | 9 | |
| 9 | 514.07 | 200 | 21 | 7 | |

## 8.4.1. Regression: the general procedure ①

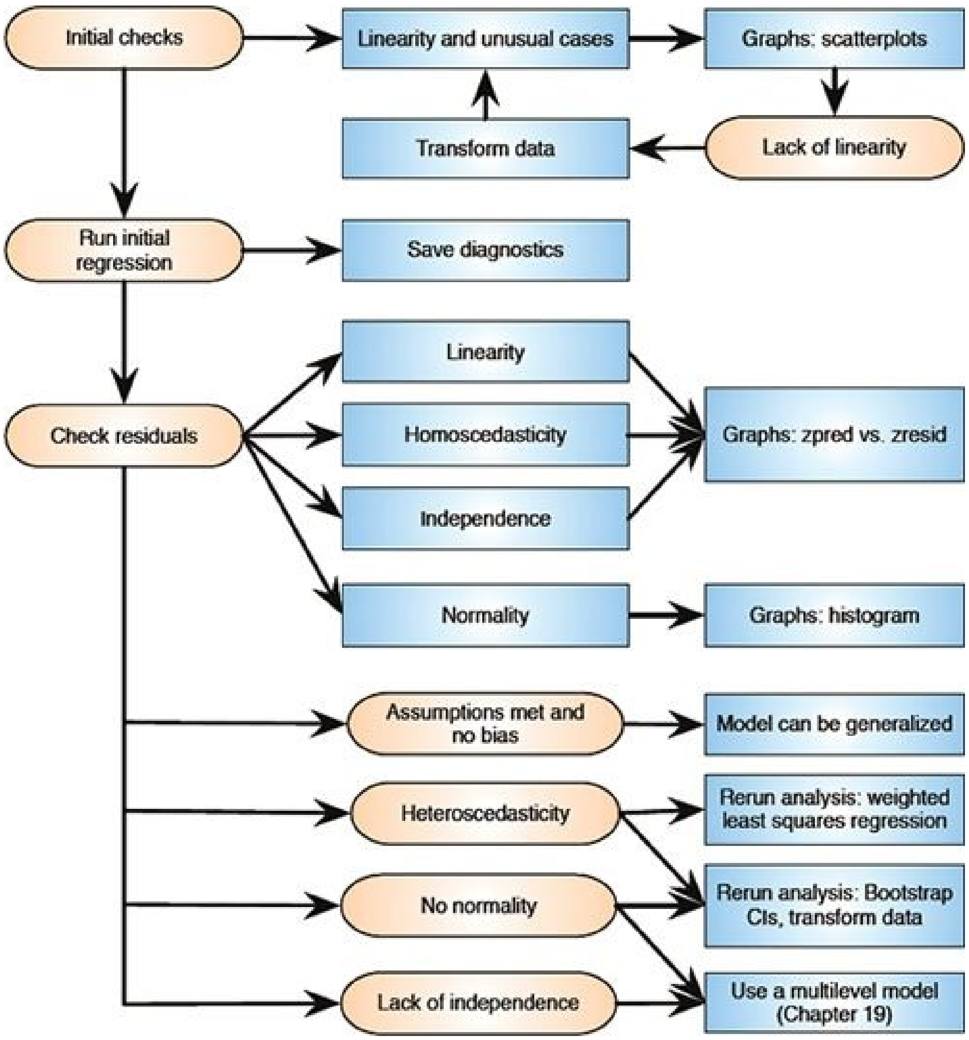
Figure 8.11 shows the general process of conducting regression analysis. First, we should produce scatterplots to get some idea of whether the assumption of linearity is met, and also to look for any outliers or obvious unusual cases. At this stage we might transform the data to correct problems. Having done this initial screen for problems, we fit a model and save the various diagnostic statistics that we discussed in Section 8.3. If we want to generalize our model beyond the sample, or we are interested in interpreting significance tests and confidence intervals, then we examine these residuals to check for homoscedasticity, normality, independence and linearity (although this will likely be fine given our earlier screening). If we find problems then we take corrective action and re-estimate the model. This process might seem complex, but it’s not as bad as it seems. Also, it’s probably wise to use bootstrapped confidence intervals when we first estimate the model because then we can basically forget about things like normality.

SELF-TEST Produce a scatterplot of sales ($y$-axis) against advertising budget ($x$-axis). Include the regression line.

The pattern of the data is shown in Figure 8.12, and it should be clear that a positive relationship exists: so, the more money spent advertising the album, the more it is likely to sell. Of course there are some albums 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.

**FIGURE 8.11**

The process of fitting a regression model.



## 8.4.2. Running a simple regression using SPSS ①

To do the analysis you need to access the main dialog box by selecting Analyze Regression ▸ Linear... . Figure 8.13 shows the resulting dialog box. There is a space labelled *Dependent* in which you should place the outcome variable (in this example **Sales**). So, select **Sales** from the list on the left-hand side, and transfer it by dragging it or clicking on ➡. There is another space labelled *Independent(s)* in which any predictor variable should be placed. In simple regression we use only one predictor (in this example, **Adverts**) and so you should select **Adverts** from the list and click on ➡ to transfer it to the list of predictors. There are a variety of options available, but these will be explored within the context of multiple regression. However, we can get bootstrapped confidence intervals for the regression coefficients by clicking on Bootstrap (see Section 5.4.3). Select to activate bootstrapping, and to get a 95% confidence interval click ☑ Perform bootstrapping. Click on OK in the main dialog box to run the basic analysis.

**FIGURE 8.12**

Scatterplot showing the relationship between album sales and the amount spent promoting the album

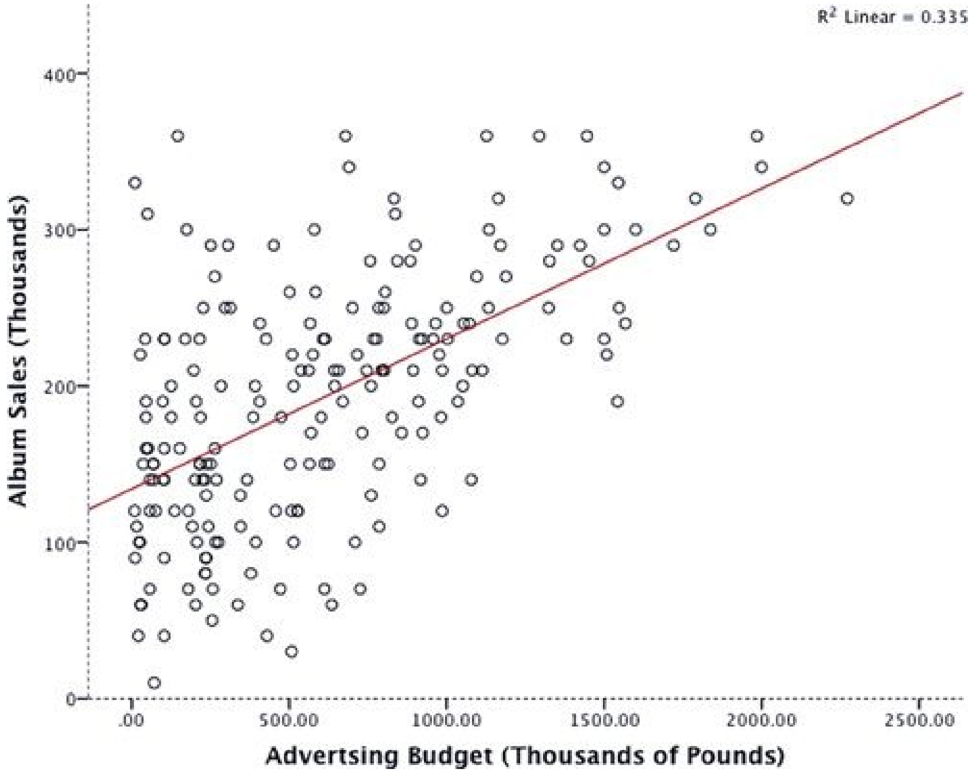
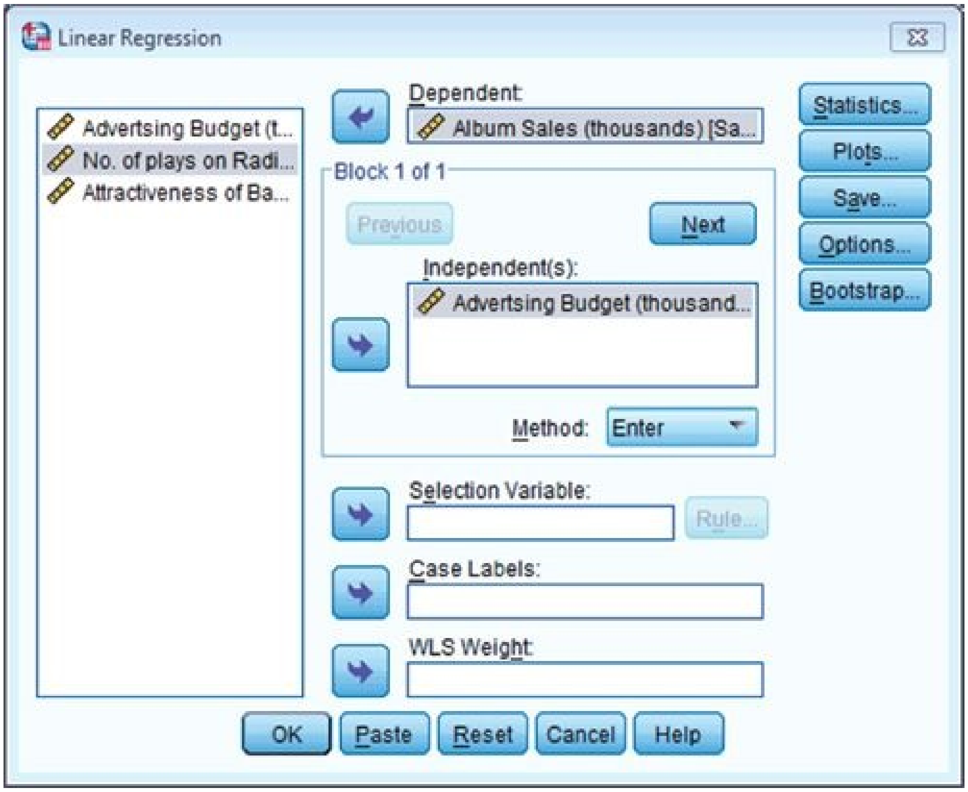


**FIGURE 8.13**

Main dialog box for regression



Dependent: Album Sales (thousands) [Sa...

Independent(s): Advertising Budget (thousand...

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

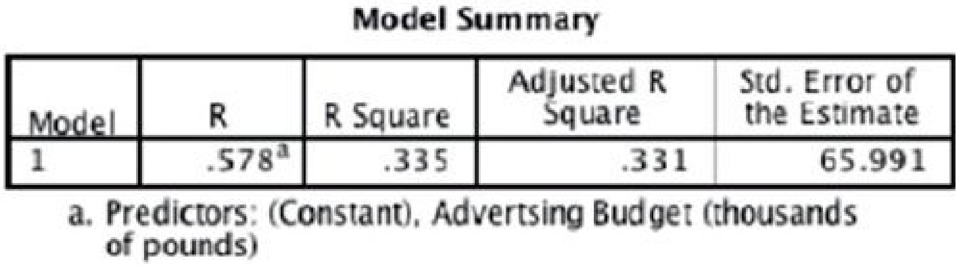
## 8.4.3. Interpreting a simple regression ①

### 8.4.3.1. Overall fit of the model ①

The first table provided by SPSS is a summary of the model (Output 8.1). This summary table provides the value of $R$ and $R^2$ for the model that has been derived. For these data, $R$ has a value of .578 and because there is only one predictor, this value represents the simple correlation between advertising and album 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 album sales. In other words, if we are trying to explain why some albums sell more than others, we can look at the variation in sales of different albums. 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 66% of the variation in album sales cannot be explained by advertising alone. Therefore, there must be other variables that have an influence also.

The next part of the output (Output 8.2) reports an analysis of variance (ANOVA – see Chapter 11). The summary table shows the various sums of squares described in Figure 8.5 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 (8.9), 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). This result tells us that there is less than a 0.1% chance that an $F$-ratio this large would happen if the null hypothesis were true. Therefore, we can conclude that our regression model results in significantly better prediction of album sales than if we used the mean value of album sales. In short, the regression model overall predicts album sales significantly well.

**OUTPUT 8.1**

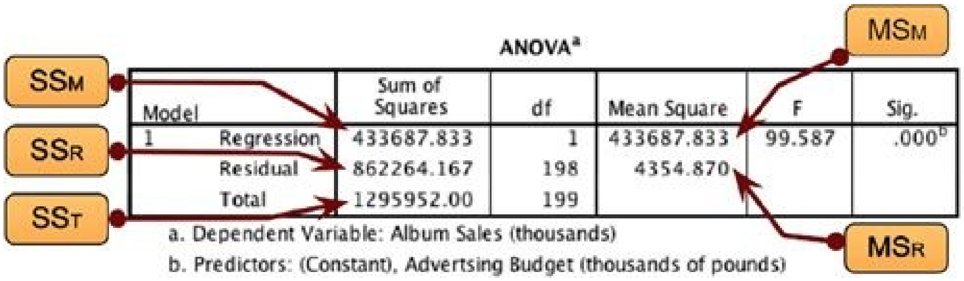


Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .578[a] | .335 | .331 | 65.991 |

a. Predictors: (Constant), Advertsing Budget (thousands of pounds)

**OUTPUT 8.2**



ANOVA[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 433687.833 | 1 | 433687.833 | 99.587 | .000[b] |
| | Residual | 862264.167 | 198 | 4354.870 | | |
| | Total | 1295952.00 | 199 | | | |

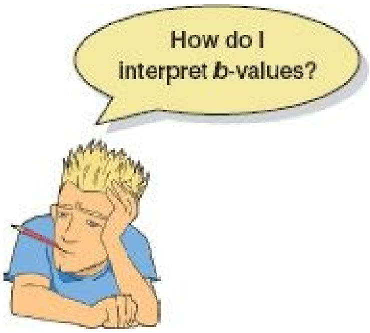
a. Dependent Variable: Album Sales (thousands)

b. Predictors: (Constant), Advertsing Budget (thousands of pounds)

SSM
SSR
SST
MSM
MSR

### 8.4.3.2. Model parameters ①

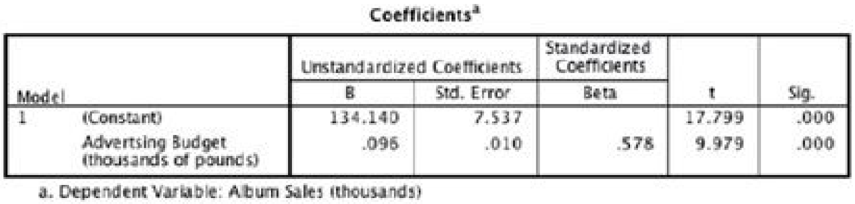
The ANOVA tells us whether the model, overall, results in a significantly good degree of prediction of the outcome variable. However, the ANOVA doesn't tell us about the individual contribution of variables in the model (although in this simple case there is only one variable in the model and so we can infer that this variable is a good predictor). The table in Output 8.3 provides estimates of the model parameter (the beta values) and the significance of these values. We saw in equation (8.1) that $b_0$ was the $Y$ intercept, and this value is the value $B$ (in the SPSS 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 albums will be sold (remember that our unit of measurement was thousands of albums). 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 it 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 albums will be sold. Our units of measurement were thousands of pounds and thousands of albums sold, so we can say that for an increase in advertising of £1000 the model predicts 96 (0.096 × 1000 = 96) extra album sales. As you might imagine, this investment is pretty bad for the album company: it invests £1000 and gets only 96 extra sales! Fortunately, as we already know, advertising accounts for only one-third of album 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 has 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. SPSS provides the exact probability that the observed value of $t$ would occur if the value of $b$ in the population were zero. If this observed significance is less than .05, then the result reflects a genuine effect (see Chapter 2). For both $t$s, the probabilities are given as .000 (zero to 3 decimal places) and so we can say that the probability of these $t$ values (or larger) occurring if the values of $b$ in the population were zero is less than .001. Therefore, the $b$s are significantly different from 0. In the case of the $b$ for advertising budget this result means that the advertising budget makes a significant contribution ($p < .001$) to predicting album sales.

The bootstrap confidence interval tells us that the population value of $b$ for advertising budget is likely to fall between .08 and .11, and because this interval doesn't include zero we would conclude that there is a genuine positive relationship between advertising budget and album sales in the population. Also, the significance associated with this confidence interval is $p = .001$, which is highly significant. Also, note that the bootstrap process involves re-estimating the standard error (it changes from .01 in the original table to a bootstrap estimate of .009). This is a very small change. For the constant, the standard error is 7.537 compared to the bootstrap estimate of 8.214, which is a difference of 0.677. The bootstrap confidence intervals and significance values are useful to report and interpret because they do not rely on assumptions of normality or homoscedasticity.

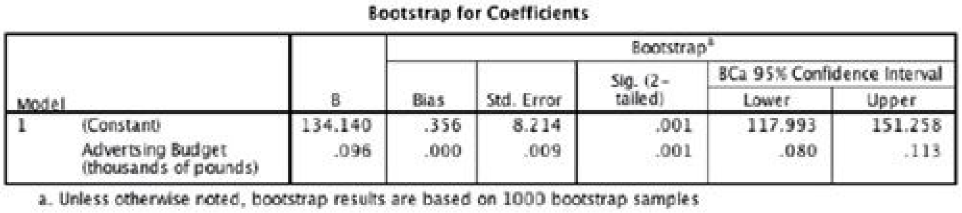
## OUTPUT 8.3



**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 134.140 | 7.537 | | 17.799 | .000 |
| | Advertsing Budget (thousands of pounds) | .096 | .010 | .578 | 9.979 | .000 |

a. Dependent Variable: Album Sales (thousands)



**Bootstrap for Coefficients**

| Model | | B | Bootstrap[a] | | | | |
|---|---|---|---|---|---|---|---|
| | | | Bias | Std. Error | Sig. (2-tailed) | BCa 95% Confidence Interval | |
| | | | | | | Lower | Upper |
| 1 | (Constant) | 134.140 | .356 | 8.214 | .001 | 117.993 | 151.258 |
| | Advertsing Budget (thousands of pounds) | .096 | .000 | .009 | .001 | .080 | .113 |

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

SELF-TEST How is the *t* in Output 8.3 calculated? Use the values in the table to see if you can get the same value as SPSS.

## 8.4.4. Using the model ①

So far, we have discovered that we have a useful model, one that significantly improves our ability to predict album 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 (8.1) with the values from the output. In addition, we can replace the *X* and *Y* with the variable names so that the model becomes:

$$\begin{aligned}\text{album sales}_i &= b_0 + b_1\text{advertising budget}_i \\ &= 134.14 + (0.096 \times \text{advertising budget}_i)\end{aligned} \tag{8.13}$$

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

$$\begin{aligned}\text{album sales}_i &= 134.14 + (0.096 \times \text{advertising budget}_i) \\ &= 134.14 + (0.096 \times 100) \\ &= 143.74\end{aligned} \tag{8.14}$$

## 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, which 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 SPSS table) then the predictor variable significantly predicts the outcome variable.

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

## 8.5. Multiple regression ②

Imagine that the record company executive was now interested in extending the model of albums sales to incorporate other variables. Before an album is released, the executive notes the amount spent on advertising, the number of times songs from the album are played on radio the week before release (**Airplay**), and the attractiveness of the band (**Attract**). He does this for 200 different albums (each made by a different band). Attractiveness was measured by asking a random sample of the target audience to rate the attractiveness of each band on a scale from 0 (hideous potato-heads) to 10 (gorgeous sex objects). The mode attractiveness given by the sample was used in the regression (because he was interested in what the majority of people thought, rather than the average of people's opinions).

When we want to build a model with several predictors, everything we have discussed so far still applies. It is important to remember that SPSS may appear to be very clever, but it is not. SPSS will happily generate output based on any garbage you decide to feed into it, it will not judge you or give any indication of whether the model is valid or generalizable. SPSS will provide the information necessary to assess these things, but we need to rely on our brains to evaluate the model – which is slightly worrying (especially if your brain is as small as mine).

The first thing to think about is what predictor variables to enter into the model. A great deal of care should be taken in selecting predictors for a model because the estimates of the regression coefficients depend upon the variables in the model. The predictors included and the way in which they are entered into the model can have a great impact. *Do not select hundreds of random predictors, bung them all into a regression analysis and hope for the best.* You should select predictors based on a sound theoretical rationale or well-conducted past research that has demonstrated their importance.[11] In our example, it seems logical that the band's image and radio play ought to affect sales, so these are sensible predictors. It would not be sensible to measure how much the album cost to make, because this won't affect sales directly: you would just be adding noise to the model. If predictors are being added that have never been looked at before (in your particular context) then select these new variables based on their substantive *theoretical* importance.

## 8.5.1. Methods of regression ②

In addition to the problem of selecting predictors, there are several ways in which variables can be entered into a model. When predictors are all completely uncorrelated, the order of variable entry has very little effect on the parameters calculated; however, we rarely have uncorrelated predictors and so the method of predictor selection is crucial.

### 8.5.1.1. Hierarchical (blockwise entry) ②

In **hierarchical regression** predictors are selected based on past work and the researcher decides in which order to enter the predictors into the model. As a general rule, known predictors (from other research) should be entered into the model first in order of their importance in predicting the outcome. After known predictors have been entered, the experimenter can add any new predictors into the model. New predictors can be entered either all in one go, in a stepwise manner, or hierarchically (such that the new predictor suspected to be the most important is entered first).

### 8.5.1.2. Forced entry ②

Forced entry (or *Enter* as it is known in SPSS) is a method in which all predictors are forced into the model simultaneously. Like hierarchical, this method relies on good theoretical reasons for including the chosen predictors, but unlike hierarchical the experimenter makes no decision about the order in which variables are entered. Some researchers believe that this method is the only appropriate method for theory testing (Studenmund & Cassidy, 1987) because stepwise techniques are influenced by random variation in the data and so seldom give replicable results if the model is retested.

### 8.5.1.3. Stepwise methods ②

Stepwise regressions are generally frowned upon by statisticians. Nevertheless, SPSS makes it very easy to do and actively encourages it in the *Automatic Linear Modelling* process (probably because this function is aimed at people who don't know better) – see Oditi's Lantern. I'm assuming that you wouldn't wade through 1000 pages of my drivel unless you wanted to know better, so we'll give stepwise a wide birth. However, you probably ought to know what it does so you can understand why to avoid it.

In **stepwise regressions** decisions about the order in which predictors are entered into the model are based on a purely mathematical criterion. In the *forward* method, an initial model is defined that contains only the constant ($b_0$). The computer then searches for the predictor (out of the ones available) that best predicts the outcome variable – it does this by selecting the predictor that has the highest simple correlation with the outcome. If this predictor significantly improves the ability of the model to predict the outcome, then this predictor is retained in the model and the computer searches for a second predictor. The criterion used for selecting this second predictor is that it is the variable that has the largest semi-partial correlation with the outcome. In plain English, imagine that the first predictor can explain 40% of the variation in the outcome variable; then there is still 60% left unexplained. The computer searches for the predictor that can explain the biggest part of the remaining 60% (it is not interested in the 40% that is already explained). As such, this semi-partial correlation gives a measure of how much 'new variance' in the outcome can be explained by each remaining predictor (see Section 7.5). The predictor that accounts for the most new variance is added to the model and, if it makes a significant contribution to the predictive power of the model, it is retained and another predictor is considered.

The *stepwise* method in SPSS is the same as the forward method, except that each time a predictor is added to the equation, a removal test is made of the least useful predictor. As such, the regression equation is being reassessed constantly to see whether any redundant predictors can be removed. The *backward* method is the opposite of the forward method in that the computer begins by placing all predictors in the model and then calculating the contribution of each one by looking at the significance value of the *t*-test for each predictor. This significance value is compared against a removal criterion (which can be either an absolute value of the test statistic or a probability value for that test statistic). If a predictor meets the removal criterion (i.e., if it is not making a statistically significant contribution to how well the model predicts the outcome variable) it is removed from the model and the model is re-estimated for the remaining predictors. The contribution of the remaining predictors is then reassessed.

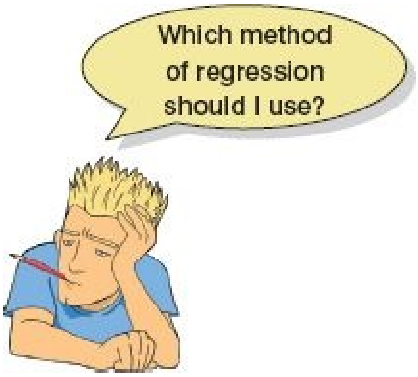
### ODITI'S LANTERN

*Automatic linear modelling*

'I, Oditi, come with a warning. Your desparation to bring me answers to numerical truths so as to gain a privileged place within my heart may lead you into the temptation that is SPSS's 'automatic linear modelling'. Automatic linear modelling promises answers without thought, and like a cat who is promised a fresh salmon, you will drool and purr in anticipation. If you want to find out more then stare into my lantern, but be warned, sometimes what looks like a juicy salmon is a rotting pilchard in disguise.'

## 8.5.1.4. Choosing a method ②

SPSS allows you to opt for any one of the methods described, and it is important to select an appropriate one. The short answer to which method to select is ‘not stepwise’, because stepwise methods rely on the computer selecting variables based upon mathematical criteria. Many writers argue that this takes many important methodological decisions out of the hands of the researcher. What’s more, the models derived by computer often take advantage of random sampling variation and so decisions about which variables should be included will be based upon slight differences in their semi-partial correlation. However, these slight statistical differences may contrast dramatically with the theoretical importance of a predictor to the model. There is also the danger of overfitting the model (having too many variables in the model that essentially make little contribution to predicting the outcome) and underfitting it (leaving out important predictors).



The main problem with stepwise methods is that they assess the fit of a variable based on the other variables in the model. Jeremy Miles (who has worked with me on other books) uses the analogy of getting dressed to describe this problem. You wake up in the morning and you need to get dressed: on your dressing table (or floor, if you’re me) you have underwear, some jeans, a T-shirt and jacket. Imagine these items are predictor variables. It’s a cold day and you’re trying to be warm. A stepwise method will put your trousers on first because they fit your goal best. It then looks around and tries the other clothes (variables). It tries to get you to put on your underwear but they won’t fit over your jeans, so it decides they are ‘a poor fit’ and discards them. You end up leaving the house without your underwear. Later on during a university seminar you stand up and your trousers fall down revealing your genitals to the room. It’s a mess. The problem is that the underwear was a poor fit only because when you tried to put them on you were already wearing jeans. In stepwise methods, variables might be considered bad predictors only because of what has already been put in the model.

For these reasons, stepwise methods are best avoided except for exploratory model building. If you do decide to use a stepwise method, then let the statistical blood be on your hands, not mine. Use the backward method rather than the forward method to minimize **suppressor effects**, which occur when a predictor has a significant effect but only when another variable is held constant. Forward selection is more likely than backward elimination to exclude predictors involved in suppressor effects. As such, the forward method runs a higher risk of making a Type II error (i.e., missing a predictor that does in fact predict the outcome). It is also advisable to cross-validate your model by splitting the data (see Section 8.3.2.2).

### 8.5.2. Comparing models ②

Hierarchical and (although obviously you’d never use them) stepwise methods involve adding predictors to the model in stages and it is, of course, useful to know whether these additions improve the model. Given that larger values of $R^2$ indicate better fit, a simple way to see whether a model has improved as a result of adding predictors to it would be to see whether $R^2$ for the new model is bigger than for the old model. In fact, it will always get bigger if we add predictors, so the issue is more whether it gets significantly bigger. We can assess the significance of the change in $R^2$ using equation (8.10), but because we're looking at the change in models we use the change in $R^2$ ($R^2_{\text{change}}$) and the $R^2$ of the newer model ($R^2_{\text{new}}$). We also use the change in the number of predictors ($k_{\text{change}}$) as well as the number of predictors in the new model ($k_{\text{new}}$). The equation is thus:

$$F_{\text{change}} = \frac{(N - k_{\text{new}} - 1)R^2_{\text{change}}}{k_{\text{change}}\left(1 - R^2_{\text{new}}\right)} \qquad (8.15)$$

We can compare models using this *F*-ratio. The problem with $R^2$ is that when you add more variables to the model, it will always go up. So, if you are deciding which of two models fits the data better, the model with more predictor variables in will always fit better. The **Akaike information criterion (AIC)**[12] is a measure of fit which penalizes the model for having more variables. If the AIC is bigger, the fit is worse; if the AIC is smaller, the fit is better. If you use the *Automated Linear Model* function in SPSS, then you can use the AIC to select models rather than the change in $R^2$. The AIC doesn't mean anything on its own: you cannot say that a value of the AIC of 10 is small, or that a value for the AIC of 1000 is large. The only thing you do with the AIC is compare it to other models with the same outcome variable: if it's getting smaller then the fit of your model is improving.

### 8.5.3. Multicollinearity ②

A final additional concern when we want to include more than one predictor in our model is multicollinearity, which exists when there is a strong correlation between two or more predictors. **Perfect collinearity** exists when at least one predictor is a perfect linear combination of the others (the simplest example being two predictors that are perfectly correlated – they have a correlation coefficient of 1). If there is perfect collinearity between predictors it becomes impossible to obtain unique estimates of the regression coefficients because there are an infinite number of combinations of coefficients that would work equally well. Put simply, if we have two predictors that are perfectly correlated, then the values of *b* for each variable are interchangeable. The good news is that perfect collinearity is rare in real-life data. The bad news is that less than perfect collinearity is virtually unavoidable. Low levels of collinearity pose little threat to the model estimates, but as collinearity increases there are three problems that arise:

- **Untrustworthy *b*s**: As collinearity increases so do the standard errors of the *b* coefficients. If you think back to what the standard error represents, then big standard errors for *b* coefficients means that these *b*s are more variable across samples. Therefore, the *b* coefficient in our sample is less likely to represent the population. Crudely put, multicollinearity means that the *b*-values are less trustworthy. Don't lend them money and don't let them go out to dinner with your boy- or girlfriend. Of course, if the *b*s are variable from sample to sample then the resulting predictor equations will be unstable across samples too.

- **It limits the size of *R***: Remember that $R$ is a measure of the correlation between the predicted values of the outcome and the observed values and that $R^2$ indicates the variance in the outcome for which the model accounts. Imagine a situation in which a single variable predicts the outcome variable fairly successfully (e.g., $R$ = .80) and a second predictor variable is then added to the model. This second variable might account for a lot of the variance in the outcome (which is why it is included in the model), but the variance it accounts for is the same variance accounted for by the first variable. In other words, once the variance accounted for by the first predictor has been removed, the second predictor accounts for very little of the remaining variance (the second variable accounts for very little *unique variance*). Hence, the overall variance in the outcome accounted for by the two predictors is little more than when only one predictor is used (so $R$ might increase from .80 to .82). This idea is connected to the notion of partial correlation that was explained in Chapter 7. If, however, the two predictors are completely uncorrelated, then the second predictor is likely to account for different variance in the outcome than that accounted for by the first predictor. So, although in itself the second predictor might account for only a little of the variance in the outcome, the variance it does account for is different to that of the other predictor (and so when both predictors are included, $R$ is substantially larger, say .95). Therefore, having uncorrelated predictors is beneficial.
- **Importance of predictors**: Multicollinearity between predictors makes it difficult to assess the individual importance of a predictor. If the predictors are highly correlated, and each accounts for similar variance in the outcome, then how can we know which of the two variables is important? Quite simply, we can't – the model could include either one, interchangeably.

One way of identifying multicollinearity is to scan a correlation matrix of the predictor variables and see if any correlate very highly (by 'very highly' I mean correlations of above .80 or .90). This is a good 'ball park' method, but misses more subtle forms of multicollinearity. Luckily, SPSS produces various collinearity diagnostics, one of which is the **variance inflation factor (VIF)**. The VIF indicates whether a predictor has a strong linear relationship with the other predictor(s). Related to the VIF is the **tolerance** statistic, which is its reciprocal (1/VIF). Although there are no hard and fast rules about what value of the VIF should cause concern, there are some general guidelines:

- If the largest VIF is greater than 10 then there is cause for concern (Bowerman & O'Connell, 1990; Myers, 1990).
- If the average VIF is substantially greater than 1 then the regression may be biased (Bowerman & O'Connell, 1990).
- Tolerance below 0.1 indicates a serious problem.
- Tolerance below 0.2 indicates a potential problem (Menard, 1995).

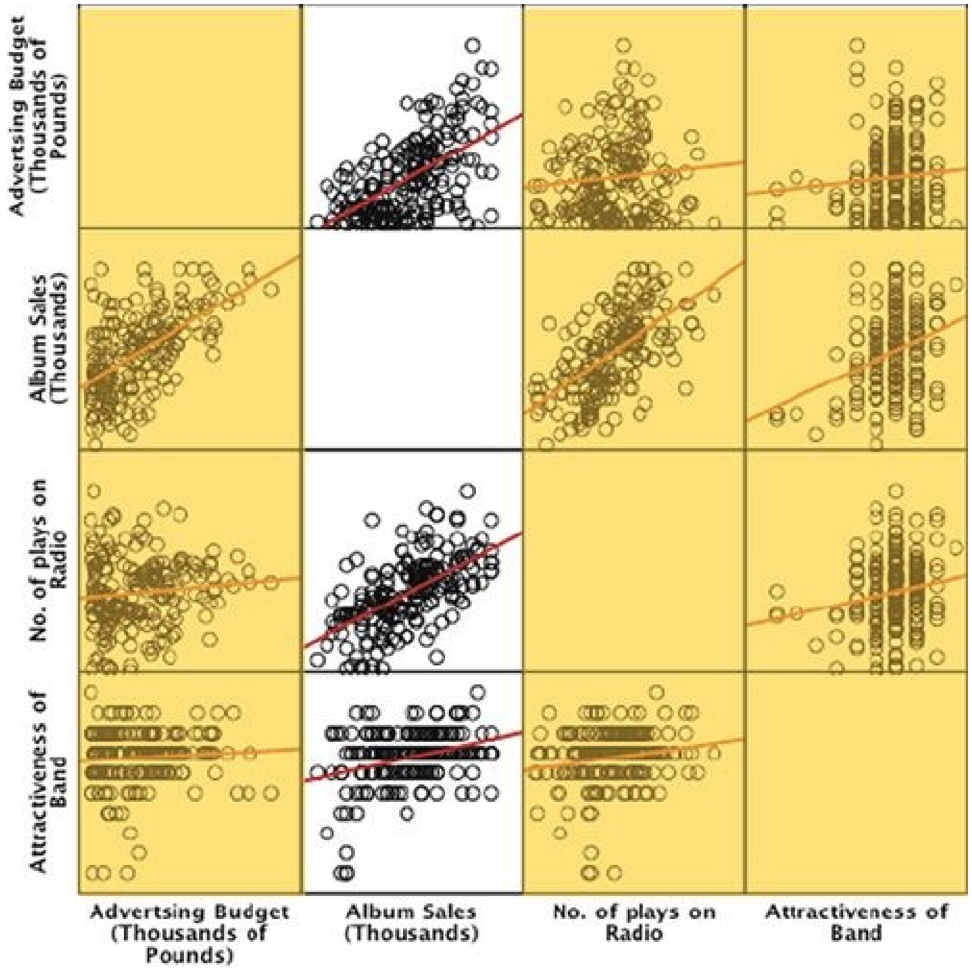
Other measures that are useful in discovering whether predictors are dependent are the *eigenvalues of the scaled, uncentred cross-products matrix*, the *condition indexes* and the *variance proportions*. These statistics are extremely complex and will be covered as part of the interpretation of SPSS output (see Section 8.7.5). If none of this has made any sense then have a look at Hutcheson and Sofroniou (1999, pp. 78–85) who give a really clear explanation of multicollinearity.

## 8.6. Regression with several predictors using SPSS ②

Remember that for any regression we need to follow the general procedure outlined in Figure 8.11. So, first we might look at some scatterplots of the relationships between the outcome variable and the predictors. The resulting scatterplots for our album sales data are shown in Figure 8.14. We need to focus on the relationship between predictors and the outcome (album sales), and in Figure 8.14 I have shaded out the other scatterplots so we can focus on the three related to album sales.[13] We can see that although the data are messy in places, the three predictors have reasonably linear relationships with the outcome (album sales) and there are no obvious outliers.

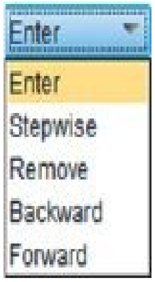
**FIGURE 8.14**

Matrix scatterplot of the relationships between advertising budget, airplay, and attractiveness of the band and album sales



SELF-TEST Produce a matrix scatterplot of **Sales Adverts, Airplay** and **Attract** including the regression line.

## 8.6.1. Main options ②

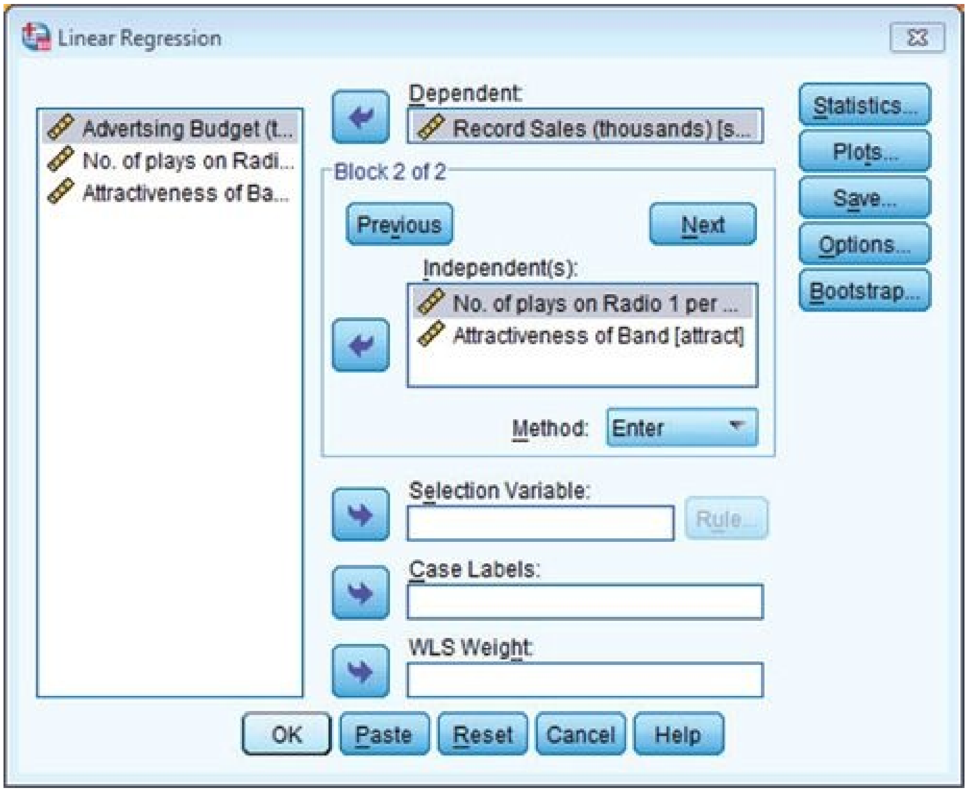


The executive has past research indicating that advertising budget is a significant predictor of album sales, and so he should include this variable in the model first. His new variables (**Airplay** and **Attract**) should, therefore, be entered into the model *after* advertising budget. This method is hierarchical (the researcher decides in which order to enter variables into the model based on past research). To do a hierarchical regression in SPSS we have to enter the variables in blocks (each block representing one step in the hierarchy). To get to the main *Regression* dialog box select Analyze Regression ▸ Linear... . We encountered this dialog box in Figure 8.13 when we looked at a model with only one predictor. Essentially, to set up the first block we do exactly what we did before. Select the outcome variable (album sales) and drag it to the box labelled *Dependent* (or click on ➡). We also need to specify the predictor variable for the first block. We've decided that advertising budget should be entered into the model first, so select this variable in the list and drag it to the box labelled *Independent(s)* (or click on ➡). Underneath the *Independent(s)* box, there is a drop-down menu for specifying the *Method* of regression (see Section 8.5.1). You can select a different method of variable entry for each block by clicking on Enter ▾, next to where it says *Method*. The default option is forced entry, and this is the option we want, but if you were carrying out more exploratory work, you might decide to use one of the stepwise methods (forward, backward, stepwise or remove).

Having specified the first block in the hierarchy, we need to move onto the second. To tell the computer that you want to specify a new block of predictors you must click on Next. This process clears the *Independent(s)* box so that you can enter the new predictors (you should also note that above this box it now reads *Block 2 of 2* indicating that you are in the second block of the two that you have so far specified). We decided that the second block would contain both of the new predictors and so you should click on **Airplay** and **Attract** (while holding down *Ctrl*, or *Cmd* if you use a Mac) in the variables list and drag them to the *Independent(s)* box or click on ➡. The dialog box should now look like Figure 8.15. To move between blocks use the Previous and Next buttons (so for example, to move back to block 1, click on Previous).

**FIGURE 8.15**

Main dialog box for block 2 of the multiple regression



Dependent: Record Sales (thousands) [s...

Block 2 of 2

Independent(s): No. of plays on Radio 1 per ... ; Attractiveness of Band [attract]

Method: Enter

It is possible to select different methods of variable entry for different blocks in a hierarchy. So although we specified forced entry for the first block, we could now specify a stepwise method for the second. Given that we have no previous research regarding the effects of attractiveness and airplay on album sales, we might be justified in requesting a stepwise method for this block. However, because of the problems with stepwise methods, I am going to stick with forced entry for both blocks in this example.

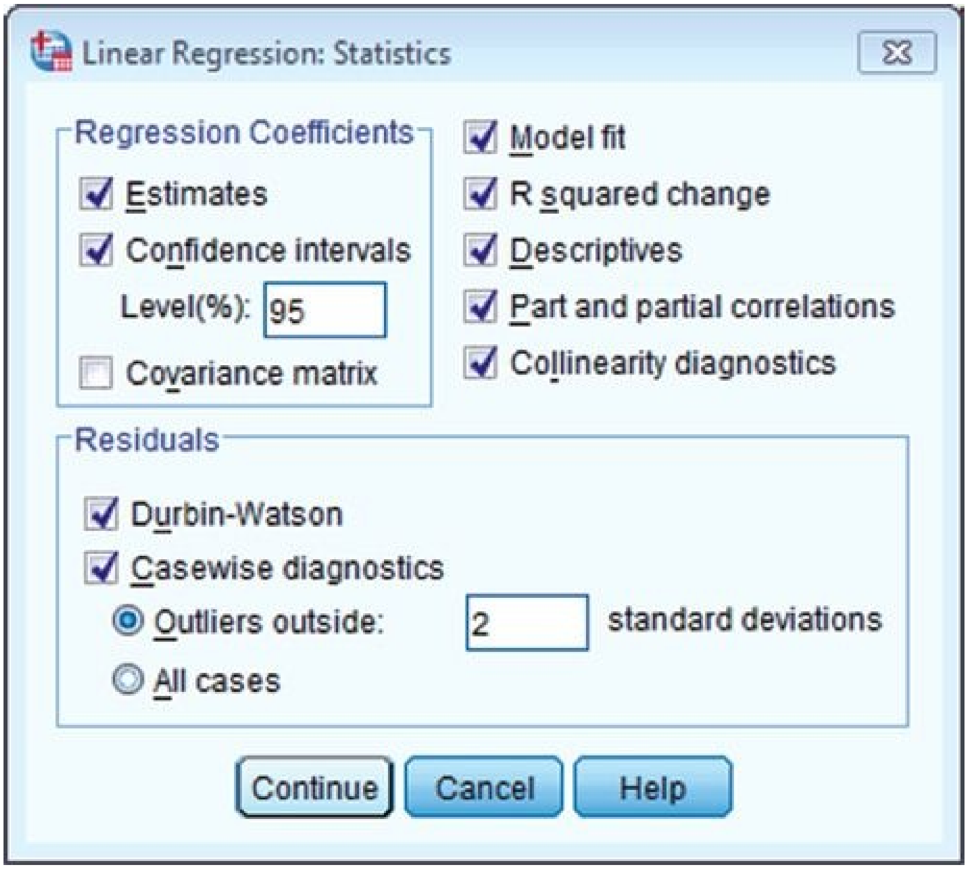
## 8.6.2. Statistics ②

In the main *Regression* dialog box click on Statistics... to open a dialog box for selecting various important options relating to the model (see list below and Figure 8.16). Most of these options relate to the parameters of the model; however, there are procedures available for checking the assumptions of no multicollinearity (collinearity diagnostics) and independence of errors (Durbin–Watson). When you have selected the statistics you require (I recommend all but the covariance matrix as a general rule), click on Continue to return to the main dialog box.

- *Estimates*: This option is selected by default because it gives us the estimated coefficients of the regression model (i.e., the estimated *b*-values). Test statistics and their significance are produced for each regression coefficient: a *t*-test is used to see whether each *b* differs significantly from zero (see Section 8.2.5).
- *Confidence intervals*: This option produces confidence intervals for each of the unstandardized regression coefficients. Remember that if the assumptions of regression are not met these confidence intervals will be inaccurate and we should use bootstrap confidence intervals instead.
- *Covariance matrix*: This option produces a matrix of the covariances, correlation coefficients and variances between the regression coefficients of each variable in the model. A variance–covariance matrix is produced with variances displayed along the diagonal and covariances displayed as off-

diagonal elements. The correlations are produced in a separate matrix.

FIGURE 8.16
*Statistics* dialog box for regression analysis



Regression Coefficients
☑ Estimates
☑ Confidence intervals
Level(%): 95
□ Covariance matrix

☑ Model fit
☑ R squared change
☑ Descriptives
☑ Part and partial correlations
☑ Collinearity diagnostics

Residuals
☑ Durbin-Watson
☑ Casewise diagnostics
● Outliers outside: 2 standard deviations
○ All cases

- *Model fit*: This option is vital and so is selected by default. It provides not only a statistical test of the model's ability to predict the outcome variable (the *F*-test described in Section 8.2.4), but also the value of *R*, the corresponding $R^2$ and the adjusted $R^2$.
- *R squared change*: This option displays the change in $R^2$ resulting from the inclusion of a new predictor (or block of predictors). This measure is a useful way to assess the contribution of new predictors (or blocks) to explaining variance in the outcome.
- *Descriptives*: If selected, this option displays a table of the mean, standard deviation and number of observations of all of the variables included in the analysis. A correlation matrix is also displayed showing the correlation between all of the variables and the one-tailed probability for each correlation coefficient. This option is extremely useful because the correlation matrix can be used to assess whether there is multicollinearity.
- *Part and partial correlations*: This option produces the zero-order correlation (the Pearson correlation) between each predictor and the outcome variable. It also produces the partial correlation between each predictor and the outcome, controlling for all other predictors in the model. Finally, it produces the part correlation (or semi-partial correlation) between each predictor and the outcome. This correlation represents the relationship between each predictor and the part of the outcome that is not explained by the other predictors in the model. As such, it measures the unique relationship between a predictor and the outcome (see Section 7.5).
- *Collinearity diagnostics*: This option is for obtaining collinearity statistics such as the VIF, tolerance, eigenvalues of the scaled, uncentred cross-products matrix, condition indexes and variance proportions (see Section 8.5.3).

- *Durbin-Watson*: This option produces the Durbin–Watson test statistic, which tests the assumption of independent errors. Unfortunately, SPSS does not provide the significance value of this test, so you must decide for yourself whether the value is different enough from 2 to be cause for concern (see Section 8.3.2.1).
- *Casewise diagnostics*: This option, if selected, lists the observed value of the outcome, the predicted value of the outcome, the difference between these values (the residual) and this difference standardized. Furthermore, it will list these values either for all cases, or just for cases for which the standardized residual is greater than 3 (when the ± sign is ignored). This criterion value of 3 can be changed, and I recommend changing it to 2 for reasons that will become apparent. A summary table of residual statistics indicating the minimum, maximum, mean and standard deviation of both the values predicted by the model and the residuals (see Section 8.6.4) is also produced.

### 8.6.3. Regression plots ②

Once you are back in the main dialog box, click on Plots... to activate the regression *Plots* dialog box shown in Figure 8.17. This dialog box provides the means to specify several graphs, which can help to establish the validity of some regression assumptions. Most of these plots involve various *residual* values, which will be described in more detail in Section 8.6.4.

On the left-hand side of the dialog box is a list of several variables.

- **DEPENDNT** (the outcome variable).
- ***ZPRED** (the standardized predicted values of the dependent variable based on the model). These values are standardized forms of the values predicted by the model.
- ***ZRESID** (the standardized residuals, or errors). These values are the standardized differences between the observed data and the values that the model predicts).
- ***DRESID** (the deleted residuals). See Section 8.3.1.1 for details.
- ***ADJPRED** (the adjusted predicted values). See Section 8.3.1.1 for details.
- ***SRESID** (the Studentized residual). See Section 8.3.1.1 for details.
- ***SDRESID** (the Studentized deleted residual). This value is the deleted residual divided by its standard error.

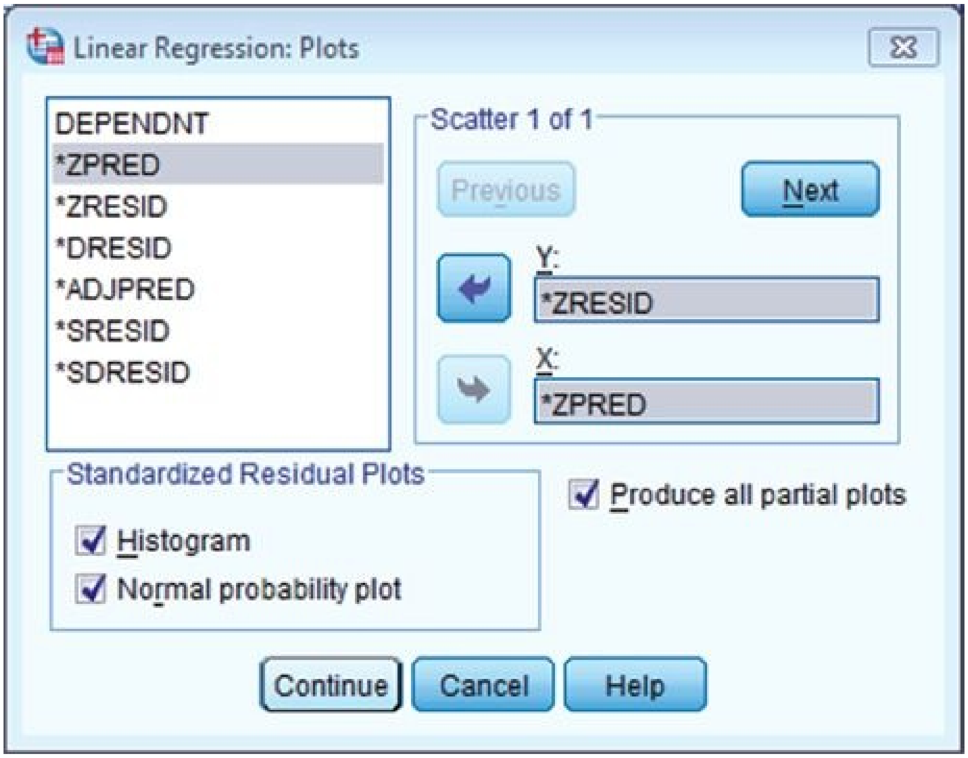
The variables listed in this dialog box all come under the general heading of residuals. In Section 5.3.3.1 we saw that a plot of *ZRESID (*y*-axis) against *ZPRED (*x*-axis) is useful for testing the assumptions of independent errors, homoscedasticity and linearity. A plot of *SRESID (*y*-axis) against *ZPRED (*x*-axis) will show up any heteroscedasticity also. Although often these two plots are virtually identical, the latter is more sensitive on a case-by-case basis. To create these plots simply select a variable from the list, and transfer it to the space labelled either *X* or *Y* (which refer to the axes) by clicking on ➡. When you have selected two variables for the first plot (as is the case in Figure 8.17) you can specify a new plot by clicking on Next. This process clears the spaces in which variables are specified. If you click on Next and would like to return to the plot that you last specified, then simply click on Previous. You can specify up to nine plots.

You can also tick the box labelled *Produce all partial plots* which will produce scatterplots of the residuals of the outcome variable and each of the predictors when both variables are regressed separately on the remaining predictors. Regardless of whether the previous sentence made any sense to you, these plots have several important characteristics that make them worth inspecting. First, the gradient of the regression line between the two residual variables is equivalent to the coefficient of the predictor in the regression equation. As such, any obvious outliers on a partial plot represent cases that might have undue influence on a predictor's regression coefficient. Second, non-linear relationships between a predictor and the outcome variable are much more detectable using these plots. Finally, they are a useful way of detecting collinearity. For these reasons, I recommend requesting them.

There are several options for plots of the standardized residuals. First, you can select a *Histogram* of the standardized residuals (this is useful for checking the assumption of normality of errors). Second, you can ask for a *Normal probability plot*, which also provides information about whether the residuals in the model are normally distributed. When you have selected the options you require, click on Continue to take you back to the main *Regression* dialog box.

**FIGURE 8.17**

The *Plots* dialog box



Y: *ZRESID

X: *ZPRED

☑ Histogram

☑ Normal probability plot

☑ Produce all partial plots

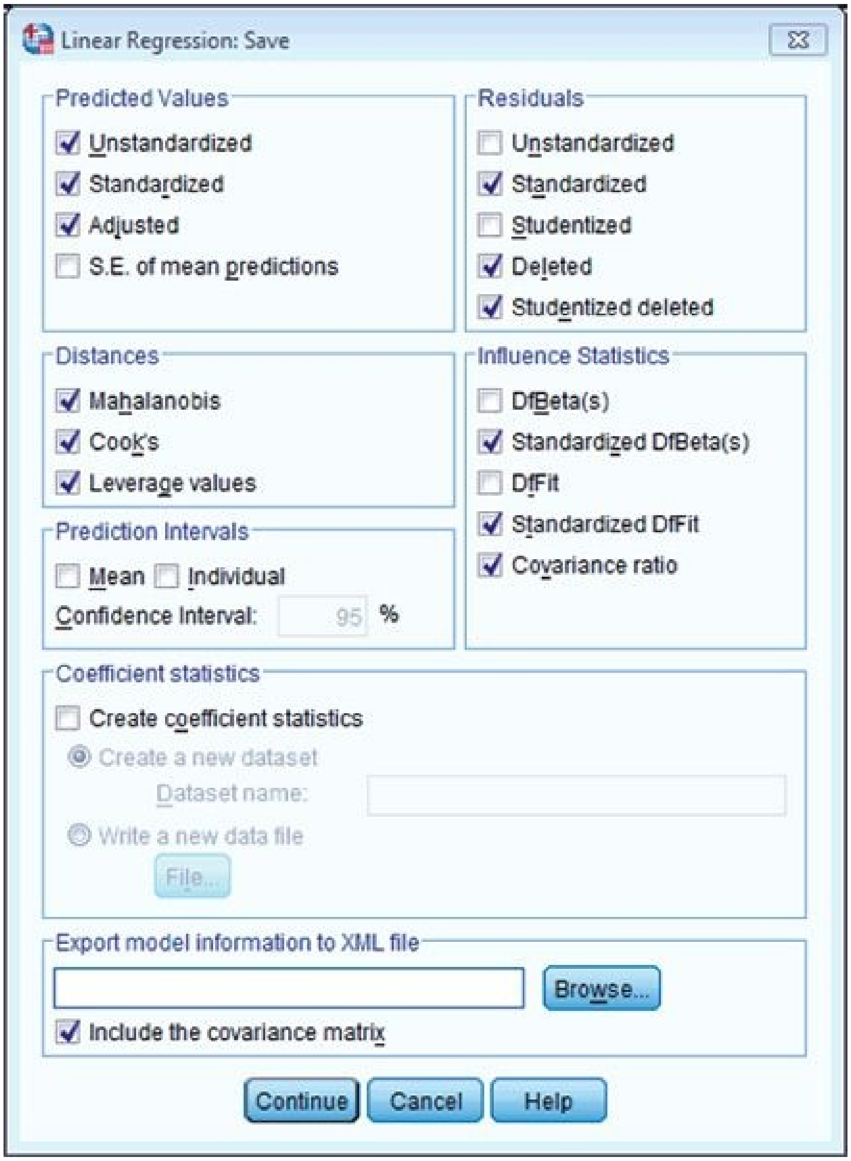
## 8.6.4. Saving regression diagnostics ②

In Section 8.3 we met two types of regression diagnostics: those that help us assess how well our model fits our sample and those that help us detect cases that have a large influence on the model generated. In SPSS we can choose to save these diagnostic variables in the data editor (so SPSS will calculate them and then create new columns in the data editor in which the values are placed).

To save regression diagnostics you need to click on Save in the main *Regression* dialog box. This process activates the *Save* new variables dialog box (see Figure 8.18). Once this dialog box is active, it is a simple matter to tick the boxes next to the required statistics. Most of the available options were explained in Section 8.3, and Figure 8.18 shows what I consider to be a fairly basic set of diagnostic statistics. Standardized (and Studentized) versions of these diagnostics are generally easier to interpret, so I suggest selecting them in preference to the unstandardized versions. Once the regression has been run, SPSS creates a column in your data editor for each statistic requested and it has a standard set of variable names to describe each one. After the name, there will be a number that refers to the analysis that has been run. So, for the first regression run on a data set the variable names will be followed by a 1, if you carry out a second regression it will create a new set of variables with names followed by a 2, and so on. The names of the variables that will be created are below. When you have selected the diagnostics you require (by clicking in the appropriate boxes), click on Continue to return to the main *Regression* dialog box.

- **pre_1**: unstandardized predicted value;
- **zpr_1**: standardized predicted value;
- **adj_1**: adjusted predicted value;
- **sep_1**: standard error of predicted value;
- **res_1**: unstandardized residual;
- **zre_1**: standardized residual;
- **sre_1**: Studentized residual;
- **dre_1**: deleted residual;
- **sdr_1**: Studentized deleted residual;
- **mah_1**: Mahalanobis distance;
- **coo_1**: Cook’s distance;
- **lev_1**: centred leverage value;
- **sdb0_1**: standardized DFBETA (intercept);
- **sdb1_1**: standardized DFBETA (predictor 1);
- **sdb2_1**: standardized DFBETA (predictor 2);
- **sdf_1**: standardized DFFIT;
- **cov_1**: covariance ratio.

**FIGURE 8.18**
Dialog box for regression diagnostics


Linear Regression: Save

Predicted Values
- ☑ Unstandardized
- ☑ Standardized
- ☑ Adjusted
- □ S.E. of mean predictions

Residuals
- □ Unstandardized
- ☑ Standardized
- □ Studentized
- ☑ Deleted
- ☑ Studentized deleted

Distances
- ☑ Mahalanobis
- ☑ Cook's
- ☑ Leverage values

Influence Statistics
- □ DfBeta(s)
- ☑ Standardized DfBeta(s)
- □ DfFit
- ☑ Standardized DfFit
- ☑ Covariance ratio

Prediction Intervals
- □ Mean □ Individual
- Confidence Interval: 95 %

Coefficient statistics
- □ Create coefficient statistics
- ● Create a new dataset
  - Dataset name:
- ○ Write a new data file
  - File...

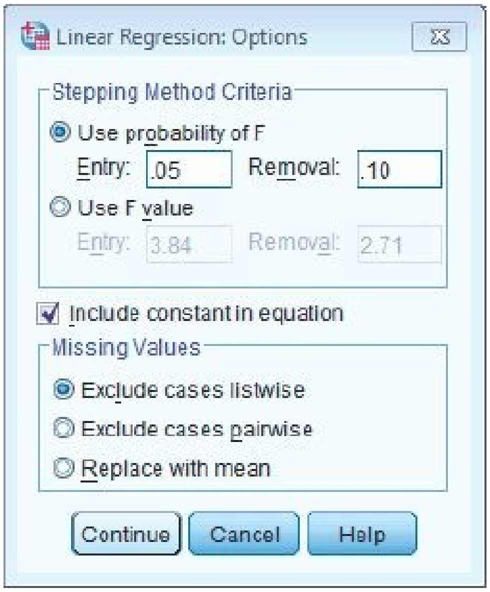
Export model information to XML file
- Browse...
- ☑ Include the covariance matrix

## 8.6.5. Further options ②

You can click on Options... to take you to the *Options* dialog box (Figure 8.19). The first set of options allows you to change the criteria used for entering variables in a stepwise regression. If you insist on doing stepwise regression, then it's probably best that you leave the default criterion of .05 probability for entry alone. However, you can make this criterion more stringent (.01). There is also the option to build a model that doesn't include a constant (i.e., has no *Y* intercept). This option should also be left alone. Finally, you can select a method for dealing with missing data points (see SPSS Tip 5.1). By default, SPSS excludes cases listwise, which in regression means that if a person has a missing value for any variable, then they are excluded from the whole analysis. So, for example, if our record company executive didn't have an attractiveness score for one of his bands, their data would not be used in the regression model. Another option is to exclude cases on a pairwise basis, which means that if a participant has a score missing for a particular variable, then their data are excluded only from calculations involving the variable for which they have no score. So, data for the band for which there was no attractiveness rating would still be used to calculate the relationships between advertising budget, airplay and album sales. However, if you do this, many of your variables may not make sense, and you can end up with absurdities such as $R^2$ either negative or greater than 1.0. So it's not a good option.

Another possibility is to replace the missing score with the average score for this variable and then include that case in the analysis (so our example band would be given an attractiveness rating equal to the average attractiveness of all bands). The problem with this final choice is that it is likely to suppress the true value of the standard deviation (and, more importantly, the standard error). The standard deviation will be suppressed because for any replaced case there will be no difference between the mean and the score, whereas if data had been collected for that case there would, almost certainly, have been some difference between the score and the mean. Obviously, if the sample is large and the number of missing values small then this is not a serious consideration. However, if there are many missing values this choice is potentially dangerous because smaller standard errors are more likely to lead to significant results that are a product of the data replacement rather than a genuine effect. The final option is to use the *Missing Value Analysis* routine in SPSS. This is for experts. It makes use of the fact that if two or more variables are present and correlated for most cases in the file, and an occasional value is missing, you can replace the missing values with estimates far better than the mean (some of these features are described in Tabachnick & Fidell, 2012, Chapter 4).

**FIGURE 8.19**
Options for linear regression



Linear Regression: Options

Stepping Method Criteria

● Use probability of F

Entry: .05 Removal: .10

○ Use F value

Entry: 3.84 Removal: 2.71

☑ Include constant in equation

Missing Values

● Exclude cases listwise

○ Exclude cases pairwise

○ Replace with mean

## 8.6.6. Robust regression ②

We can get bootstrapped confidence intervals for the regression coefficients by clicking on Bootstrap... (see Section 5.4.3). However, this function doesn't work when we have used the Save... option to save residuals, so we can't use it now. We will return to robust regression in Section 8.8.

### ODITI’S LANTERN

*Regression*

‘I, Oditi, wish to predict when I can take over the world, and rule you pathetic mortals with will of pure iron … erm.. ahem, I mean, I wish to predict how to save cute kittens from the jaws of rabid dogs, because I’m nice like that, and have no aspirations to take over the world. This chapter is so long that some of you will die before you reach the end, so ignore the author’s bumbling drivel and stare instead into my lantern of wonderment.’

## 8.7. Interpreting multiple regression ②

Having selected all of the relevant options and returned to the main dialog box, we need to click on OK to run the analysis. SPSS will spew out copious amounts of output in the viewer window, and we now turn to look at how to make sense of this information.

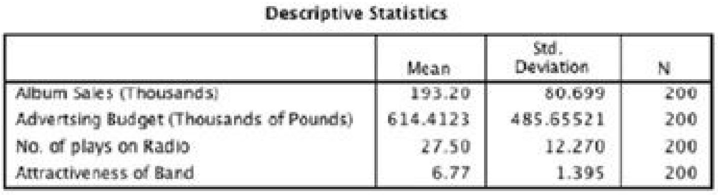
### 8.7.1. Descriptives ②

The output described in this section is produced using the options in the *Statistics* dialog box (see Figure 8.16). To begin with, if you selected the *Descriptives* option, SPSS will produce the table seen in Output 8.4. This table tells us the mean and standard deviation of each variable in our data set, so we now know that the average number of album sales was 193,200. This table isn’t necessary for interpreting the regression model, but it is a useful summary of the data. In addition to the descriptive statistics, selecting this option produces a correlation matrix. This table shows three things. First, it shows the value of Pearson’s correlation coefficient between every pair of variables (e.g., we can see that the advertising budget had a large positive correlation with album sales, $r$ = .578). Second, the one-tailed significance of each correlation is displayed (e.g., the correlation above is significant, $p$ < .001). Finally, the number of cases contributing to each correlation ($N$ = 200) is shown.

You might notice that along the diagonal of the matrix the values for the correlation coefficients are all 1.00 (i.e., a perfect positive correlation). The reason for this is that these values represent the correlation of each variable with itself, so obviously the resulting values are 1. The correlation matrix is extremely useful for getting a rough idea of the relationships between predictors and the outcome, and for a preliminary look for multicollinearity. If there is no multicollinearity in the data then there should be no substantial correlations ($r$ > .9) between predictors.

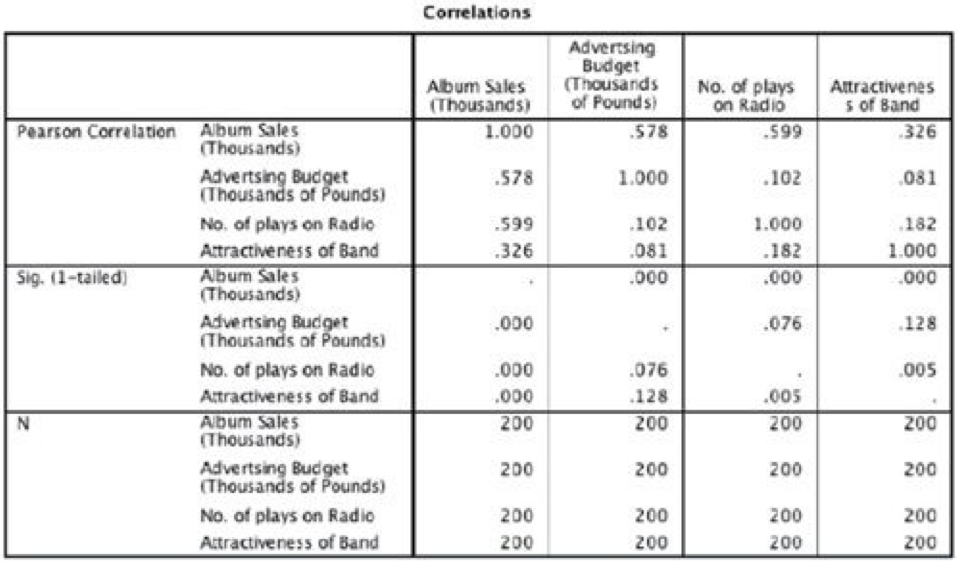
**OUTPUT 8.4**

Descriptive statistics for regression analysis



**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| Album Sales (Thousands) | 193.20 | 80.699 | 200 |
| Advertsing Budget (Thousands of Pounds) | 614.4123 | 485.65521 | 200 |
| No. of plays on Radio | 27.50 | 12.270 | 200 |
| Attractiveness of Band | 6.77 | 1.395 | 200 |



**Correlations**

| | | Album Sales (Thousands) | Advertsing Budget (Thousands of Pounds) | No. of plays on Radio | Attractiveness of Band |
|---|---|---|---|---|---|
| Pearson Correlation | Album Sales (Thousands) | 1.000 | .578 | .599 | .326 |
| | Advertsing Budget (Thousands of Pounds) | .578 | 1.000 | .102 | .081 |
| | No. of plays on Radio | .599 | .102 | 1.000 | .182 |
| | Attractiveness of Band | .326 | .081 | .182 | 1.000 |
| Sig. (1-tailed) | Album Sales (Thousands) | . | .000 | .000 | .000 |
| | Advertsing Budget (Thousands of Pounds) | .000 | . | .076 | .128 |
| | No. of plays on Radio | .000 | .076 | . | .005 |
| | Attractiveness of Band | .000 | .128 | .005 | . |
| N | Album Sales (Thousands) | 200 | 200 | 200 | 200 |
| | Advertsing Budget (Thousands of Pounds) | 200 | 200 | 200 | 200 |
| | No. of plays on Radio | 200 | 200 | 200 | 200 |
| | Attractiveness of Band | 200 | 200 | 200 | 200 |

If we look only at the predictors (ignore album sales) then the highest correlation is between the attractiveness of the band and the amount of airplay, which is significant at a .01 level (*r* = .182, *p* = .005). Despite the significance of this correlation, the coefficient is small and so it looks as though our predictors are measuring different things (there is no collinearity). We can see also that of all of the predictors the number of plays on radio correlates best with the outcome (*r* = .599, *p* < .001) and so it is likely that this variable will best predict album sales.

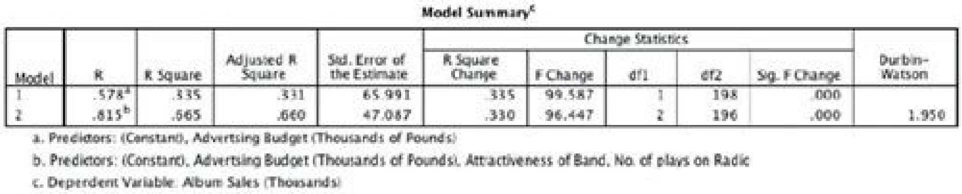
**CRAMMING SAM'S TIPS** **Descriptive statistics**

- Use the descriptive statistics to check the correlation matrix for multicollinearity – that is, predictors that correlate too highly with each other, r > .9.

## 8.7.2. Summary of model ②

The next section of output describes the overall model (so it tells us whether the model is successful in predicting album sales). Remember that we chose a hierarchical method and so each set of summary statistics is repeated for each stage in the hierarchy. In Output 8.5 you should note that there are two models. Model 1 refers to the first stage in the hierarchy when only advertising budget is used as a predictor. Model 2 refers to when all three predictors are used. Output 8.5 is the *model summary* and this table was produced using the *Model fit* option. This option is selected by default in SPSS because it provides us with some very important information about the model: the values of $R$, $R^2$ and the adjusted $R^2$. If the *R squared change* and *Durbin-Watson* options were selected, then these values are included also (if they weren't selected you'll find that you have a smaller table).

Under the model summary table shown in Output 8.5 you should notice that SPSS tells us what the dependent variable (outcome) was and what the predictors were in each of the two models. In the column labelled $R$ are the values of the multiple correlation coefficient between the predictors and the outcome. When only advertising budget is used as a predictor, this is the simple correlation between advertising and album sales (.578). In fact all of the statistics for model 1 are the same as the simple regression model earlier (see Section 8.4.3). The next column gives us a value of $R^2$, which we already know is a measure of how much of the variability in the outcome is accounted for by the predictors. For the first model its value is .335, which means that advertising budget accounts for 33.5% of the variation in album sales. However, when the other two predictors are included as well (model 2), this value increases to .665 or 66.5% of the variance in album sales. Therefore, if advertising accounts for 33.5%, we can tell that attractiveness and radio play account for an additional 33%.[14] So, the inclusion of the two new predictors has explained quite a large amount of the variation in album sales.



Model Summary$^c$

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics |  |  |  |  | Durbin-Watson |
|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  | R Square Change | F Change | df1 | df2 | Sig. F Change |  |
| 1 | .578$^a$ | .335 | .331 | 65.991 | .335 | 99.587 | 1 | 198 | .000 |  |
| 2 | .815$^b$ | .665 | .660 | 47.087 | .330 | 96.447 | 2 | 196 | .000 | 1.950 |

a. Predictors: (Constant), Advertising Budget (Thousands of Pounds)

b. Predictors: (Constant), Advertising Budget (Thousands of Pounds), Attractiveness of Band, No. of plays on Radio

c. Dependent Variable: Album Sales (Thousands)

**OUTPUT 8.5** Regression model summary

The adjusted $R^2$ gives us some idea of how well our model generalizes and ideally we would like its value to be the same as, or very close to, the value of $R^2$. In this example the difference for the final model is small (in fact the difference between the values is .665 − .660 = .005 or 0.5%). This shrinkage means that if the model were derived from the population rather than a sample it would account for approximately 0.5% less variance in the outcome. If you apply Stein's formula you'll get an adjusted value of .653 (Jane Superbrain Box 8.2), which is very close to the observed value of $R^2$ (.665) indicating that the cross-validity of this model is very good.

## JANE SUPERBRAIN 8.2

### *Maths frenzy* ③

We can have a look at how some of the values in the output are computed by thinking back to the theory part of the chapter. For example, looking at the change in $R^2$ for the first model, we have only one predictor (so $k = 1$) and 200 cases ($N = 200$), so the $F$ comes from equation (8.10):[15]

$$F_{\text{Model1}} = \frac{(200-1-1)0.334648}{1(1-0.334648)} = 99.59$$

In model 2 in Output 8.5 two predictors have been added (attractiveness and radio play), so the new model has 3 predictors ($k_{\text{new}}$) and the previous model had only 1, which is a change of 2 ($k$change). The addition of these two predictors increases $R^2$ by .330 ($R^2$ change), making the $R^2$ of the new model .665 ($R^2_{\text{new}}$).[16] The $F$-ratio for this change comes from equation (8.15):

$$F_{\text{change}} = \frac{(N-3-1)0.33}{2(1-0.664668)} = 96.44$$

We can also apply Stein's formula (equation (8.12)) to $R^2$ to get some idea of its likely value in different samples. We replace $n$ with the sample size (200) and $k$ with the number of predictors (3):

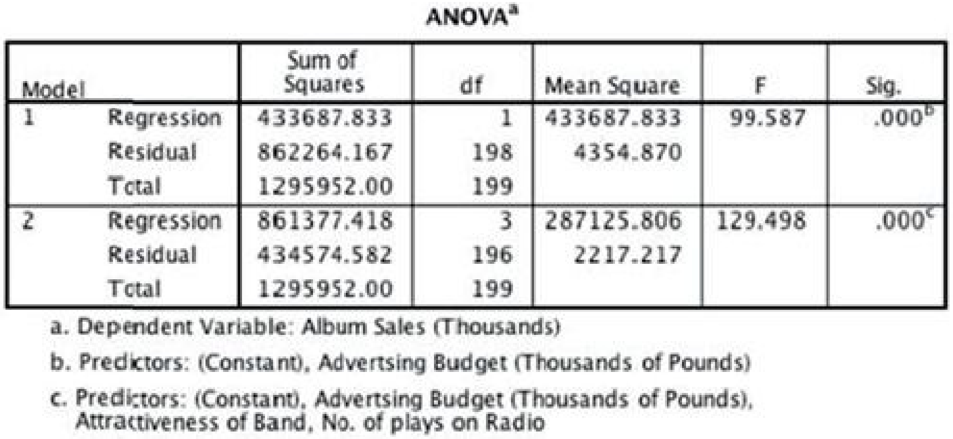
$$\begin{aligned}
\text{adjusted } R^2 &= 1 - \left[\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\right](1-R^2) \\
&= 1 - \left[(1.015)(1.015)(1.005)\right](0.335) \\
&= 1 - 0.347 \\
&= .653
\end{aligned}$$

The change statistics are provided only if requested, and these tell us whether the change in $R^2$ is significant. In Output 8.5, the change is reported for each block of the hierarchy. So, model 1 causes $R^2$ to change from 0 to .335, and this change in the amount of variance explained gives rise to an $F$-ratio of 99.59, which is significant with a probability less than .001. In model 2, in which attractiveness and radio play have been added as predictors, $R^2$ increases by .330, making the $R^2$ of the new model .665. This increase yields an $F$-ratio of 96.44 (Jane Superbrain Box 8.2), which is significant ($p < .001$). The change statistics therefore tell us about the difference made by adding new predictors to the model.

Finally, if you requested the Durbin–Watson statistic it will be found in the last column of the table in Output 8.5. This statistic informs us about whether the assumption of independent errors is tenable (see Section 8.3.2.1). As a conservative rule I suggested that values less than 1 or greater than 3 should definitely raise alarm bells (although I urge you to look up precise values for the situation of interest). The closer to 2 that the value is, the better, and for these data the value is 1.950, which is so close to 2 that the assumption has almost certainly been met.

Output 8.6 shows the next part of the output, which contains an ANOVA that tests whether the model is significantly better at predicting the outcome than using the mean as a 'best guess'. Specifically, the $F$-ratio represents the ratio of the improvement in prediction that results from fitting the model, relative to the inaccuracy that still exists in the model (see Section 8.2.4). This table is again split into two sections, one for each model. We are told the value of the sum of squares for the model (this value is $SS_M$ in Section 8.2.4 and represents the improvement in prediction resulting from fitting a regression line to the data rather than using the mean as an estimate of the outcome). We are also told the residual sum of squares (this value is $SS_R$ in Section 8.2.4 and represents the total difference between the model and the observed data). We are also told the degrees of freedom (*df*) for each term. In the case of the improvement due to the model, this value is equal to the number of predictors (1 for the first model and 3 for the second), and for $SS_R$ it is the number of observations (200) minus the number of coefficients in the regression model. The first model has two coefficients (one for the predictor and one for the constant) whereas the second has four (one for each of the three predictors and one for the constant). Therefore, model 1 has 198 degrees of freedom whereas model 2 has 196. The average sum of squares (MS) is then calculated for each term by dividing the SS by the *df*. The *F*-ratio is calculated by dividing the average improvement in prediction by the model ($MS_M$) by the average difference between the model and the observed data ($MS_R$). If the improvement due to fitting the regression model is much greater than the inaccuracy within the model then the value of *F* will be greater than 1, and SPSS calculates the exact probability of obtaining the value of *F* by chance. For the initial model the *F*-ratio is 99.59, $p < .001$. For the second the *F*-ratio is 129.498 – also highly significant ($p < .001$). We can interpret these results as meaning that both models significantly improved our ability to predict the outcome variable compared to not fitting the model.

**OUTPUT 8.6**



**ANOVA**[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 433687.833 | 1 | 433687.833 | 99.587 | .000[b] |
| | Residual | 862264.167 | 198 | 4354.870 | | |
| | Total | 1295952.00 | 199 | | | |
| 2 | Regression | 861377.418 | 3 | 287125.806 | 129.498 | .000[c] |
| | Residual | 434574.582 | 196 | 2217.217 | | |
| | Total | 1295952.00 | 199 | | | |

a. Dependent Variable: Album Sales (Thousands)

b. Predictors: (Constant), Advertsing Budget (Thousands of Pounds)

c. Predictors: (Constant), Advertsing Budget (Thousands of Pounds), Attractiveness of Band, No. of plays on Radio

## CRAMMING SAM'S TIPS The model summary

- The fit of the regression model can be assessed using the *Model Summary* and *ANOVA* tables from SPSS.
- Look for the $R^2$ to tell you the proportion of variance explained by the model.
- If you have done a hierarchical regression then assess the improvement of the model at each stage of the analysis by looking at the change in $R^2$ and whether this change is significant (look for values less than .05 in the column labelled *Sig F Change*).
- The ANOVA also tells us whether the model is a significant fit of the data overall (look for values less than .05 in the column labelled *Sig.)*.
- The assumption that errors are independent is likely to be met if the Durbin–Watson statistic is close to 2 (and between 1 and 3).

## 8.7.3. Model parameters ②

So far we have looked at whether or not the model has improved our ability to predict the outcome variable. The next part of the output is concerned with the parameters of the model. Output 8.7 shows the model parameters for both steps in the hierarchy. Now, the first step in our hierarchy was to include advertising budget (as we did for the simple regression earlier in this chapter) and so the parameters for the first model are identical to the parameters obtained in Output 8.3. Therefore, we will discuss only the parameters for the final model (in which all predictors were included). The format of the table of coefficients will depend on the options selected. The confidence interval for the *b*-values, collinearity diagnostics and the part and partial correlations will be present only if selected in the dialog box in Figure 8.16.

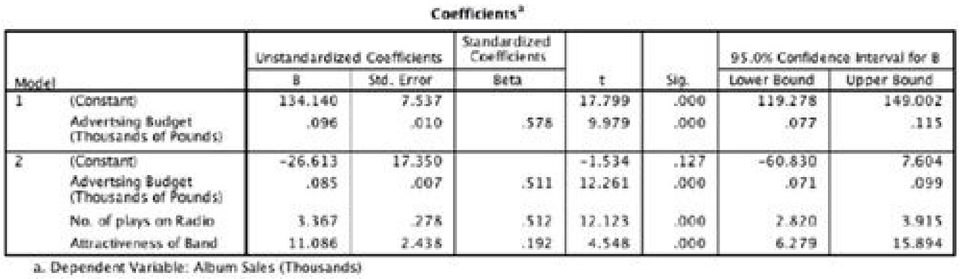
Remember that in multiple regression the model takes the form of equation (8.6), and in that equation there are several unknown parameters (the *b*-values). The first part of the table gives us estimates for these *b*-values, and these values indicate the individual contribution of each predictor to the model. By replacing the *b*-values in equation (8.6) we can define our specific model as:

$$\begin{aligned}\text{sales}_i &= b_0 + b_1\text{advertising}_i + b_2\text{airplay}_i + b_3\text{attractiveness}_i \\ &= -26.61 + (0.08\ \text{advertising}_i) + (3.37\ \text{airplay}_i) + (11.09\ \text{attractiveness}_i) \qquad (8.16)\end{aligned}$$

The *b*-values tell us about the relationship between album sales and each predictor. If the value is positive we can tell that there is a positive relationship between the predictor and the outcome, whereas a negative coefficient represents a negative relationship. For these data all three predictors have positive *b*-values indicating positive relationships. So, as advertising budget increases, album sales increase; as plays on the radio increase, so do album sales; and finally, more attractive bands will sell more albums. The *b*-values tell us more than this, though. They tell us to what degree each predictor affects the outcome *if the effects of all other predictors are held constant*.

## OUTPUT 8.7

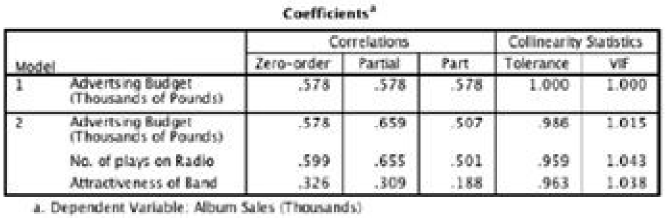
Coefficients of the regression model[17]



**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 134.140 | 7.537 | | 17.799 | .000 | 119.278 | 149.002 |
| | Advertsing Budget (Thousands of Pounds) | .096 | .010 | .578 | 9.979 | .000 | .077 | .115 |
| 2 | (Constant) | -26.613 | 17.350 | | -1.534 | .127 | -60.830 | 7.604 |
| | Advertsing Budget (Thousands of Pounds) | .085 | .007 | .511 | 12.261 | .000 | .071 | .099 |
| | No. of plays on Radio | 3.367 | .278 | .512 | 12.123 | .000 | 2.820 | 3.915 |
| | Attractiveness of Band | 11.086 | 2.438 | .192 | 4.548 | .000 | 6.279 | 15.894 |

a. Dependent Variable: Album Sales (Thousands)



**Coefficients**[a]

| Model | | Correlations | | | Collinearity Statistics | |
|---|---|---|---|---|---|---|
| | | Zero-order | Partial | Part | Tolerance | VIF |
| 1 | Advertsing Budget (Thousands of Pounds) | .578 | .578 | .578 | 1.000 | 1.000 |
| 2 | Advertsing Budget (Thousands of Pounds) | .578 | .659 | .507 | .986 | 1.015 |
| | No. of plays on Radio | .599 | .655 | .501 | .959 | 1.043 |
| | Attractiveness of Band | .326 | .309 | .188 | .963 | 1.038 |

a. Dependent Variable: Album Sales (Thousands)

- **Advertising budget** (*b* = 0.085): This value indicates that as advertising budget increases by one unit, album sales increase by 0.085 units. Both variables were measured in thousands; therefore, for every £1000 more spent on advertising, an extra 0.085 thousand albums (85 albums) are sold. This interpretation is true only if the effects of attractiveness of the band and airplay are held constant.
- **Airplay** (*b* = 3.367): This value indicates that as the number of plays on radio in the week before release increases by one, album sales increase by 3.367 units. Therefore, every additional play of a song on radio (in the week before release) is associated with an extra 3.367 thousand albums (3367 albums) being sold. This interpretation is true only if the effects of attractiveness of the band and

advertising are held constant.

- **Attractiveness** ($b$ = 11.086): This value indicates that a band rated one unit higher on the attractiveness scale can expect additional album sales of 11.086 units. Therefore, every unit increase in the attractiveness of the band is associated with an extra 11.086 thousand albums (11,086 albums) being sold. This interpretation is true only if the effects of radio airplay and advertising are held constant.

Each of the beta values has an associated standard error indicating to what extent these values would vary across different samples, and these standard errors are used to determine whether or not the *b*-value differs significantly from zero. As we saw in Section 8.4.3.2, a *t*-statistic can be derived that tests whether a *b*-value is significantly different from 0. With only one predictor a significant value of *t* indicates that the slope of the regression line is significantly different from horizontal, but with many predictors it is not so easy to visualize what the value tells us. Instead, it is easiest to conceptualize the *t*-tests as measures of whether the predictor is making a significant contribution to the model. Therefore, if the *t*-test associated with a *b*-value is significant (if the value in the column labelled *Sig.* is less than .05) then the predictor is making a significant contribution to the model. The smaller the value of *Sig.* (and the larger the value of *t*), the greater the contribution of that predictor. For this model, the advertising budget, $t(196) = 12.26$, $p < .001$, the amount of radio play prior to release, $t(196) = 12.12$, $p < .001$ and attractiveness of the band, $t(196) = 4.55$, $p < .001$, are all significant predictors of album sales.[18] Remember that these significance tests are accurate only if the assumptions discussed in Chapter 5 are met. From the magnitude of the *t*-statistics we can see that the advertising budget and radio play had a similar impact, whereas the attractiveness of the band had less impact.

The *b*-values and their significance are important statistics to look at; however, the standardized versions of the *b*-values are probably easier to interpret (because they are not dependent on the units of measurement of the variables). The standardized beta values (labelled as Beta, $b_i$) tell us the number of standard deviations that the outcome will change as a result of one standard deviation change in the predictor. The standardized beta values are all measured in standard deviation units and so are directly comparable: therefore, they provide a better insight into the ‘importance’ of a predictor in the model. The standardized beta values for airplay and advertising budget are virtually identical (.512 and .511 respectively) indicating that both variables have a comparable degree of importance in the model (this concurs with what the magnitude of the *t*-statistics told us). To interpret these values literally, we need to know the standard deviations of all of the variables, and these values can be found in Output 8.4.

- **Advertising budget** (*standardized ß* = .511): This value indicates that as advertising budget increases by one standard deviation (£485,655), album sales increase by 0.511 standard deviations. The standard deviation for album sales is 80,699 and so this constitutes a change of 41,240 sales (0.511 × 80,699). Therefore, for every £485,655 more spent on advertising, an extra 41,240 albums are sold. This interpretation is true only if the effects of attractiveness of the band and airplay are held constant.
- **Airplay** (*standardized ß* = .512): This value indicates that as the number of plays on radio in the week before release increases by one standard deviation (12.27), album sales increase by 0.512 standard deviations. The standard deviation for album sales is 80,699 and so this constitutes a change of 41,320 sales (0.512 × 80,699). Therefore, if Radio 1 plays the song an extra 12.27 times in the week before release, 41,320 extra album sales can be expected. This interpretation is true only if the effects of attractiveness of the band and advertising are held constant.
- **Attractiveness** (*standardized ß* = .192): This value indicates that a band rated one standard deviation (1.40 units) higher on the attractiveness scale can expect additional album sales of 0.192

standard deviations units. This constitutes a change of 15,490 sales (0.192 × 80,699). Therefore, a band with an attractiveness rating 1.40 higher than another band can expect 15,490 additional sales. This interpretation is true only if the effects of radio airplay and advertising are held constant.

SELF-TEST Think back to what the confidence interval of the mean represented (Section 2.5.2). Can you work out what the confidence interval for *b* represents?

We are also given the confidence intervals for the betas (again these are accurate only if the assumptions discussed in Chapter 5 are met). Imagine that we collected 100 samples of data measuring the same variables as our current model. For each sample we could create a regression model to represent the data. If the model is reliable then we hope to find very similar parameters (*b*s) in all samples. The confidence intervals of the unstandardized beta values are boundaries constructed such that in 95% of samples these boundaries contain the population value of *b* (see Section 2.5.2). Therefore, if we'd collected 100 samples, and calculated the confidence intervals for *b*, we are saying that 95% of these confidence intervals would contain the true value of *b*. Therefore, we can be fairly confident that the confidence interval we have constructed for this sample will contain the true value of *b* in the population. This being so, a good model will have a small confidence interval, indicating that the value of *b* in this sample is close to the true value of *b* in the population. The sign (positive or negative) of the *b*-values tells us about the direction of the relationship between the predictor and the outcome. Therefore, we would expect a very bad model to have confidence intervals that cross zero, indicating that in the population the predictor could have a negative relationship to the outcome but could also have a positive relationship. In this model the two best predictors (advertising and airplay) have very tight confidence intervals, indicating that the estimates for the current model are likely to be representative of the true population values. The interval for attractiveness is wider (but still does not cross zero), indicating that the parameter for this variable is less representative, but nevertheless significant.

If you asked for part and partial correlations, then they will appear in the output in separate columns of the table. The zero-order correlations are the simple Pearson's correlation coefficients (and so correspond to the values in Output 8.4). The partial correlations represent the relationships between each predictor and the outcome variable, controlling for the effects of the other two predictors. The part correlations represent the relationship between each predictor and the outcome, controlling for the effect that the other two variables have on the outcome. In effect, these part correlations represent the unique relationship that each predictor has with the outcome. If you opt to do a stepwise regression, you would find that variable entry is based initially on the variable with the largest zero-order correlation and then on the part correlations of the remaining variables. Therefore, airplay would be entered first (because it has the largest zero-order correlation), then advertising budget (because its part correlation is bigger than attractiveness) and then finally attractiveness – try running a forward stepwise regression on these data to see if I'm right. Finally, we are given details of the collinearity statistics, but these will be discussed in Section 8.7.5.

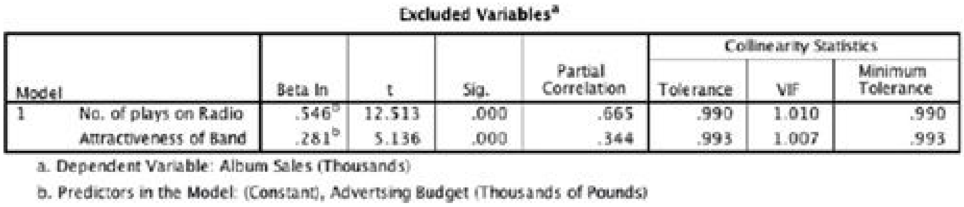
## CRAMMING SAM'S TIPS Model parameters

- The individual contribution of variables to the regression model can be found in the *Coefficients* table from SPSS. If you have done a hierarchical regression then look at the values for the final model.
- For each predictor variable, you can see if it has made a significant contribution to predicting the outcome by looking at the column labelled *Sig.* (values less than .05 are significant).
- The standardized beta values tell you the importance of each predictor (bigger absolute value = more important).
- The tolerance and VIF values will also come in handy later on, so make a note of them.

## 8.7.4. Excluded variables ②

At each stage of a regression analysis SPSS provides a summary of any variables that have not yet been entered into the model. In a hierarchical model, this summary has details of the variables that have been specified to be entered in subsequent steps, and in stepwise regression this table contains summaries of the variables that SPSS is considering entering into the model. For this example, there is a summary of the excluded variables (Output 8.8) for the first stage of the hierarchy (there is no summary for the second stage because all predictors are in the model). The summary gives an estimate of each predictor's beta value if it was entered into the equation at this point and calculates a *t*-test for this value. In a stepwise regression, SPSS should enter the predictor with the highest *t*-statistic and will continue entering predictors until there are none left with *t*-statistics that have significance values less than .05. The partial correlation also provides some indication as to what contribution (if any) an excluded predictor would make if it were entered into the model.

**OUTPUT 8.8**



Excluded Variables[a]

| Model | | Beta In | t | Sig. | Partial Correlation | Collinearity Statistics | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Tolerance | VIF | Minimum Tolerance |
| 1 | No. of plays on Radio | .546[b] | 12.513 | .000 | .665 | .990 | 1.010 | .990 |
| | Attractiveness of Band | .281[b] | 5.136 | .000 | .344 | .993 | 1.007 | .993 |

a. Dependent Variable: Album Sales (Thousands)

b. Predictors in the Model: (Constant), Advertsing Budget (Thousands of Pounds)

## 8.7.5. Assessing multicollinearity ②

Output 8.7 provided some measures of whether there is collinearity in the data. Specifically, it provided the VIF and tolerance statistics (with tolerance being 1 divided by the VIF). We can apply the guidelines from Section 8.5.3 to our model. The VIF values are all well below 10 and the tolerance statistics all well above 0.2; therefore, we can safely conclude that there is no collinearity within our data. To calculate the average VIF we simply add the VIF values for each predictor and divide by the number of predictors (*k*):

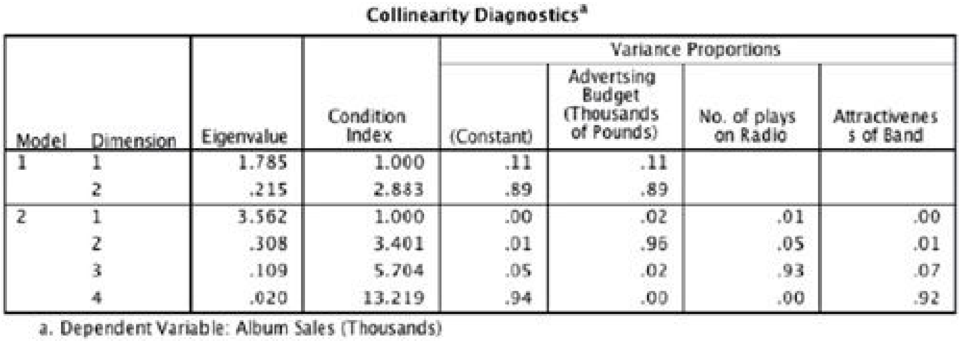
$$\overline{\text{VIF}} = \frac{\sum_{i=1}^{k}\text{VIF}_i}{k} = \frac{1.015+1.043+1.038}{3} = 1.032$$

The average VIF is very close to 1 and this confirms that collinearity is not a problem for this model.

SPSS also produces a table of eigenvalues of the scaled, uncentred cross-products matrix, condition indexes and variance proportions. There is a lengthy discussion, and example, of collinearity in Section 19.8.2 and how to detect it using variance proportions, so I will limit myself now to saying that we are looking for large variance proportions on the same *small* eigenvalues (Jane Superbrain Box 8.3). Therefore, in Output 8.9 we look at the bottom few rows of the table (these are the small eigenvalues) and look for any variables that both have high variance proportions for that eigenvalue. The variance proportions vary between 0 and 1, and for each predictor should be distributed across different dimensions (or eigenvalues). For this model, you can see that each predictor has most of its variance loading onto a different dimension (advertising has 96% of variance on dimension 2, airplay has 93% of variance on dimension 3 and attractiveness has 92% of variance on dimension 4).

These data represent a classic example of no multicollinearity. For an example of when collinearity exists in the data and some suggestions about what can be done, see Chapters 19 (Section 19.8.2) and 17 (Section 17.3.3.3).

## OUTPUT 8.9



Collinearity Diagnostics[a]

| Model | Dimension | Eigenvalue | Condition Index | Variance Proportions | | | |
|---|---|---|---|---|---|---|---|
| | | | | (Constant) | Advertising Budget (Thousands of Pounds) | No. of plays on Radio | Attractiveness of Band |
| 1 | 1 | 1.785 | 1.000 | .11 | .11 | | |
| | 2 | .215 | 2.883 | .89 | .89 | | |
| 2 | 1 | 3.562 | 1.000 | .00 | .02 | .01 | .00 |
| | 2 | .308 | 3.401 | .01 | .96 | .05 | .01 |
| | 3 | .109 | 5.704 | .05 | .02 | .93 | .07 |
| | 4 | .020 | 13.219 | .94 | .00 | .00 | .92 |

a. Dependent Variable: Album Sales (Thousands)

## CRAMMING SAM'S TIPS Multicollinearity

- To check for multicollinearity, use the VIF values from the table labelled *Coefficients* in the SPSS output.
- If these values are less than 10, then there probably isn't cause for concern.
- If you take the average of VIF values, and it is not substantially greater than 1, then there's also no cause for concern.

## JANE SUPERBRAIN 8.3

### *What are eigenvectors and eigenvalues?* ④

The definitions and mathematics of eigenvalues and eigenvectors are very complicated and most of us need not worry about them (although they do crop up again in Chapters 16 and 17). However, although the mathematics is hard, they are quite easy to visualize. Imagine we have two variables: the salary a supermodel earns in a year, and how attractive she is. Also imagine these two variables are normally distributed and so can be considered together as a bivariate normal distribution. If these variables are correlated, then their scatterplot forms an ellipse: if we draw a dashed line around the outer values of the scatterplot we get something oval shaped (Figure 8.20). We can draw two lines to measure the length and height of this ellipse. These lines are the *eigenvectors* of the original correlation matrix for these two variables (a vector is just a set of numbers that tells us the location of a line in geometric space). Note that the two lines we've drawn (one for height and one for width of the oval) are perpendicular; that is, they are at 90 degrees to each other, which means that they are independent of one another). So, with two variables, eigenvectors are just lines measuring the length and height of the ellipse that surrounds the scatterplot of data for those variables.

If we add a third variable (e.g., the length of experience of the supermodel) then all that happens is our scatterplot gets a third dimension, the ellipse turns into something shaped like a rugby ball (or American football), and because we now have a third dimension (height, width and depth) we get an extra eigenvector to measure this extra dimension. If we add a fourth variable, a similar logic applies (although it's harder to visualize): we get an extra dimension, and an eigenvector to measure that dimension. Each eigenvector has an *eigenvalue* that tells us its length (i.e., the distance from one end of the eigenvector to the other). So, by looking at all of the eigenvalues for a data set, we know the dimensions of the ellipse or rugby ball: put more generally, we know the dimensions of the data. Therefore, the eigenvalues show how evenly (or otherwise) the variances of the matrix are distributed.

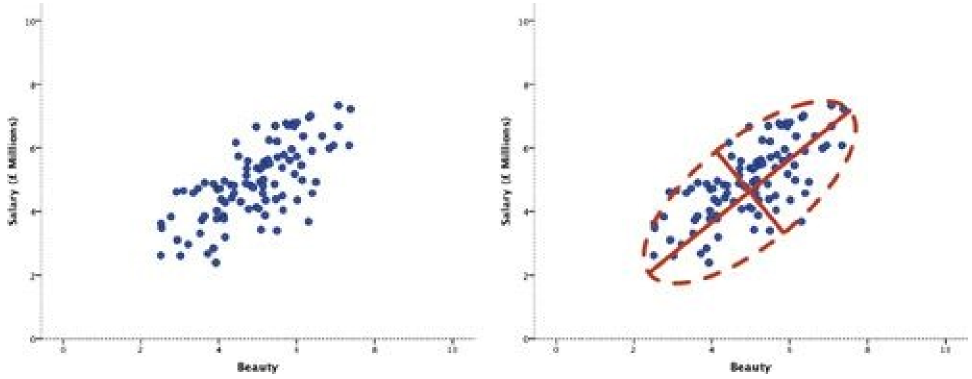


**FIGURE 8.20** A scatterplot of two variables forms an ellipse

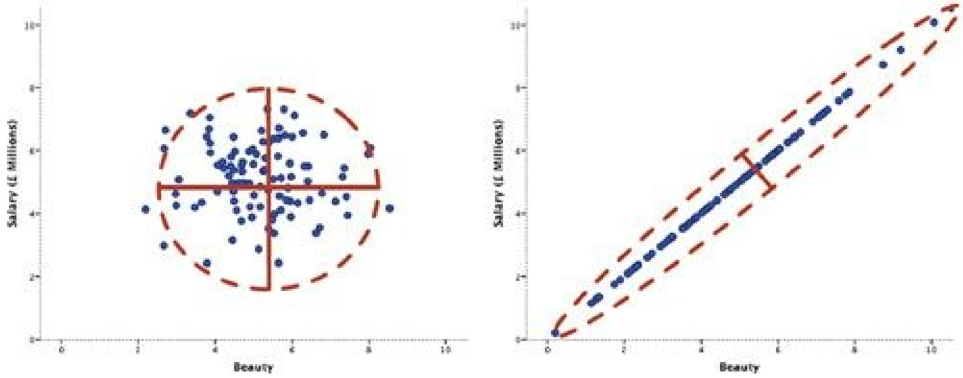


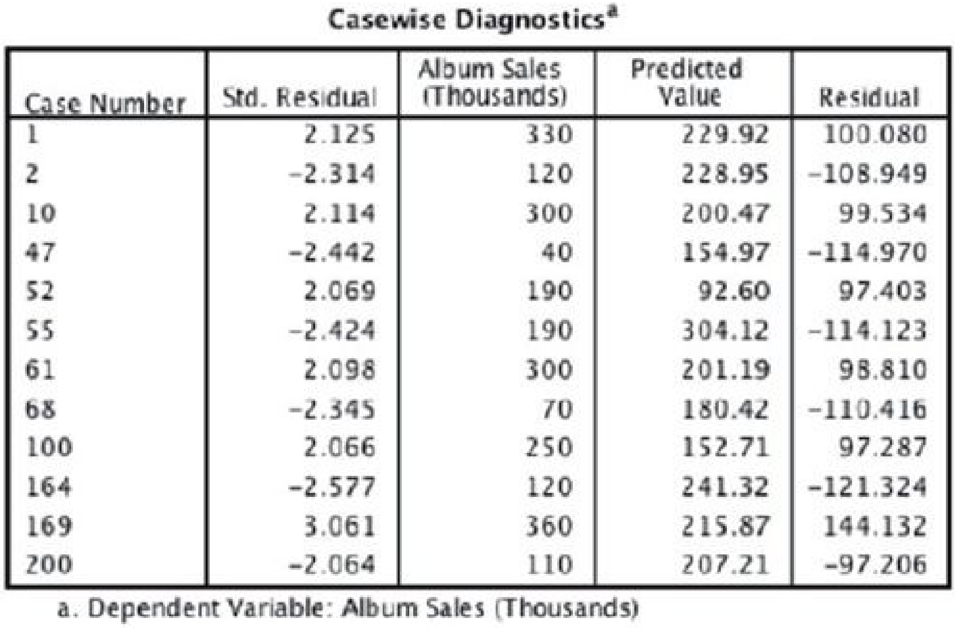
**FIGURE 8.21** Perfectly uncorrelated (left) and correlated (right) variables

In the case of two variables, the *condition* of the data is related to the ratio of the larger eigenvalue to the smaller. Figure 8.21 shows the two extremes: when there is no relationship at all between variables (left), and when there is a perfect relationship (right). When there is no relationship, the scatterplot will be contained roughly within a circle (or a sphere if we had three variables). If we draw lines that measure the height and width of this circle we'll find that these lines are the same length. The eigenvalues measure the length, therefore the eigenvalues will also be the same. So, when we divide the largest eigenvalue by the smallest we'll get a value of 1 (because the eigenvalues are the same). When the variables are perfectly correlated (i.e., there is perfect collinearity) then the scatterplot forms a straight line and the ellipse surrounding it will also collapse to a straight line. Therefore, the height of the ellipse will be very small indeed (it will approach zero). Therefore, when we divide the largest eigenvalue by the smallest we'll get a value that tends to infinity (because the smallest eigenvalue is close to zero). Therefore, an infinite condition index is a sign of deep trouble.

## 8.7.6. Bias in the model: casewise diagnostics ②

The final stage of the general procedure outlined in Figure 8.11 is to check the residuals for evidence of bias. We do this in two stages. The first is to examine the casewise diagnostics, and the second is to check the assumptions discussed in Chapter 5. SPSS produces a summary table of the residual statistics, and these should be examined for extreme cases. Output 8.10 shows any cases that have a standardized residual less than −2 or greater than 2 (remember that we changed the default criterion from 3 to 2 in Figure 8.16). I mentioned in Section 8.3.1.1 that in an ordinary sample we would expect 95% of cases to have standardized residuals within about ±2. We have a sample of 200, therefore it is reasonable to expect about 10 cases (5%) to have standardized residuals outside of these limits. From Output 8.10 we can see that we have 12 cases (6%) that are outside the limits: therefore, our sample is within 1% of what we would expect. In addition, 99% of cases should lie within ±2.5 and so we would expect only 1% of cases to lie outside these limits. From the cases listed here, it is clear that two cases (1%) lie outside of the limits (cases 164 and 169). Therefore, our sample appears to conform to what we would expect for a fairly accurate model. These diagnostics give us no real cause for concern except that case 169 has a standardized residual greater than 3, which is probably large enough for us to investigate further.

### OUTPUT 8.10



Casewise Diagnostics[a]

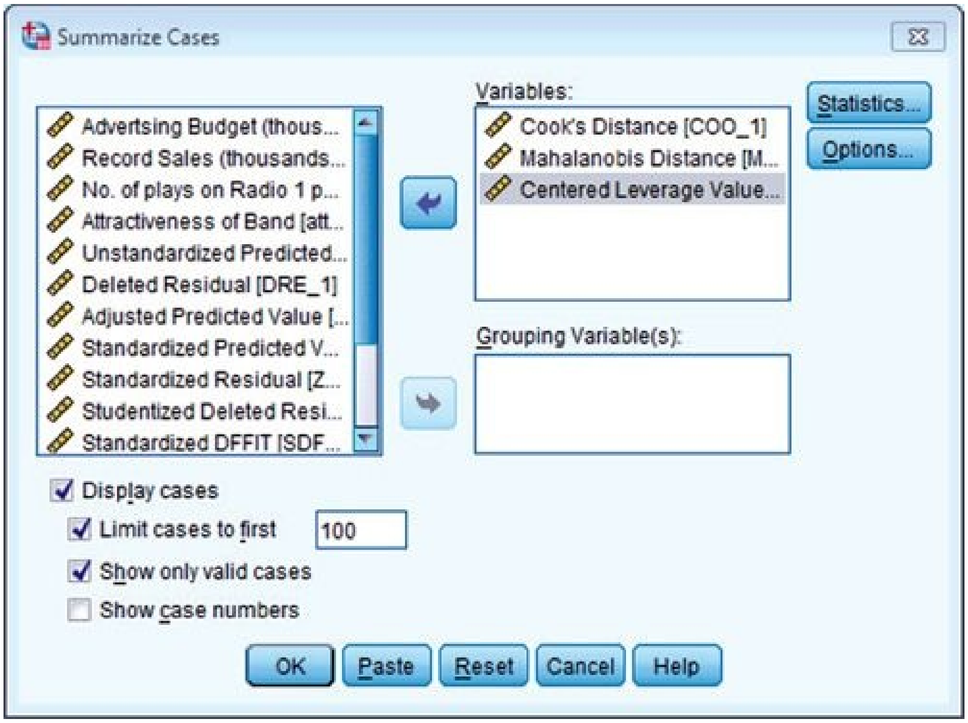
| Case Number | Std. Residual | Album Sales (Thousands) | Predicted Value | Residual |
|---|---|---|---|---|
| 1 | 2.125 | 330 | 229.92 | 100.080 |
| 2 | -2.314 | 120 | 228.95 | -108.949 |
| 10 | 2.114 | 300 | 200.47 | 99.534 |
| 47 | -2.442 | 40 | 154.97 | -114.970 |
| 52 | 2.069 | 190 | 92.60 | 97.403 |
| 55 | -2.424 | 190 | 304.12 | -114.123 |
| 61 | 2.098 | 300 | 201.19 | 98.810 |
| 68 | -2.345 | 70 | 180.42 | -110.416 |
| 100 | 2.066 | 250 | 152.71 | 97.287 |
| 164 | -2.577 | 120 | 241.32 | -121.324 |
| 169 | 3.061 | 360 | 215.87 | 144.132 |
| 200 | -2.064 | 110 | 207.21 | -97.206 |

a. Dependent Variable: Album Sales (Thousands)

You may remember that in Section 8.6.4 we asked SPSS to save various diagnostic statistics. You should find that the data editor now contains columns for these variables. It is perfectly acceptable to check these values in the data editor, but you can also get SPSS to list the values in your viewer window too. To list variables you need to use the *Case Summaries* command, which can be found by selecting Analyze Reports ▸ Case Summaries... . Figure 8.22 shows the dialog box for this function. Simply select the variables that you want to list and transfer them to the box labelled *Variables* by clicking on ➡. By default, SPSS will limit the output to the first 100 cases, but if you want to list all of your cases then deselect this option (see also SPSS Tip 8.1). It is also very important to select the *Show case numbers* option to enable you to tell the case number of any problematic cases.

To save space, Output 8.11 shows the influence statistics for 12 cases that I selected. None of them have a Cook's distance greater than 1 (even case 169 is well below this criterion) and so none of the cases has an undue influence on the model. The average leverage can be calculated as $(k + 1)/n = 4/200 = 0.02$, and so we are looking for values either twice as large as this (0.04) or three times as large (0.06) depending on which statistician you trust most (see Section 8.3.1.2). All cases are within the boundary of three times the average and only case 1 is close to two times the average.

**FIGURE 8.22**
The *Summarize Cases* dialog box



Advertsing Budget (thous...
Record Sales (thousands...
No. of plays on Radio 1 p...
Attractiveness of Band [att...
Unstandardized Predicted...
Deleted Residual [DRE_1]
Adjusted Predicted Value [...
Standardized Predicted V...
Standardized Residual [Z...
Studentized Deleted Resi...
Standardized DFFIT [SDF...

Variables:
Cook's Distance [COO_1]
Mahalanobis Distance [M...
Centered Leverage Value...

Grouping Variable(s):

☑ Display cases
☑ Limit cases to first 100
☑ Show only valid cases
☐ Show case numbers

## SPSS TIP 8.1 Selecting cases ③

In large data sets, a useful strategy when summarizing cases is to use SPSS's *Select Cases* function (see Section 5.4.2) and to set conditions that will select problematic cases. For example, you could create a variable that selects cases with a Cook's distance greater than 1 by running this syntax:

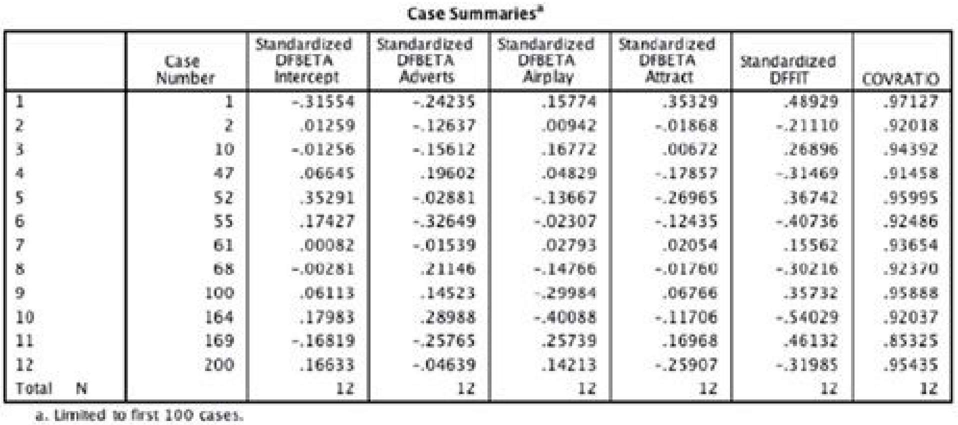
```
USE ALL.
```

```
COMPUTE cook_problem=(COO_1 > 1).

VARIABLE LABELS cook_problem ‘Cooks distance greater than 1’.

VALUE LABELS cook_problem 0 ‘Not Selected’ 1 ‘Selected’.

FILTER BY cook_problem.

EXECUTE.
```

This syntax creates a variable called cook_problem, based on whether Cook’s distance is greater than 1 (the *compute* command), it labels this variable as ‘Cooks distance greater than 1’ (the *variable labels* command), sets value labels to be 1 = include, 0 = exclude (the *value labels* command), and finally filters the data set by this new variable (the *filter by* command). Having selected cases, you can use case summaries to see which cases meet the condition you set (in this case having Cook’s distance greater than 1).

Finally, from our guidelines for the Mahalanobis distance we saw that with a sample of 100 and three predictors, values greater than 15 were problematic. Also, with three predictors, values greater than 7.81 are significant ($p < .05$). None of our cases come close to exceeding the criterion of 15, although a few would be deemed ‘significant’ (e.g., case 1). The evidence does not suggest major problems with no influential cases within our data (although all cases would need to be examined to confirm this fact).

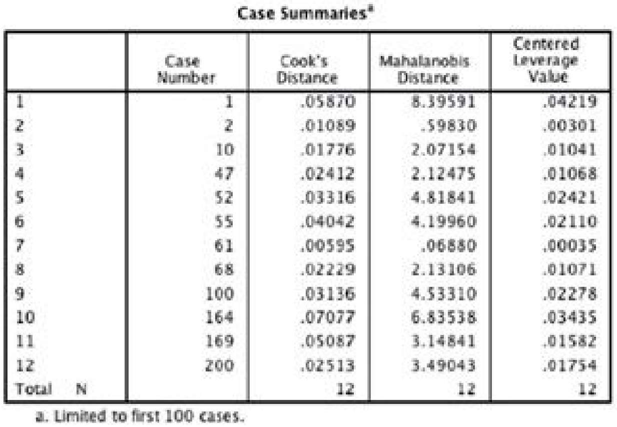
## OUTPUT 8.11



Case Summaries[a]

| | | Case Number | Standardized DFBETA Intercept | Standardized DFBETA Adverts | Standardized DFBETA Airplay | Standardized DFBETA Attract | Standardized DFFIT | COVRATIO |
|---|---|---|---|---|---|---|---|---|
| 1 | | 1 | -.31554 | -.24235 | .15774 | .35329 | .48929 | .97127 |
| 2 | | 2 | .01259 | -.12637 | .00942 | -.01868 | -.21110 | .92018 |
| 3 | | 10 | -.01256 | -.15612 | .16772 | .00672 | .26896 | .94392 |
| 4 | | 47 | .06645 | .19602 | .04829 | -.17857 | -.31469 | .91458 |
| 5 | | 52 | .35291 | -.02881 | -.13667 | -.26965 | .36742 | .95995 |
| 6 | | 55 | .17427 | -.32649 | -.02307 | -.12435 | -.40736 | .92486 |
| 7 | | 61 | .00082 | -.01539 | .02793 | .02054 | .15562 | .93654 |
| 8 | | 68 | -.00281 | .21146 | -.14766 | -.01760 | -.30216 | .92370 |
| 9 | | 100 | .06113 | .14523 | -.29984 | .06766 | .35732 | .95888 |
| 10 | | 164 | .17983 | .28988 | -.40088 | -.11706 | -.54029 | .92037 |
| 11 | | 169 | -.16819 | -.25765 | .25739 | .16968 | .46132 | .85325 |
| 12 | | 200 | .16633 | -.04639 | .14213 | -.25907 | -.31985 | .95435 |
| Total | N | | 12 | 12 | 12 | 12 | 12 | 12 |

a. Limited to first 100 cases.



Case Summaries[a]

| | | Case Number | Cook’s Distance | Mahalanobis Distance | Centered Leverage Value |
|---|---|---|---|---|---|
| 1 | | 1 | .05870 | 8.39591 | .04219 |
| 2 | | 2 | .01089 | .59830 | .00301 |
| 3 | | 10 | .01776 | 2.07154 | .01041 |
| 4 | | 47 | .02412 | 2.12475 | .01068 |
| 5 | | 52 | .03316 | 4.81841 | .02421 |
| 6 | | 55 | .04042 | 4.19960 | .02110 |
| 7 | | 61 | .00595 | .06880 | .00035 |
| 8 | | 68 | .02229 | 2.13106 | .01071 |
| 9 | | 100 | .03136 | 4.53310 | .02278 |
| 10 | | 164 | .07077 | 6.83538 | .03435 |
| 11 | | 169 | .05087 | 3.14841 | .01582 |
| 12 | | 200 | .02513 | 3.49043 | .01754 |
| Total | N | | 12 | 12 | 12 |

a. Limited to first 100 cases.

We can look also at the DFBeta statistics to see whether any case would have a large influence on the regression parameters. An absolute value greater than 1 is a problem and in all cases the values lie within ±1, which shows that these cases have no undue influence over the regression parameters.

There is also a column for the covariance ratio. We saw in Section 8.3.1.2 that we need to use the following criteria:

- $\text{CVR}_i > 1 + [3(k + 1)/n] = 1 + [3(3 + 1)/200] = 1.06$,
- $\text{CVR}_i < 1 - [3(k + 1)/n] = 1 - [3(3 + 1)/200] = 0.94$.

Therefore, we are looking for any cases that deviate substantially from these boundaries. Most of our 12 potential outliers have CVR values within or just outside these boundaries. The only case that causes concern is case 169 (again) whose CVR is some way below the bottom limit. However, given the Cook's distance for this case, there is probably little cause for alarm.

You would have requested other diagnostic statistics, and from what you know from the earlier discussion of them you would be well advised to glance over them in case of any unusual cases in the data. However, from this minimal set of diagnostics we appear to have a fairly reliable model that has not been unduly influenced by any subset of cases.

## CRAMMING SAM'S TIPS Residuals

You need to look for cases that might be influencing the regression model:

- Look at standardized residuals and check that no more than 5% of cases have absolute values above 2, and that no more than about 1% have absolute values above 2.5. Any case with a value above about 3 could be an outlier.
- Look in the data editor for the values of Cook's distance: any value above 1 indicates a case that might be influencing the model.
- Calculate the average leverage (the number of predictors plus 1, divided by the sample size) and then look for values greater than twice or three times this average value.
- For Mahalanobis distance, a crude check is to look for values above 25 in large samples (500) and values above 15 in smaller samples (100). However, Barnett and Lewis (1978) should be consulted for more detailed analysis.
- Look for absolute values of DFBeta greater than 1.
- Calculate the upper and lower limit of acceptable values for the covariance ratio, CVR. The upper limit is 1 plus three times the average leverage, while the lower limit is 1 minus three times the average leverage. Cases that have a CVR that falls outside these limits may be problematic.

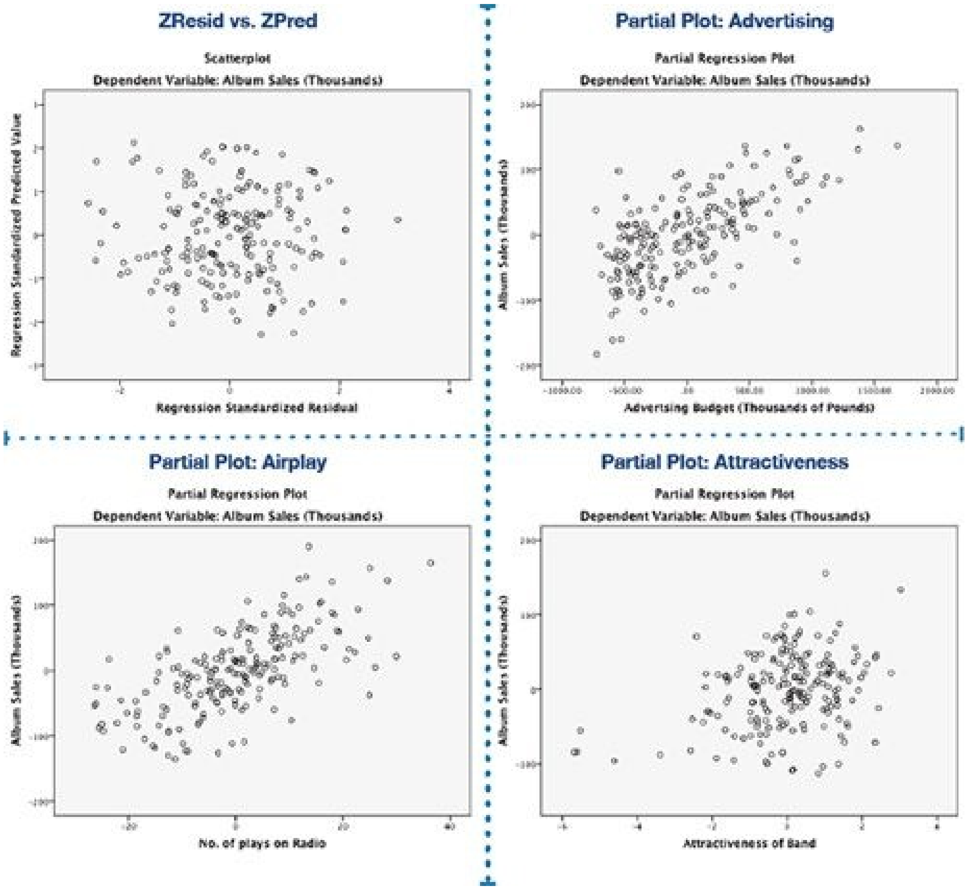
## 8.7.7. Bias in the model: assumptions ②

The general procedure outlined in Figure 8.11 suggests that, having fitted a model, we need to look for evidence of bias, and the second stage of this process is to check some assumptions. I urge you to review Chapter 5 to remind yourself of the main assumptions and the implications of violating them. We have already looked for collinearity within the data and used Durbin–Watson to check whether the residuals in the model are independent. We saw in Section 5.3.3.1 that we can look for heteroscedasticity and non-linearity using a plot of standardized residuals against standardized predicted values. We asked for this plot in Section 8.6.3. If everything is OK then this graph should look like a random array of dots, if the graph funnels out then that is a sign of heteroscedasticity and any curve suggests non-linearity (see Figure 5.20). Figure 8.23 (top left) shows the graph for our model. Note how the points are randomly and evenly dispersed throughout the plot. This pattern is indicative of a situation in which the assumptions of linearity and homoscedasticity have been met. Compare this with the examples in Figure 5.20.

Figure 8.23 also shows the partial plots, which are scatterplots of the residuals of the outcome variable and each of the predictors when both variables are regressed separately on the remaining predictors. Obvious outliers on a partial plot represent cases that might have undue influence on a predictor's regression coefficient, and non-linear relationships and heteroscedasticity can be detected using these plots as well. For advertising budget (Figure 8.23, top right) the partial plot shows the strong positive relationship to album sales. There are no obvious outliers on this plot, and the cloud of dots is evenly spaced out around the line, indicating homoscedasticity. For airplay (Figure 8.23, bottom left) the partial plot shows a strong positive relationship to album sales. The pattern of the residuals is similar to advertising (which would be expected, given the similarity of the standardized betas of these predictors). There are no obvious outliers on this plot, and the cloud of dots is evenly spaced around the line, indicating homoscedasticity. For attractiveness (Figure 8.23, bottom right) the plot again shows a positive relationship to album sales. The relationship looks less linear than for the other predictors, and the dots show some funnelling, indicating greater spread at high levels of attractiveness. There are no obvious outliers on this plot, but the funnel-shaped cloud of dots might indicate a violation of the assumption of homoscedasticity.

## FIGURE 8.23

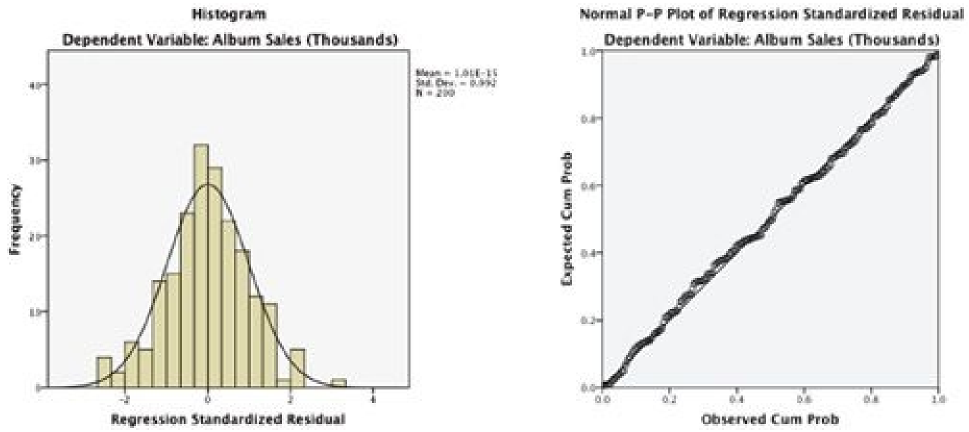
Plot of standardized predicted values against standardized residuals (top left), and partial plots of album sales against advertising (top right), airplay (bottom left) and attractiveness of the band (bottom right)



To test the normality of residuals, we look at the histogram and normal probability plot selected in Figure 8.17. Figure 8.24 shows the histogram and normal probability plot of the data for the current example. Compare these to examples of non-normality in Section 5.3.2.1. For the album sales data, the distribution is very normal: the histogram is symmetrical and approximately bell-shaped. The P-P plot shows up deviations from normality as deviations from the diagonal line (see Section 5.3.2.1). For our model, the dots lie almost exactly along the diagonal, which as we know indicates a normal distribution: hence this plot also suggests that the residuals are normally distributed.

**FIGURE 8.24**

Histograms and normal P-P plots of normally distributed residuals (left-hand side) and non-normally distributed residuals (right-hand side)



## CRAMMING SAM'S TIPS Model assumptions

- Look at the graph of ZRESID* plotted against ZPRED*. If it looks like a random array of dots then this is good. If the dots seem to get more or less spread out over the graph (look like a funnel) then this is probably a violation of the assumption of homogeneity of variance. If the dots have a pattern to them (i.e., a curved shape) then this is probably a violation of the assumption of linearity. If the dots seem to have a pattern and are more spread out at some points on the plot than others then this probably reflects violations of both homogeneity of variance *and* linearity. Any of these scenarios puts the validity of your model into question. Repeat the above for all partial plots too.
- Look at histograms and P-P plots. If the histograms look like normal distributions (and the P-P plot looks like a diagonal line), then all is well. If the histogram looks non-normal and the P-P plot looks like a wiggly snake curving around a diagonal line then things are less good. Be warned, though: distributions can look very non-normal in small samples even when they are normal.

## 8.8. What if I violate an assumption? Robust regression ②

We could summarize by saying that our model appears, in most senses, to be both accurate for the sample and generalizable to the population. The only slight glitch is some concern over whether attractiveness ratings had violated the assumption of homoscedasticity. Therefore, we could conclude that in our sample, advertising budget and airplay are fairly equally important in predicting album sales. Attractiveness of the band is a significant predictor of album sales but is less important than the other two predictors (and probably needs verification because of possible heteroscedasticity). The assumptions seem to have been met and so we can probably assume that this model would generalize to any album being released. However, this won't always be the case: there will be times when you uncover problems. It's worth looking carefully at Chapter 5 to see exactly what the implications are of violating assumptions, but in brief it will invalidate significance tests, confidence intervals and generalization of the model. These problems can be largely overcome by using robust methods such as bootstrapping (Section 5.4.3) to generate confidence intervals and significance tests of the model parameters. Therefore, if you uncover problems, rerun your regression, select the same options as before, but click Bootstrap... in the main dialog box (Figure 8.13) to access the bootstrap function. We discussed this dialog box in Section 5.4.3; to recap, select ☑ Perform bootstrapping to activate bootstrapping, and to get a 95% confidence interval click ◉ Percentile or ◉ Bias corrected accelerated (BCa). For this analysis, let's ask for a bias corrected and accelerated (BCa) confidence interval. The other thing is that bootstrapping doesn't appear to work if you ask SPSS to save diagnostics; therefore, click on Save... to open the dialog box in Figure 8.18 and *make sure that everything is deselected*. Back in the main dialog box, click on OK to run the analysis.

## LABCOAT LENI'S REAL RESEARCH 8.1

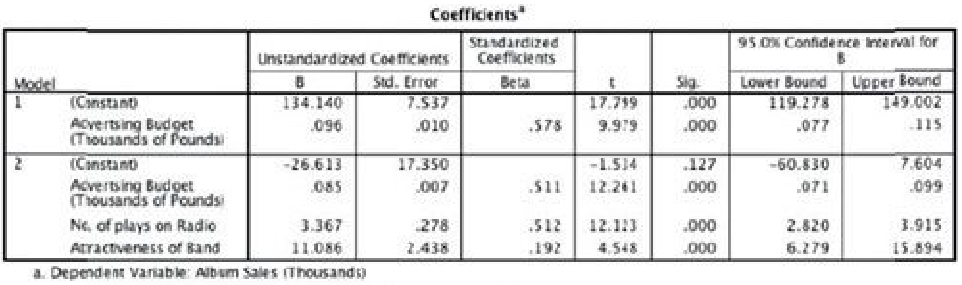
### *I want to be loved (on Facebook)* ①

Social media websites such as Facebook seem to have taken over the world. These websites offer an unusual opportunity to carefully manage your self-presentation to others (i.e., you can try to appear to be cool when in fact you write statistics books, appear attractive when you have huge pustules all over your face, fashionable when you wear 1980s heavy metal band T-shirts, and so on). Ong et al. (2011) conducted an interesting study that examined the relationship between narcissism and behaviour on Facebook in 275 adolescents. They measured the Age, Gender and Grade (at school), as well as extroversion and narcissism. They also measured how often (per week) these people updated their Facebook status (FB_Status), and also how they rated their own profile picture on each of four dimensions: coolness, glamour, fashionableness and attractiveness. These ratings were summed as an indicator of how positively they perceived the profile picture they had selected for their page (FB_Profile_TOT). They hypothesized that narcissism would predict, above and beyond the other variables, the frequency of status updates, and how positive a profile picture the person chose. To test this, they conducted two hierarchical regressions: one with FB_Status as the outcome and one with FB_Profile_TOT as the outcome. In both models they entered Age, Gender and Grade in the first block, then added extroversion (NEO_ FFI) in a second block, and finally narcissism (NPQC_R) in a third block. The data from this study are in the file Ong et al. (2011).sav. Labcoat Leni wants you to replicate their two hierarchical regressions and create a table of the results for each. Answers are on the companion website (or look at Table 2 in the original article).

ONG, E. Y. L., ET AL. (2011). *PERSONALITY AND INDIVIDUAL DIFFERENCES*, 50(2), 180–185.

The main difference will be a table of bootstrap confidence intervals for each predictor and their significance value.[19] These tell us that advertising, $b$ = 0.09 [0.07, 0.10], $p$ = .001, airplay, $b$ = 3.37 [2.74, 4.02], $p$ = .001, and attractiveness of the band, $b$ = 11.09 [6.46, 15.01], $p$ = .001, all significantly predict album sales. Note that as before, the bootstrapping process involves re-estimating the standard errors, so these have changed for each predictor (although not dramatically). The main benefit of the bootstrap confidence intervals and significance values is that they do not rely on assumptions of normality or homoscedasticity, so they give us an accurate estimate of the true population value of $b$ for each predictor.

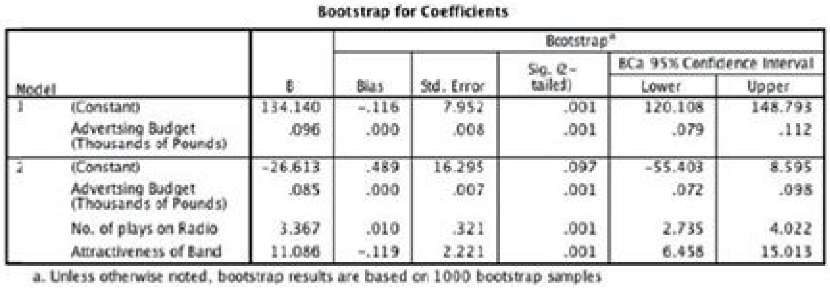
## OUTPUT 8.12



**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 134.140 | 7.537 | | 17.799 | .000 | 119.278 | 149.002 |
| | Advertsing Budget (Thousands of Pounds) | .096 | .010 | .578 | 9.979 | .000 | .077 | .115 |
| 2 | (Constant) | -26.613 | 17.350 | | -1.534 | .127 | -60.830 | 7.604 |
| | Advertsing Budget (Thousands of Pounds) | .085 | .007 | .511 | 12.261 | .000 | .071 | .099 |
| | No. of plays on Radio | 3.367 | .278 | .512 | 12.123 | .000 | 2.820 | 3.915 |
| | Attractiveness of Band | 11.086 | 2.438 | .192 | 4.548 | .000 | 6.279 | 15.894 |

a. Dependent Variable: Album Sales (Thousands)



**Bootstrap for Coefficients**

| Model | | B | Bootstrap[a] | | | | |
|---|---|---|---|---|---|---|---|
| | | | Bias | Std. Error | Sig. (2-tailed) | BCa 95% Confidence Interval | |
| | | | | | | Lower | Upper |
| 1 | (Constant) | 134.140 | -.116 | 7.952 | .001 | 120.108 | 148.793 |
| | Advertsing Budget (Thousands of Pounds) | .096 | .000 | .008 | .001 | .079 | .112 |
| 2 | (Constant) | -26.613 | .489 | 16.295 | .097 | -55.403 | 8.595 |
| | Advertsing Budget (Thousands of Pounds) | .085 | .000 | .007 | .001 | .072 | .098 |
| | No. of plays on Radio | 3.367 | .010 | .321 | .001 | 2.735 | 4.022 |
| | Attractiveness of Band | 11.086 | -.119 | 2.221 | .001 | 6.458 | 15.013 |

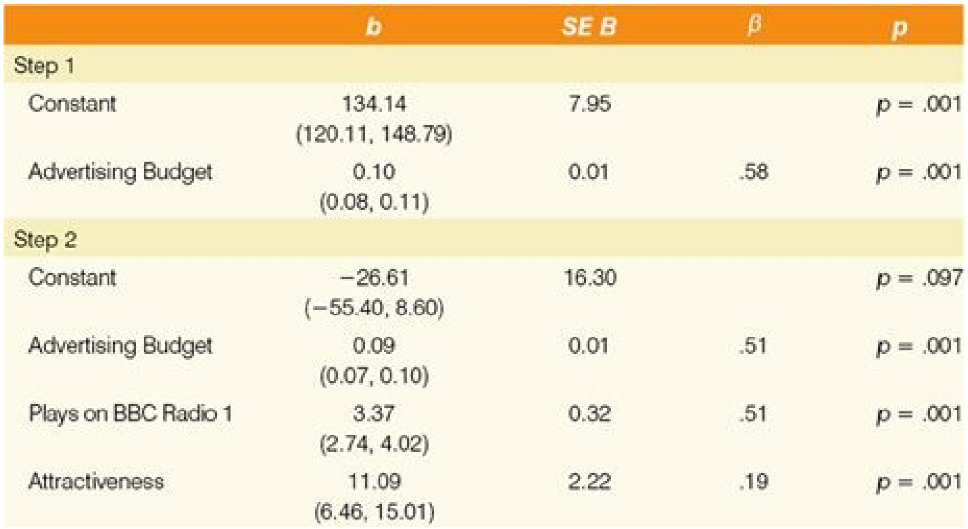
a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

## 8.9. How to report multiple regression ②

If your model has several predictors then you can't really beat a summary table as a concise way to report your model. As a bare minimum, report the betas, their confidence interval, significance value and some general statistics about the model (such as the $R^2$). The standardized beta values and the standard errors are also very useful. Personally I like to see the constant as well because then readers of your work can construct the full regression model if they need to. For hierarchical regression you should report these values at each stage of the hierarchy. So, basically, you want to reproduce the table labelled *Coefficients* from the SPSS output and omit some of the non-essential information. For the example in this chapter we might produce a table like that in Table 8.2.

Look back through the SPSS output in this chapter and see if you can work out from where the values came. Things to note are: (1) I've rounded off to 2 decimal places throughout because this is a reasonable level of precision given the variables measured; (2) for the standardized betas there is no zero before the decimal point (because these values shouldn't exceed 1) but for all other values less than 1 the zero is present; (3) often you'll see that the significance of the variable is denoted by an asterisk with a footnote to indicate the significance level being used, but it's better practice to report exact *p*-values; (4) the $R^2$ for the initial model and the change in $R^2$ (denoted as $\Delta R^2$) for each subsequent step of the model are reported below the table; and (5) in the title I have mentioned that confidence intervals and standard errors in the table are based on bootstrapping – this information is important for readers to know.

**TABLE 8.2** Linear model of predictors of album sales, with 95% bias corrected and accelerated confidence intervals reported in parentheses. Confidence intervals and standard errors based on 1000 bootstrap samples



| | *b* | *SE B* | *β* | *p* |
|---|---|---|---|---|
| Step 1 | | | | |
| Constant | 134.14<br>(120.11, 148.79) | 7.95 | | *p* = .001 |
| Advertising Budget | 0.10<br>(0.08, 0.11) | 0.01 | .58 | *p* = .001 |
| Step 2 | | | | |
| Constant | −26.61<br>(−55.40, 8.60) | 16.30 | | *p* = .097 |
| Advertising Budget | 0.09<br>(0.07, 0.10) | 0.01 | .51 | *p* = .001 |
| Plays on BBC Radio 1 | 3.37<br>(2.74, 4.02) | 0.32 | .51 | *p* = .001 |
| Attractiveness | 11.09<br>(6.46, 15.01) | 2.22 | .19 | *p* = .001 |

*Note.* $R^2$ = .34 for Step 1; $\Delta R^2$ = .33 for Step 2 (*p*s < .001).

## LABCOAT LENI’S REAL RESEARCH 8.2

### *Why do you like your lecturers?* ①

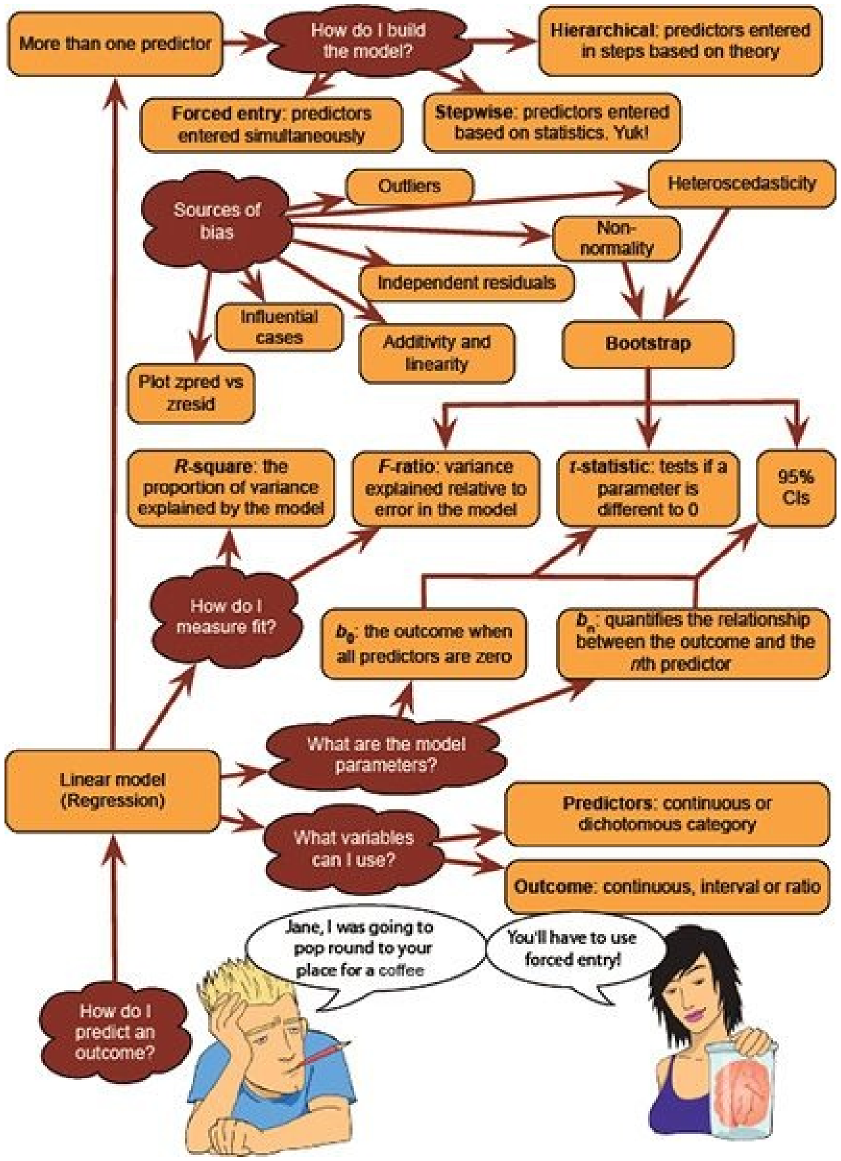
In the previous chapter we encountered a study by Chamorro-Premuzic et al. in which they measured students’ personality characteristics and asked them to rate how much they wanted these same characteristics in their lecturers (see Labcoat Leni’s Real Research 7.1 for a full description). In that chapter we correlated these scores; however, we could go a step further and see whether students’ personality characteristics predict the characteristics that they would like to see in their lecturers.

The data from this study are in the file ChamorroPremuzic.sav. Labcoat Leni wants you to carry out five multiple regression analyses: the outcome variable in each of the five analyses is the ratings of how much students want to see neuroticism, extroversion, openness to experience, agreeableness and conscientiousness. For each of these outcomes, force age and gender into the analysis in the first step of the hierarchy, then in the second block force in the five student personality traits (neuroticism, extroversion, openness to experience, agreeableness and conscientiousness). For each analysis create a table of the results. Answers are on the companion website (or look at Table 4 in the original article).

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

# 8.10. Brian’s attempt to woo Jane ①

**FIGURE 8.25** What Brian learnt from this chapter


More than one predictor
How do I build the model?
Hierarchical: predictors entered in steps based on theory
Forced entry: predictors entered simultaneously
Stepwise: predictors entered based on statistics. Yuk!
Sources of bias
Outliers
Heteroscedasticity
Non-normality
Independent residuals
Influential cases
Additivity and linearity
Bootstrap
Plot zpred vs zresid
R-square: the proportion of variance explained by the model
F-ratio: variance explained relative to error in the model
t-statistic: tests if a parameter is different to 0
95% CIs
How do I measure fit?
$b_0$: the outcome when all predictors are zero
$b_n$: quantifies the relationship between the outcome and the nth predictor
What are the model parameters?
Linear model (Regression)
What variables can I use?
Predictors: continuous or dichotomous category
Outcome: continuous, interval or ratio
Jane, I was going to pop round to your place for a coffee
You'll have to use forced entry!
How do I predict an outcome?

## 8.11. What next? ①

This chapter is possibly the longest book chapter ever written, and if you feel like you aged several years while reading it then, well, you probably have (look around, there are cobwebs in the room, you have a long beard, and when you go outside you'll discover a second ice age has been and gone, leaving only you and a few woolly mammoths to populate the planet). However, on the plus side, you now know more or less everything you ever need to know about statistics. Really, it's true; you'll discover in the coming chapters that everything else we discuss is basically a variation of this chapter. So, although you may be near death having spent your life reading this chapter (and I'm certainly near death having written it) you are officially a stats genius – well done!

We started the chapter by discovering that at 8 years old I could have really done with regression analysis to tell me which variables are important in predicting talent competition success. Unfortunately I didn't have regression, but fortunately I had my dad instead (and he's better than regression). He correctly predicted the recipe for superstardom, but in doing so he made me hungry for more. I was starting to get a taste for the rock-idol lifestyle: I had friends, a fortune (well, two gold-plated winner's medals), fast cars (a bike) and dodgy-looking 8-year-olds were giving me suitcases full of lemon sherbet to lick off of mirrors. The only things needed to complete the job were a platinum selling album and a heroin addiction. However, before that my parents and teachers were about to impress reality upon my young mind …

## 8.12. Key terms that I’ve discovered

Adjusted predicted value

Adjusted $R^2$

Autocorrelation

$b_i$

$\beta_i$

Cook’s distance

Covariance ratio (CVR)

Cross-validation

Deleted residual

DFBeta

DFFit

Dummy variables

Durbin–Watson test

*F*-ratio

Generalization

Goodness of fit

Hat values

Heteroscedasticity

Hierarchical regression

Homoscedasticity

Independent errors

Leverage

Mahalanobis distances

Mean squares

Model sum of squares

Multicollinearity

Multiple *r*

Multiple regression

Ordinary least squares (OLS)

Outcome variable

Perfect collinearity

Predicted value

Predictor variable

Residual

Residualsum of squares

Shrinkage

Simple regression

Standardized DFBeta

Standardized DFFit

Standardized residuals

Stepwise regression

Studentized deleted residuals

Studentized residuals

Suppressor effects

*t*-statistic
Tolerance
Total sum of squares
Unstandardized residuals
Variance inflation factor (VIF)

## 8.13. Smart Alex’s tasks

- **Task 1**: In Chapter 3 (Task 6) we looked at data based on findings that the number of cups of tea drunk was related to cognitive functioning (Feng et al., 2010). The data are in the file **Tea Makes You Brainy 716.sav**. Using the model that predicts cognitive functioning from tea drinking, what would cognitive functioning be if someone drank 10 cups of tea? Is there a significant effect? ①
- **Task 2**: Run a regression analysis for the **pubs.sav** data in Jane Superbrain Box 8.1 predicting **mortality** from the number of **pubs**. Try repeating the analysis but bootstrapping the confidence intervals. ②
- **Task 3:** In Jane Superbrain Box 2.1 we saw some data (**HonestyLab.sav**) relating to people’s ratings of dishonest acts and the likeableness of the perpetrator. Run a regression using bootstrapping to predict ratings of dishonesty from the likeableness of the perpetrator. ②
- **Task 4**: A fashion student was interested in factors that predicted the salaries of cat-walk models. She collected data from 231 models. For each model she asked them their salary per day on days when they were working (**Salary**), their age (**Age**), how many years they had worked as a model (**Years**), and then got a panel of experts from modelling agencies to rate the attractiveness of each model as a percentage, with 100% being perfectly attractive (**Beauty**). The data are in the file **Supermodel.sav**. Unfortunately, this fashion student bought a substandard statistics textbook and so doesn’t know how to analyse her data. ☺ Can you help her out by conducting a multiple regression to see which variables predict a model’s salary? How valid is the regression model? ②
- **Task 5**: A study was carried out to explore the relationship between **Aggression** and several potential predicting factors in 666 children who had an older sibling. Variables measured were **Parenting_Style** (high score = bad parenting practices), **Computer_ Games** (high score = more time spent playing computer games), **Television** (high score = more time spent watching television), **Diet** (high score = the child has a good diet low in harmful additives), and **Sibling_Aggression** (high score = more aggression seen in their older sibling). Past research indicated that parenting style and sibling aggression were good predictors of the level of aggression in the younger child. All other variables were treated in an exploratory fashion. The data are in the file **Child Aggression.sav**. Analyse them with multiple regression. ②
- **Task 6:** Repeat the analysis in Labcoat Leni’s Real Research 8.1 using bootstrapping for the confidence intervals. What are the confidence intervals for the regression parameters? ①
- **Task 7:** Coldwell, Pike, and Dunn (2006) investigated whether household chaos predicted children’s problem behaviour over and above parenting. From 118 families they recorded the age

and gender of the youngest child (**Child_age** and **Child_ gender**). They then interviewed the child about their relationship with their mum using the Berkeley Puppet Interview (BPI), which measures (1) warmth/enjoyment (**Child_warmth**), and (2) anger/hostility (**Child_anger**). Higher scores indicate more anger/hostility and warmth/enjoyment, respectively. Each mum was interviewed about their relationship with the child resulting in scores for relationship positivity (**Mum_pos)** and relationship negativity (**Mum_neg**). Household chaos (**Chaos**) was assessed using the Confusion, Hubbub, and Order Scale. The outcome variable was the child's adjustment (**sdq**): the higher the score, the more problem behaviour the child is reported to be displaying. The data are in the file **Coldwell et al. (2006).sav**. Conduct a hierarchical regression in three steps: (1) enter child age and gender; (2) add the variables measuring parent–child positivity, parent–child negativity, parent – child warmth and parent–child anger; (3) add chaos. Is household chaos predictive of children's problem behaviour over and above parenting? ③

Answers can be found on the companion website.

## 8.14. Further reading

guley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioural sciences*. Basingstoke: Palgrave Macmillan.

verman, B. L., & O'Connell, R. T. (1990). *Linear statistical models: An applied approach* (2nd ed.). Belmont, CA: Duxbury. (This text is only for the mathematically minded or postgraduate students, but provides an extremely thorough exposition of regression analysis.)

es, J. N. V., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. London: Sage. (This is an extremely readable text that covers regression in loads of detail but with minimum pain – highly recommended.)

[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 (see Oditi's Lantern).

[3] In case you're interested, by standardizing $b$, as we do when we compute a correlation coefficient, we're estimating $b$ for standardized versions of the predictor and outcome variables (i.e., versions of these variables that have a mean of 0 and standard deviation of 1). In this situation $b_0$ drops out of the equation because it is the value of the outcome when the predictor is 0, and when the predictor and outcome are standardized then when the predictor is 0, the outcome (and hence $b_0$) will be 0 also.

[4] For example, you'll sometimes see equation (8.1) written as $Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_{i.}$ The only difference is that this equation has βs in it instead of *b*s. Both versions are the same thing, they just use different letters to represent the coefficients.

[5] This is the correlation between the green dots and the blue dots in Figure 8.4. With only one predictor in the model this value will be the same as the Pearson correlation coefficient between the predictor and outcome variable.

[6] When the model contains more than one predictor, people sometimes refer to $R^2$ as multiple $R^2$. This is another example of how people attempt to make statistics more confusing than it needs to be by referring to the same thing in different ways. The meaning and interpretation of $R^2$ are the same regardless of how many predictors you have in the model or whether you choose to call it multiple $R^2$: it is the squared correlation between values of the outcome predicted by the model and the values observed in the data.

[7] You may come across the average leverage denoted as $p/n$ in which $p$ is the number of parameters being estimated. In regression, we estimate parameters for each predictor and also for a constant and so $p$ is equivalent to the number of predictors plus one ($k + 1$).

[8] The value of $b_1$ is reduced because the data no longer have a perfect linear relationship and so there is now variance that the model cannot explain.

[9] Some authors refer to these external variables as part of an error term that includes any random factor in the way in which the outcome varies. However, to avoid confusion with the residual terms in the regression equations I have chosen the label 'external variables'. Although this term implicitly washes over any random factors, I acknowledge their presence here.

[10] I used the program G*Power, mentioned in Section 2.6.1.7, to compute these values.

[11] I might cynically qualify this suggestion by proposing that predictors be chosen based on past research that has utilized good methodology. If basing such decisions on regression analyses, select predictors based only on past research that has used regression appropriately and yielded reliable, generalizable models.

[12] Hirotsugu Akaike (pronounced ‘A-ka-ee-kay’) was a Japanese statistician who gave his name to the AIC, which is used in a huge range of different places.

[13] We can see that although the data are messy in places, the three predictors have reasonably linear relationships with the outcome (album sales) and there are no obvious outliers.

[14] That is, 33% = 66.5% - 33.5% (this value is the *R Square Change* in the table).

[15] To get the same values as SPSS we have to use the exact value of $R^2$, which is 0.3346480676231 (if you don’t believe me double-click on the table in the SPSS output that reports this value, then double-click on the cell of the table containing the value of $R^2$ and you’ll see that .335 becomes the value just mentioned).

[16] The more precise value is 0.664668.

[17] To spare your eyesight I have split this part of the output into two tables; however, it should appear as one long table in the SPSS viewer.

[18] For all of these predictors I wrote $t(196)$. The number in brackets is the degrees of freedom. We saw in Section 8.2.5 that in regression the degrees of freedom are $N - p - 1$, where $N$ is the total sample size (in this case 200) and $p$ is the number of predictors (in this case 3). For these data we get $200 - 3 - 1 = 196$.

[19] Remember that because of how bootstrapping works the values in your output will be slightly different than mine, and different again if you rerun the analysis.