9 The Linear Model (Regression)

Field, Andy P. Discovering Statistics Using IBM SPSS Statistics. 5th edition, North American edition. Thousand Oaks, California: Sage Publications Inc., 2018. Print.

9.1 What will this chapter tell me?

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 eight. 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. Obviously actual talent wouldn’t be a factor. A linear model (regression) fitted to these data would enable us to predict the future (success in next week’s competition) based on values of the variables we’d measured. If, for example, singing was an important factor in getting a good audience evaluation, I could sing again the following week; but if jokers tended to do better then I might switch to a comedy routine. When I was eight I wasn’t the pathetic nerd that I am today, so I didn’t know about linear models (nor did I wish to); 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.

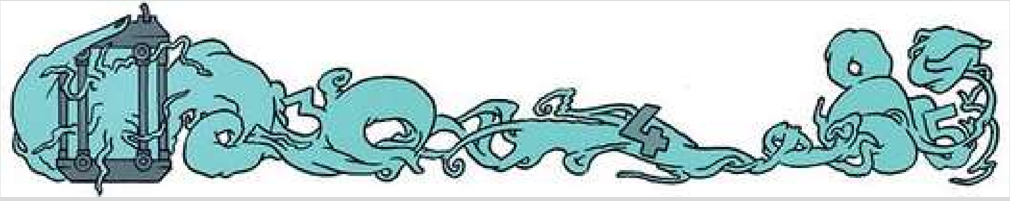
[1] It appears that even then I had a passion for lowering the tone.

[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).

Figure 9.1 Me playing with my ding-a-ling in the Holimarine Talent Show. Note the groupies queuing up at the front

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.’



9.2 An introduction to the linear model (regression)

🟩⬛⬛⬛

9.2.1 The linear model with one predictor 🟩⬛⬛⬛

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):

$$\text{outcome}_i = (b_1 X_i) + \text{error}_i \tag{9.1}$$

I mentioned then that if we work with raw scores we must add information about where the outcome variable is centred. I wrote that we add a constant, b_0, known as the intercept to the model that represents the value of the outcome when the predictor is absent (i.e., it is zero). The resulting model is:

$$\text{outcome}_i = (b_0 + b_1 X_i) + \text{error}_i \tag{9.2}$$

$$Y_i = (b_0 + b_1 X_i) + \varepsilon_i$$

This equation keeps the fundamental idea that an outcome for a person can be predicted from a model (the stuff in parentheses) and some error associated with that prediction (ε_i). We still predict 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_1) and consequently we include a parameter, b_0, that tells us the value of the outcome when the predictor is zero.

As a quick diversion, 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. 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 equations look familiar? 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, Russian or from some other country that has a different variant of the equation of a straight line. The different forms of the equation illustrate how the symbols or letters in an equation can be somewhat arbitrary choices.[3] 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?

[3] For example, you'll sometimes see equation (9.2) written as $Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_i$. The only difference is that this equation has got βs in it instead of bs. Both versions are the same thing, they just use different letters to represent the coefficients.

I have talked throughout this book about fitting 'linear models', and linear simply means ‘straight line’. All the equations above are forms of the equation of 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 throughout this book, where you see them referred to generally as *b* (without any subscript) or ***bi*** (meaning the *b* associated with variable *i*). Figure 9.2 (left) shows a set of lines that have the same intercept but different gradients. For these three models, b_0 is the same in each but b_1 is different for each line. Figure 9.2 (right) shows models that have the same gradients (b_1 is the same in each model) but different intercepts (b_0 is different in each model).

In Chapter 8 we saw how relationships can be either positive or negative (and I don’t mean whether 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 9.2 (left), the orange 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) locates the model in geometric space.

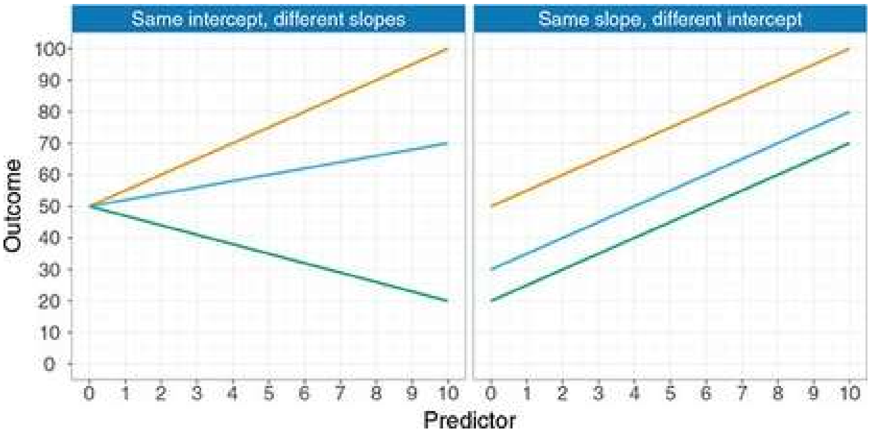
Let’s look at an example. Imagine that I was interested in predicting physical and downloaded album sales (outcome) from the amount of money spent advertising that album (predictor). We could adapt the linear model (equation (9.2)) by replacing the predictor and outcome with our variable names:

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \text{album sales}_i &= b_0 + b_1 \,\text{advertising budget}_i + \varepsilon_i \end{aligned} \tag{9.3}$$

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:

$$\text{album sales}_i = 50 + (100 \times \text{advertising budget}_i) + \varepsilon_i \tag{9.4}$$

Figure 9.2 Lines that share the same intercept but have different gradients, and lines with the same gradients but different intercepts



Note that I have replaced the *b*s 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:

$$\begin{aligned}\text{album sales}_i &= 50 + (100 \times 5) + \varepsilon_i \\ &= 550 + \varepsilon_i\end{aligned} \tag{9.5}$$

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**.

9.2.2 The linear model with several predictors

Life is usually complicated and there will be numerous variables that might be related to the outcome that you want to predict. To take our album sales example, variables other than advertising are likely to affect sales. For example, how much someone hears songs from the album on the radio, or the 'look' of the band. One of the beautiful things about the linear model is that it expands to include as many predictors as you like. We hinted at this in Chapter 2 (equation (2.4)). An additional predictor can be placed in the model and given a *b* to estimate its relationship to the outcome:

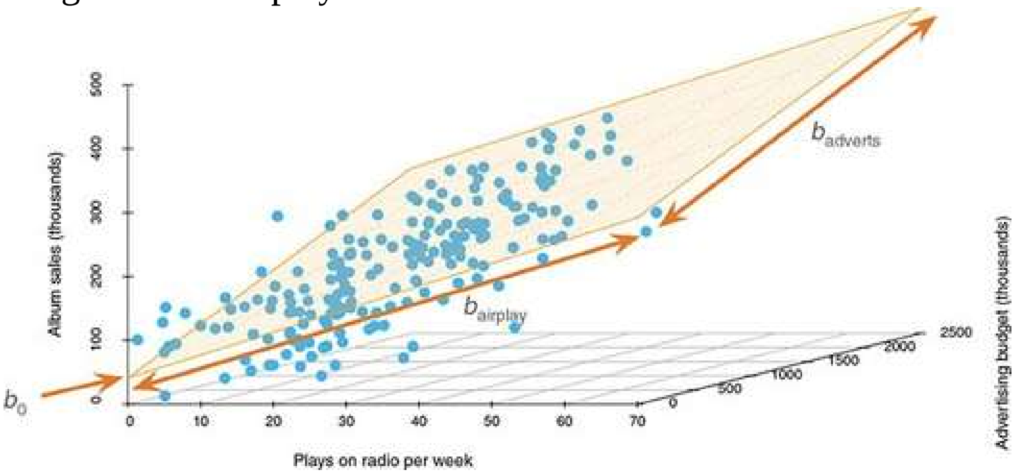
What is the difference between simple and multiple regression?

$$Y_i = (b_0 + b_1X_{1i} + b_2X_{2i}) + \varepsilon_i \tag{9.6}$$

All that has changed is the addition of a second predictor (X_2) and an associated parameter (b_2). To make things more concrete, if we add the number of plays of the band on the radio per week (airplay) to the model in equation (9.3), we get:

$$\text{album sales}_i = b_0 + b_1\text{advertising budget}_i + b_2\text{airplay}_i + \varepsilon_i \tag{9.7}$$

Figure 9.3 Scatterplot of the relationship between album sales, advertising budget and radio play



The new model includes a *b*-value for both predictors (and, of course, the constant, b_0). By estimating the *b*-values, we can make predictions about album sales based not only on the amount spent on advertising but also on airplay.

The resulting model is visualized in Figure 9.3. The tinted trapezium (the regression *plane*) is described by equation (9.7) 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 differences between the model and the real-life data (this fact is evident because most of the dots do not lie exactly on the plane). The vertical distances between the 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 plane, whereas the *b*-value for airplay describes the slope of the top and bottom of the plane. Just like with one predictor, these two slopes describe the shape of the model (what it looks like) and the intercept locates the model in space.

It is easy enough to visualize a linear model with two predictors, because it is possible to plot the plane using a 3-D scatterplot. However, with three, four or even more predictors you can’t immediately visualize what the model looks like, or what the *b*-values represent, but you can apply the principles of these basic models to more complex scenarios. For example, in general, we can add as many predictors as we like, provided we give them a *b*, and the linear model expands accordingly:

$$Y_i = \left(b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni}\right) + \varepsilon_i \tag{9.8}$$

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 nth predictor (X_{ni}), and ε_i is the error for the ith entity. (The parentheses aren't necessary, they're there to make the connection to equation (9.2).) This equation illustrates that we can add predictors to the model until we reach the final one (X_n), and each time we add one, we assign it a regression coefficient (b).

To sum up, regression analysis is a term for fitting a linear model to data and using 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 with several predictors it is called **multiple regression**. Both are merely terms for the linear model.

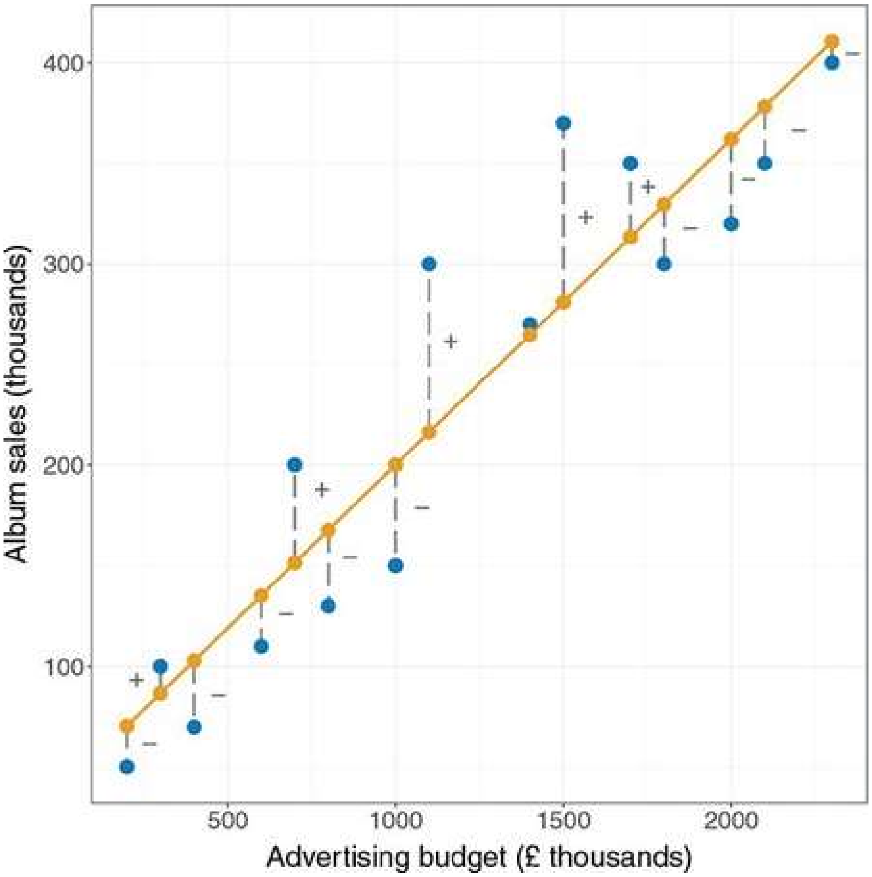
9.2.3 Estimating the model

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 (bs). 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 2.6. 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 data collected. These deviations were the vertical distances between what the model predicted and each data point that was observed. We can do the same to assess the fit of a regression line (or plane).

Figure 9.4 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 orange 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 (9.4)), then insert different values for advertising budget, the predicted values are the resulting estimates of album sales. If we insert the observed values

of advertising budget into the model to get these predicted values, then we can gauge how well the model fits (i.e., makes accurate predictions). If the model is a perfect fit to the data then for a given value of the predictor(s) the model will predict the same value of the outcome as was observed. In terms of Figure 9.4 this would mean that the orange and blue dots fall in the same locations. They don't, because the model is not perfect (and it never will be): sometimes it overestimates the observed value of the outcome and sometimes it underestimates it. With the linear model 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); they are the vertical dashed lines in Figure 9.4.

Figure 9.4 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 observed data



We saw in Chapter 2, equation (2.11), that to calculate the total error in a model we square the differences between the observed values of the outcome, and the predicted values that come from the model:

$$\text{total error} = \sum_{i=1}^{n} (\text{observed}_i - \text{model}_i)^2 \tag{9.9}$$

Sometimes the predicted value of the outcome is less than the actual value and sometimes it is greater. Consequently, some residuals are positive but others are negative, and if we summed them they would cancel out. The solution is to square them before we add them up (this idea should be familiar from Section 2.5.2). Therefore, to assess the error in a linear model, just like when we assessed the fit of the mean using the variance, we use a sum of squared errors, and because we call these errors residuals, this total is called the *sum of squared residuals* or **residual sum of squares** (SS_R). The residual sum of squares is a gauge of how well a linear model fits the data: if the squared differences are large, the model is not representative of the data (there is a lot of error in prediction); if the squared differences are small, the line is representative.

Let’s get back to how we estimate the *b*-values. If you were particularly bored, you could draw every possible straight line (linear model) through your data and calculate the residual sum of squares for each one. You could then compare these ‘goodness-of-fit’ measures and keep the line with the smallest SS_R because it would be the best-fitting model. We have better things to do, so like when we estimate the mean, we use the method of least squares to estimate the parameters (*b*) that define the regression 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 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 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. I’m pretty sure that’s how the method of least squares works.

9.2.4 Assessing the goodness of fit, sums of squares, R and R^2 🟩⬛⬛⬛

Once Nephwick and Clungglewad have found the values of *b* that define the model of best fit we assess how well this model fits the observed data (i.e., the **goodness of fit**). We do this because even though the model is the best one available, it can still be a lousy fit (the best of a bad bunch). We saw above that the residual sum of squares measures how much error there is in the model: it quantifies the error in prediction, but it doesn't tell us whether using the model is better than nothing. We need to compare the model against a baseline to see whether it 'improves' how well we can predict the outcome. So, we fit a baseline model and use equation (9.9) to calculate the fit of this model. Then we fit the best model, and calculate the error, SS_R, within it using equation (9.9). If the best model is any good, it should have significantly less error within it than the baseline model.

What would be a good baseline model? Let's go back to our example of predicting album sales (Y) from the amount of money spent advertising that album (X). In my fictional world where I am a statistician employed by a record company or my favourite football team, my boss one day bursts into my office. He says, 'Andy, I know you wanted to be a rock star but have 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 satisfy a brainless record company executive (who didn't offer my band a record contract). The next day he bursts in again and demands to know how many albums we will sell if we spend £1 on advertising. In the absence of any better information, I'm again going to have to say the average number of sales (200,000). This is getting embarrassing for me: whatever amount of money is spent on advertising, I predict the same levels of sales. My boss will think I'm an idiot.

The mean of the outcome is a model of 'no relationship' between the variables: as one variable changes the prediction for the other remains constant (see Section 3.7.2). I hope this illustrates that the mean of the outcome is a good baseline of 'no relationship'. Using the mean of the outcome as a baseline model, we can calculate the difference between the observed values and the values predicted by the mean (equation (9.9)). We saw in Section 2.5.1 that we square these differences to give us the sum of squared differences. This sum of squared differences is known as the **total sum of squares** (denoted by SS_T) and it represents how good the mean is as a model of the observed outcome scores (Figure 9.5, top left).

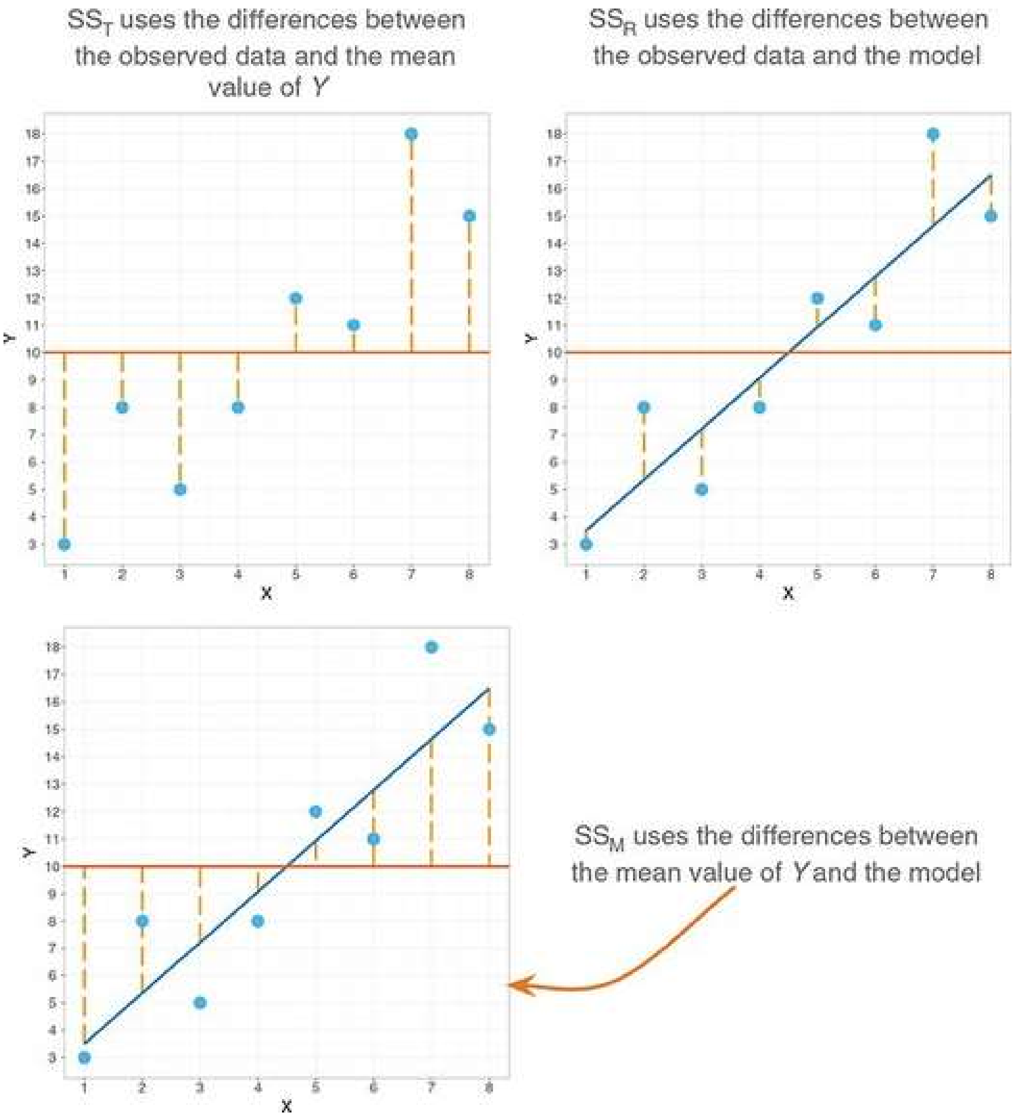
We then fit a more sophisticated model to the data, such as a linear model, and again work out the differences between what this new model predicts and the observed data (again using equation (9.9)). This value is the residual sum of squares (SS_R) discussed in the previous section. It represents the degree of inaccuracy when the best model is fitted to the data (Figure 9.5, top right).

We can use the values of SS_T and SS_R to calculate how much better the linear model is than the baseline model of 'no relationship'. The improvement in prediction resulting from using the linear model rather than the mean is calculated as the difference between SS_T and SS_R (Figure 9.5, bottom). 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 9.5 shows each sum of squares graphically where the model is a line (i.e., one predictor) but the same principles apply with more than one predictor.

If the value of SS_M is large, the linear model is very different from using the mean to predict the outcome variable. This implies that the linear model has made a big improvement to predicting the outcome variable. If SS_M is small then using the linear model is little better than using the mean (i.e., the best model is no better than predicting from 'no relationship'). A useful measure arising from these sums of squares is the proportion of improvement due to the model. This is calculated by dividing the sum of squares for the model by the total sum of squares to give a quantity called R^2:

$$R^2 = \frac{SS_M}{SS_T} \tag{9.10}$$

Figure 9.5 Diagram showing from where the 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 model

SS_M uses the differences between the mean value of Y and the model

To express this value as a percentage multiply it by 100. R^2 represents the amount of variance in the outcome explained by the model (SS_M) relative to how much variation there was to explain in the first place (SS_T); it is the same as the R^2 we met in Section 8.4.2 and it is interpreted in the same way: it represents the proportion of the variation in the outcome that can be predicted from 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 observed values of the outcome.[4] 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.[5]

[4] This is the correlation between the orange and blue dots in Figure 9.4. With

only one predictor in the model this value will be the same as the Pearson correlation coefficient between the predictor and outcome variable.

[5] 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.

A second use of the sums of squares in assessing the model is 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 error in the model (SS_R). I say 'based upon' because the sums of squares depend on the number of differences that were added up, and so the average sums of squares (referred to as the **mean squares** or MS) are used to compute *F*. The mean sum of squares is the sum of squares divided by the associated degrees of freedom (this is comparable to calculating the variance from the sums of squares – see Section 2.5.2). For SS_M the degrees of freedom are the number of predictors in the model (*k*), and for SS_R they are the number of observations (*N*) minus the number of parameters being estimated (i.e., the number of *b* coefficients including the constant). We estimate a *b* for each predictor and the intercept (b_0), so the total number of *b*s estimated will be $k + 1$, giving us degrees of freedom of $N - (k + 1)$ or, more simply, $N - k - 1$. Thus

$$\text{MS}_\text{M} = \frac{\text{SS}_\text{M}}{k} \qquad \text{MS}_\text{R} = \frac{\text{SS}_\text{R}}{N - k - 1} \tag{9.11}$$

There is more on mean squares in Chapter 12. The ***F*-statistic** computed from these mean squares,

$$F = \frac{\text{MS}_\text{M}}{\text{MS}_\text{R}} \tag{9.12}$$

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 the improvement in prediction from using the model should be large (MS_M will be large) and the difference between the model and the observed data should be small (MS_R will be small). In short, for a good model the numerator in equation (9.12) will be bigger than the denominator, resulting in a large *F*-statistic (greater than 1 at least).

This *F* has an associated probability distribution from which a *p*-value can be derived to tell us the probability of getting an *F* at least as big as the one we have if the null hypothesis were true. The null hypothesis in this case is a flat model (predicted values of the outcome are the same regardless of the value of the predictors). If you want to go old school, you can compare the *F*-statistic against critical values for the corresponding degrees of freedom (as in the Appendix).

The *F*-statistic is also used to calculate the significance of R^2 using the following equation:

$$F = \frac{(N-k-1)R^2}{k(1-R^2)} \tag{9.13}$$

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).

9.2.5 Assessing individual predictors

We've seen that any predictor in a linear model has a coefficient (b_1). The value of *b* represents the change in the outcome resulting from a unit change in a predictor. If a predictor was useless at predicting the outcome, then what might we expect the change in the outcome to be as values of the predictor change? If a predictor had 'no relationship' with the outcome then the change would be zero. Think back to Figure 9.5. In the panel representing SS_T we saw that the line representing 'no relationship' or 'mean of the outcome' is flat: as the predictor variable changes, the predicted value of the outcome does *not* change (it is a constant value). A 'flat' model, a model in which the same predicted value arises from all values of the predictor variables, will have *b*-values 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 is constant); and (2) the linear model is 'flat' (the line or plane doesn't deviate from the horizontal). Therefore, logically, if a variable significantly predicts an outcome, it should have a *b*-value that is *different* from zero. This hypothesis is tested using a ***t*-statistic** that tests the null hypothesis that the value of *b* is 0. If the test is significant, we might interpret this information as supporting a 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 based on the ratio of explained variance against unexplained variance or error. 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. Remember that the standard error for *b* tells us something about how different *b*-values would be across different samples (think back to Section 2.7). If the standard error is very small, then most samples are likely to have a *b*-value similar to the one in our sample (because there is little variation across samples). Therefore, the standard error is a good estimate of how much error there is likely to be in our *b*.

The following equation shows how the *t*-test is calculated:

$$t = \frac{b_{\text{observed}} - b_{\text{expected}}}{SE_b} = \frac{b_{\text{observed}}}{SE_b} \qquad (9.14)$$

You'll find a general version of this equation in Section 10.5.1 (equation (10.5)). The b_{expected} is the value of *b* that we would expect to obtain if the null hypothesis were true. The null hypothesis is that *b* is 0, and so this value is replaced by 0 and drops out of the equation. The resulting *t* is the observed value of *b* divided by the standard error with which it is associated. The *t*, therefore, tells us whether the observed *b* is different from 0 relative to the variation in *b*s across samples. When the standard error is small even a small deviation from zero can reflect a significant difference because *b* is representative of the majority of possible samples.

The statistic *t* has a probability distribution that differs according to the degrees of freedom for the test. In this context, the degrees of freedom are $N - k - 1$, where *N* is the total sample size and *k* is the number of predictors. With only one predictor, this reduces to $N - 2$. Using the appropriate *t*-distribution, it's possible to calculate a *p*-value that indicates the probability of getting a *t* at least as large as the one we observed if the null hypothesis were true (i.e., if *b* was in fact 0 in the population). If this observed *p*-value is less than 0.05, then scientists tend to assume that *b* is significantly different from 0; put another way, the predictor makes a significant contribution to predicting the outcome. However, remember the potential pitfalls of blindly applying this 0.05 rule. If you want to pretend it's 1935 then instead of computing an exact *p*, you can compare your observed *t* against critical values in a table (in the Appendix).

9.3 Bias in linear models?

In Chapter 6 we saw that statistical models can be biased by unusual cases or by failing to meet certain assumptions. Therefore, the next questions to ask are whether the model: (1) is influenced by a small number of cases; and (2) generalizes 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** (Section 9.4) is a critical additional step, and if we find that our model is not generalizable, then we must restrict any conclusions to the sample used. First, let's look at bias. 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 9.1).

9.3.1 Outliers

An outlier is a case that differs substantially from the main trend in the data (see Section 6.3). Outliers can affect the estimates of the regression coefficients. For example, Figure 9.6 uses the same data as Figure 9.4 except that the score of one album has been changed to be an outlier (in this case an album that sold relatively few copies despite a very large advertising budget). The blue line shows the original model, and the orange line shows the model with the outlier included. The outlier makes the line flatter (i.e., b_1 gets smaller) and increases the intercept (b_0 gets larger). If outliers affect the estimates of the *b*s that define the model then it is important to detect them. But how?

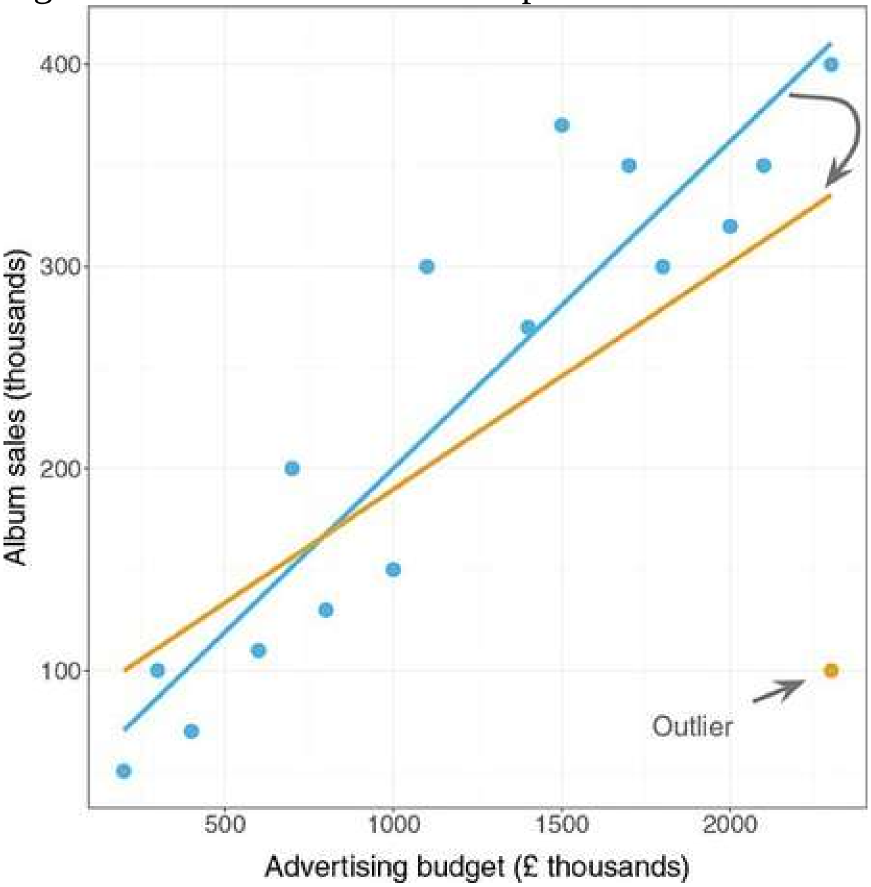
An outlier, by its nature, is very different from the other scores. In which case, do you think that the model will predict an outlier's score very accurately? Probably not: in Figure 9.6 it's evident that even though the outlier has dragged the model towards it, the model still predicts it very badly (the line is a long way from the outlier). Therefore, if we compute the residuals (the differences between the observed values of the outcome and the values predicted by the model), outliers could be spotted because they'd have large values. In other words, we'd look for cases that the model predicts inaccurately.

SELF TEST

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

Remember that 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 to the sample data then the residuals will be large. Up to now we have discussed *normal* or **unstandardized residuals**. These are the raw differences between predicted and observed values of the outcome variable. They are measured in the same units as the outcome variable, which makes it difficult to apply general rules (because what constitutes 'large' depends on the outcome variable). All we can do is to look for residuals that stand out as being particularly large.

Figure 9.6 Graph demonstrating the effect of an outlier. The blue line represents the original regression line for these data, whereas the orange line represents the regression line when an outlier is present



To overcome this problem, we can use **standardized residuals**, which are the residuals converted to *z*-scores (see Section 1.8.6) and so are expressed in standard deviation units. Regardless of the variables in your model, standardized residuals (like any *z*-scores) are distributed around a mean of 0 with a standard deviation of 1. Therefore, we can compare standardized residuals from different models and use what we know about *z*-scores to apply universal guidelines for what is expected. For example, 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 (see Chapter 1). Based on this: (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 (2.5 will do) there is evidence that the level of error within our model may be unacceptable; and (3) if more than 5% of cases have standardized residuals with an absolute value greater than 1.96 (2 for convenience) then the model may be a poor representation of the 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.

9.3.2 Influential cases

It is also possible to look at whether certain cases exert undue influence over the parameters of the model. In other words, if we were to delete a certain case, how different would the regression coefficients be? This analysis helps to determine whether the model is stable across the sample, or whether it is biased by a few influential cases. This process can also unveil outliers.

There are several statistics used to assess the influence of a case. The **adjusted predicted value** for a case is the predicted value of the outcome for that case from a model in which the case is excluded. In effect, you estimate the model parameters excluding a particular case and use this new model to predict the outcome for the case that was excluded. If a case does not exert a large influence over the model then the adjusted predicted value should be 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 that case was used to estimate the model.

We can also look at the **deleted residual**, which is the difference between the adjusted predicted value and the original observed value. 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.

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). **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.

The **leverage** (sometimes called **hat values**) 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 cases.[6] The maximum value for leverage is $(N - 1)/N$; however, IBM SPSS Statistics calculates a version of the leverage that has a maximum value of 1 (indicating that the case has complete influence over prediction).

[6] 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 $(k + 1)$.

- If no cases exert undue influence over the model then all leverage values should be close to the average value $((k + 1)/n)$.
- We should investigate cases with values greater than twice the average, $2(k + 1)/n$ (Hoaglin & Welsch, 1978), or three times the average, $3(k + 1)/n$ (Stevens, 2002).

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, not 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 $p = 0.05$ and 0.01 are in the Appendix). For example, with three predictors, a distance greater than 7.81 ($p = 0.05$) or 11.34 ($p = 0.01$) would be cause for concern. As general context, based on Barnett and Lewis (1978), with large samples ($N = 500$) and five predictors, values above 25 are cause for concern. In smaller samples ($N = 100$) and fewer predictors (namely, three), values greater than 15 are problematic. In very small samples ($N = 30$) with only two predictors, values greater than 11 should be examined.

Another approach is to look at how the estimates of b in a model change as a result of excluding a case (i.e., compare the values of b estimated from the full data to those estimated from the data excluding the particular case). The change in *bs* tells us how much influence a case has on the parameters of the model. To take a hypothetical example, imagine two variables that have a perfect negative relationship except for a single case (case 30). These data are in the file **DFBeta.sav.**

Figure 9.7 Prasanta Chandra Mahalanobis staring into his distances

SELF TEST

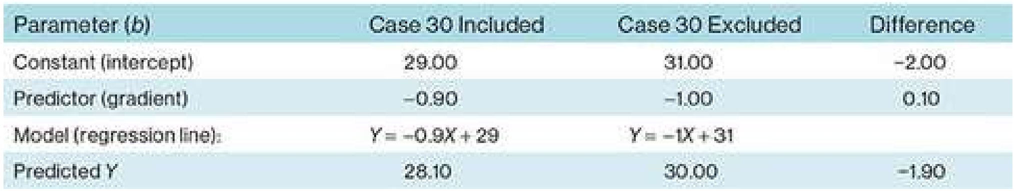
Once you have read Section 9.7, fit a linear model first with all the cases included and then with case 30 deleted.

The results of these two models are summarized in Table 9.1, 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* predicted from participant 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, and the coefficient for the constant (the intercept, b_0) is 31. However, when case 30 is included, both parameters are reduced[7] 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. The values of DFBeta help us to identify cases that have a large influence on the parameters of the model. The units of measurement used will affect these values, and so you can use **standardized DFBeta** to apply universal cut-off points. Standardized DFBetas with 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).

7 The value of b_1 is reduced because the variables no longer have a perfect linear relationship and so there is now variance that the predictor cannot explain.

A related statistic is the **DFFit**, which is the difference between the predicted values for a case when the model is estimated including or excluding that case: in this example the value is −1.90 (see Table 9.1). If a case has no influence then its DFFit should be zero – hence, we expect non-influential cases to have small DFFit values. As with DFBeta, 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. To overcome this problem we can look at standardized versions of the DFFit values (**standardized DFFit**) which are expressed in standard deviation units. A final measure is the **covariance ratio (CVR)**, which quantifies the degree to which 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 is having very little influence on the variances of the model parameters. Belsey, Kuh, & Welsch (1980) recommend the following:

Table 9.1 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

- 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 inequalities, *k* is the number of predictors, CVR_i is the covariance ratio for the *i*th participant, and *n* is the sample size.

9.3.3 A final comment on diagnostic statistics

I’ll conclude this section with a point made by Belsey *et al.* (1980): diagnostics are tools to see how well your model fits the sampled data and *not* 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) similarly notes that if a case is a significant outlier but is not having an influence (e.g., Cook’s distance is less than 1, DFBetas and DFFit are small) there is no real need to worry about that point because it’s not having a large impact on the model parameters. Nevertheless, you should still be interested in *why* the case didn’t fit the model.

9.4 Generalizing the model

The linear model produces an equation that is correct for the sample of observed values. However, we are usually interested in generalizing our findings beyond the sample. For a linear model to generalize the underlying assumptions must be met, and to test whether the model does generalize we can cross-validate it.

9.4.1 Assumptions of the linear model

We have already looked at the main assumptions of the linear model and how to assess them in Chapter 6. The main ones in order of importance (Field & Wilcox, 2017; Gelman & Hill, 2007) are:

- *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 6).
- **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 this assumption then the model standard errors will be invalid, as will the confidence intervals and significance tests based upon them. In terms of the model parameters themselves, the estimates from the method of least squares will be valid but not optimal (see Section 6.8). 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. As such it is affected by the order of cases and only makes sense when your cases have a

meaningful order (which they don’t in the album sales example). The test statistic varies 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. If this test is relevant to you, look up the critical values in Durbin and Watson (1951). As a very conservative rule of thumb, values less than 1 or greater than 3 are cause for concern.

- *Homoscedasticity* (see Section 6.7): At each level of the predictor variable(s), the variance of the residual terms should be constant. This assumption 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 confidence intervals and significance tests; estimates of the model parameters (*b*) using the method of least squares are valid but not optimal. This problem is overcome using weighted least squares regression, in which each case is weighted by a function of its variance, or using robust regression.
- *Normally distributed errors* (see Section 6.6): It can be helpful if the residuals in the model are random, normally distributed variables with a mean of 0. This assumption means that the differences between the predicted and 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, which they don’t. In small samples a lack of normality invalidates confidence intervals and significance tests, whereas 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 can ignore this assumption.

Jane Superbrain 9.1 The difference between residuals and influence statistics

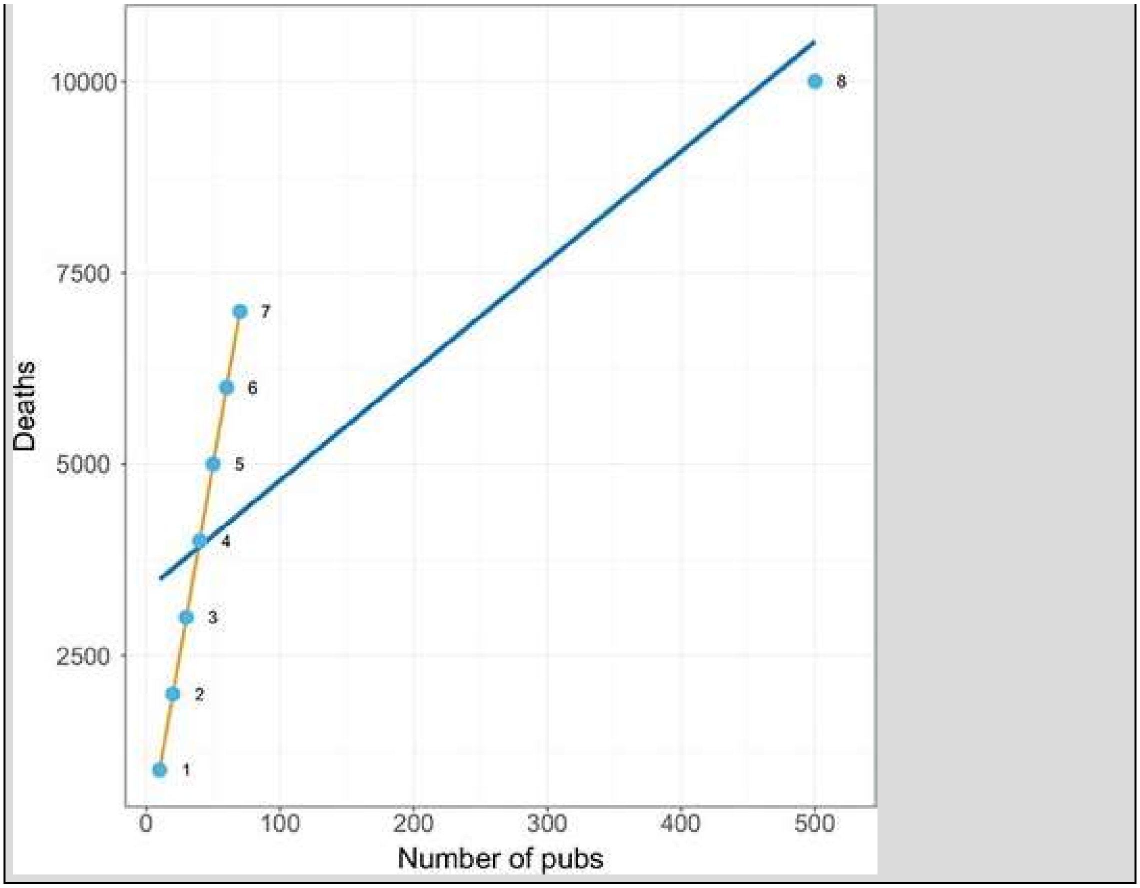
To illustrate how residuals and influence statistics differ, imagine that the Mayor of London in 1900 was interested in how drinking affected mortality. London is divided up into different regions called boroughs, and so he measured the number of pubs and the number of deaths over a period of time in eight of his boroughs. The data are in a file called **pubs.sav**.

The scatterplot of these data (Figure 9.8) reveals that without the last case there is a perfect linear relationship (the orange 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 – fit the model and see for yourself).

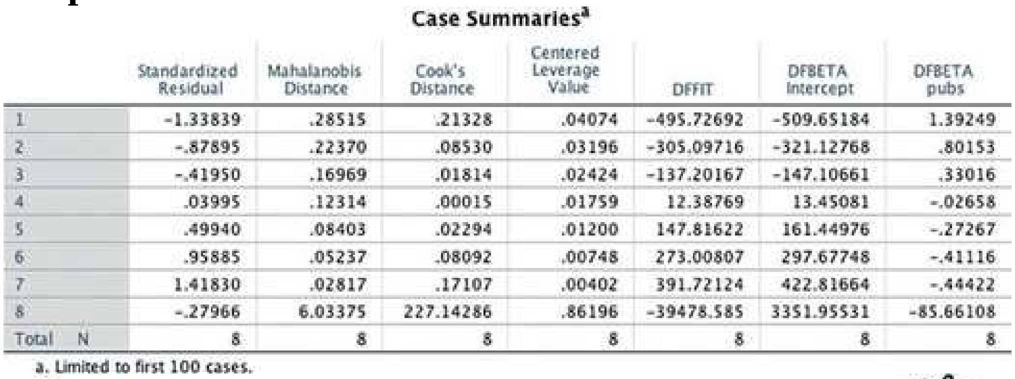
The residuals and influence statistics are interesting (Output 9.1). The standardized residual for case 8 is the second *smallest*: it produces a very small residual (most of the non-outliers have larger residuals) because it sits very close to the line that has been fitted to the data. According to the residual it is not an outlier, but how is that possible when it is so different from the rest of the data? The answer lies in the influence statistics, which are all massive for case 8: it exerts a huge influence over the model – so huge that the model predicts that case very well.

When you see a statistical oddity like this, ask what's happening in the real world. District 8 is 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 needed pubs. Therefore, there was a massive number of pubs. (I'm very grateful to David Hitchin for this example, and he in turn got it from Dr Richard Roberts.)

Figure 9.8 Relationship between the number of pubs and the number of deaths in 8 London districts



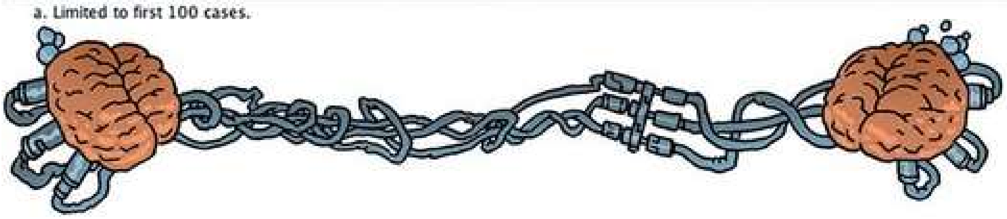
Output 9.1



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	-39478.585	3351.95531	-85.66108
Total N	8	8	8	8	8	8	8

a. Limited to first 100 cases.



There are other considerations that we haven't touched on (see Berry, 1993):

- *Predictors are uncorrelated with 'external variables'*: *External variables* are variables that haven't been included in the model and that influence the outcome variable.[8] These variables are like the 'third variable' that we discussed in the correlation chapter. 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 9.9.3).
- *Non-zero variance*: The predictors should have some variation in value (i.e., they should not have variances of 0). This is self-evident really.

[8] 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.

As we saw in Chapter 6, 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 95% confidence interval for a *b* tells us the boundaries within which the population values of that *b* are likely to fall.[9] Therefore, if confidence intervals are inaccurate (as they are when these assumptions are broken) we cannot accurately estimate the likely population value. In other words, 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 is not the same as the population model – but the likelihood of them being the same is increased.

[9] Assuming your sample is one of the 95% that generates a confidence interval containing the population value. Yes, I do have to keep making this point – it's important.

9.4.2 Cross-validation of the model

Even if we can't be confident that the model derived from our sample accurately represents the population, we can assess how well our model might 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 does *not* generalize. First, we should collect enough data to obtain a reliable model (see the next section). Once we have a estimated the model there are two main methods of cross-validation:

- *Adjusted R^2*: Whereas R^2 tells us how much of the variance in *Y* overlaps with predicted values from the model in our sample, **adjusted *R*2** 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 model would predict scores of a different sample of data from the same population. Stein's formula,

$$\text{adjusted } R^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\right]\left(1 - R^2\right) \tag{9.15}$$

- does tell us how well the model cross-validates (see Stevens, 2002), and the more mathematically minded of you might want to try using it instead of what SPSS chugs out. In Stein's formula, R^2 is the unadjusted value, *n* is the number of cases and *k* is the number of predictors in the model.

- *Data splitting*: This approach involves randomly splitting your sample data, estimating the model in both halves of the data and comparing the resulting models. When using stepwise methods (see Section 9.9.1), 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 R^2 and b in the two samples you can tell how well the original model generalizes (see Tabachnick & Fidell, 2012).

9.5 Sample size and the linear model

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 have normal sampling distributions because of the central limit theorem (Section 6.6.1). 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. These rules are very pervasive but they oversimplify the issue to the point of being useless. The sample size required depends 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, $R = 6/(21 - 1) = 0.3$ (a medium effect size by Cohen's criteria described in Section 3.7.2). 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 0.06).

Figure 9.9 The sample size required to test the overall regression model depending on the number of predictors and the size of expected effect, $R^2 = 0.02$ (small), 0.13 (medium) and 0.26 (large)

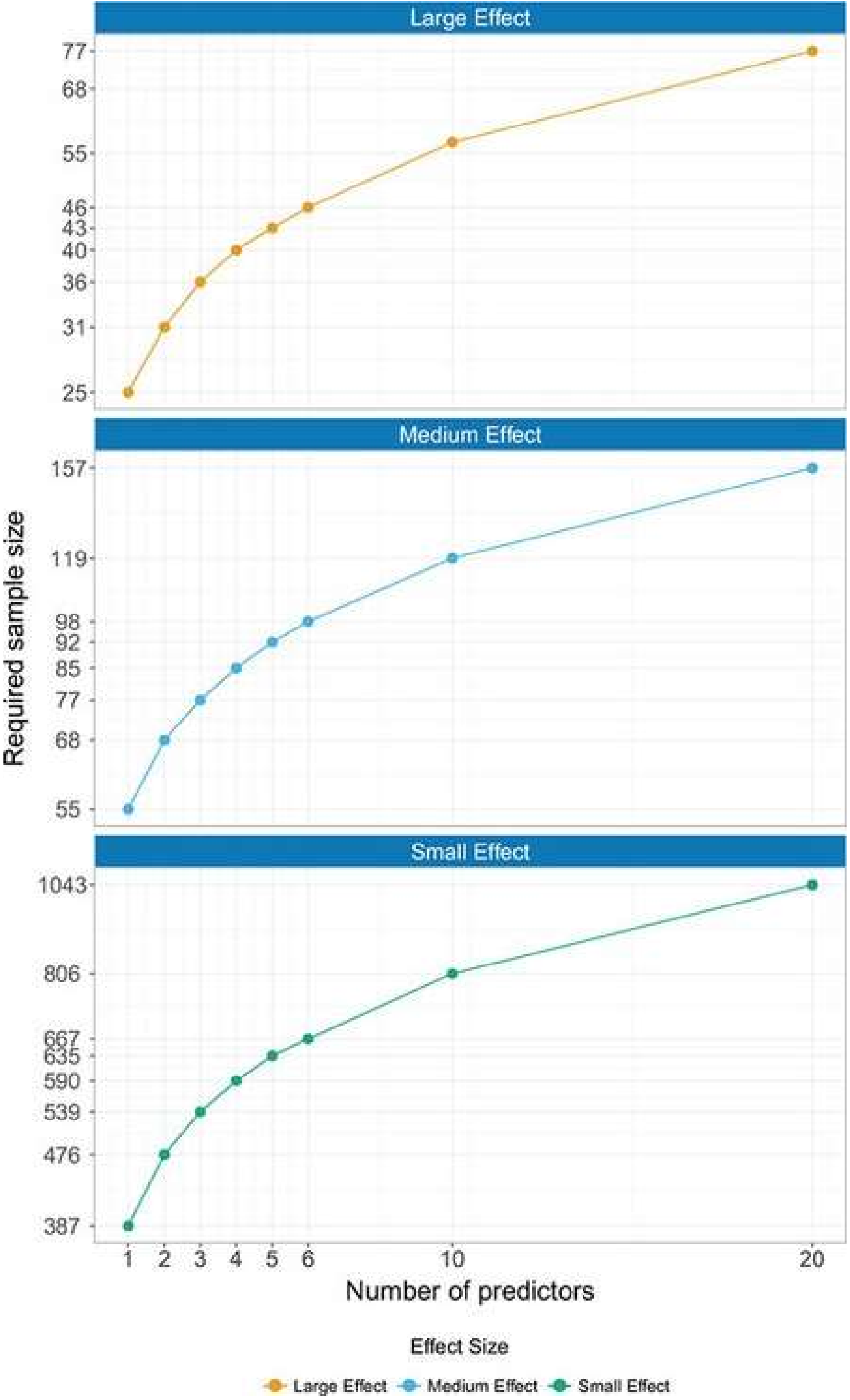


Figure 9.9 shows the sample size required[10] to achieve a high level of power (I've taken Cohen's, 1988, benchmark of 0.8) to test that the model is significant

overall (i.e., R^2 is not equal to zero). I've varied the number of predictors and the size of expected effect: I used $R^2 = 0.02$ (small), 0.13 (medium) and 0.26 (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 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.

[10] I used the program G*power, mentioned in Section 2.9.8, to compute these values.

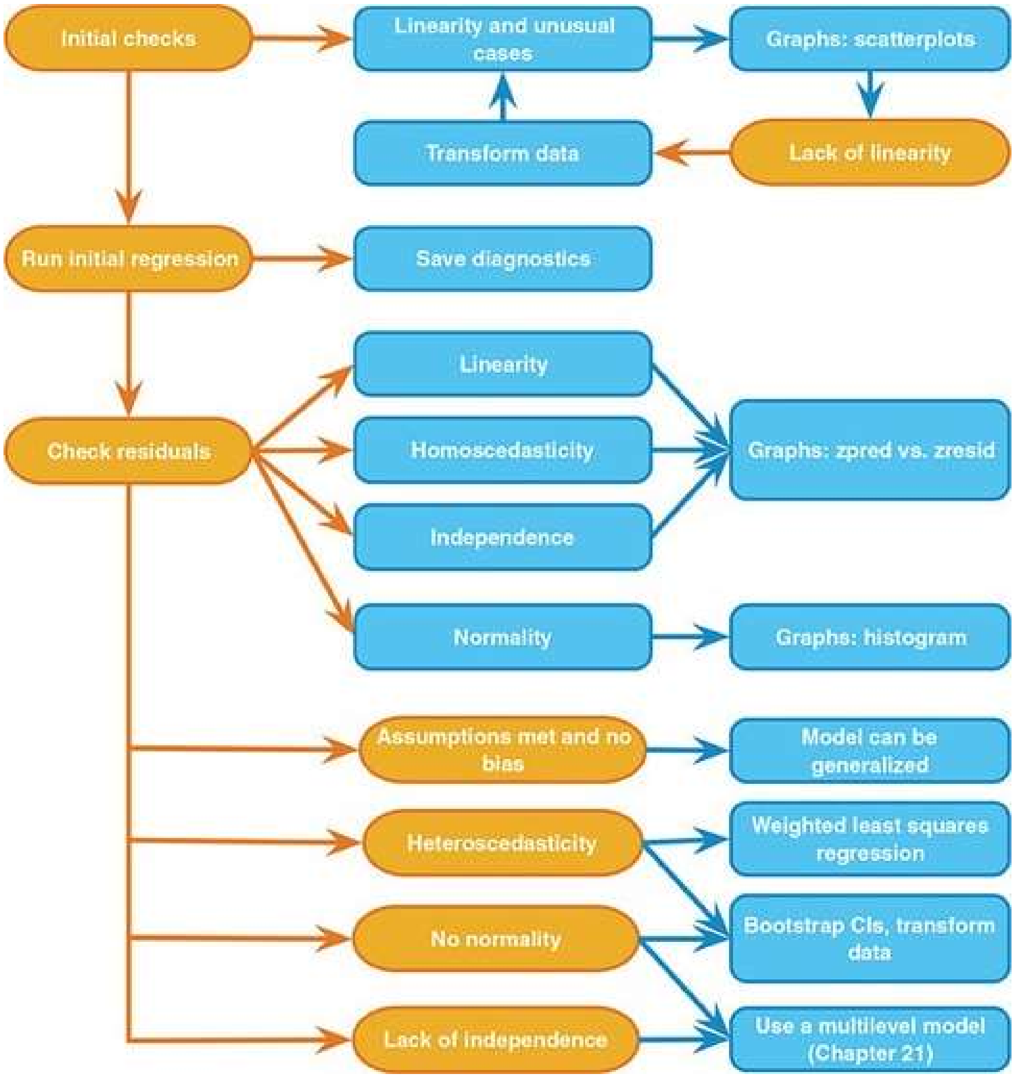
9.6 Fitting linear models: the general procedure

Figure 9.10 shows the general process of fitting linear models. First, we should produce scatterplots to get some idea of whether the assumption of linearity is met, and to look for 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 9.3. 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.

9.7 Using SPSS Statistics to fit a linear model with one predictor

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 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 a prominent national radio station in the week before release (**Airplay**), and how attractive people found the band's image out of 10 (**Image**). Ignore these last two variables for now; we'll use them later. Note how the data are laid out (Figure 9.11): 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 the week before release, and was made by a band with a pretty sick image.

Figure 9.10 The process of fitting a regression model

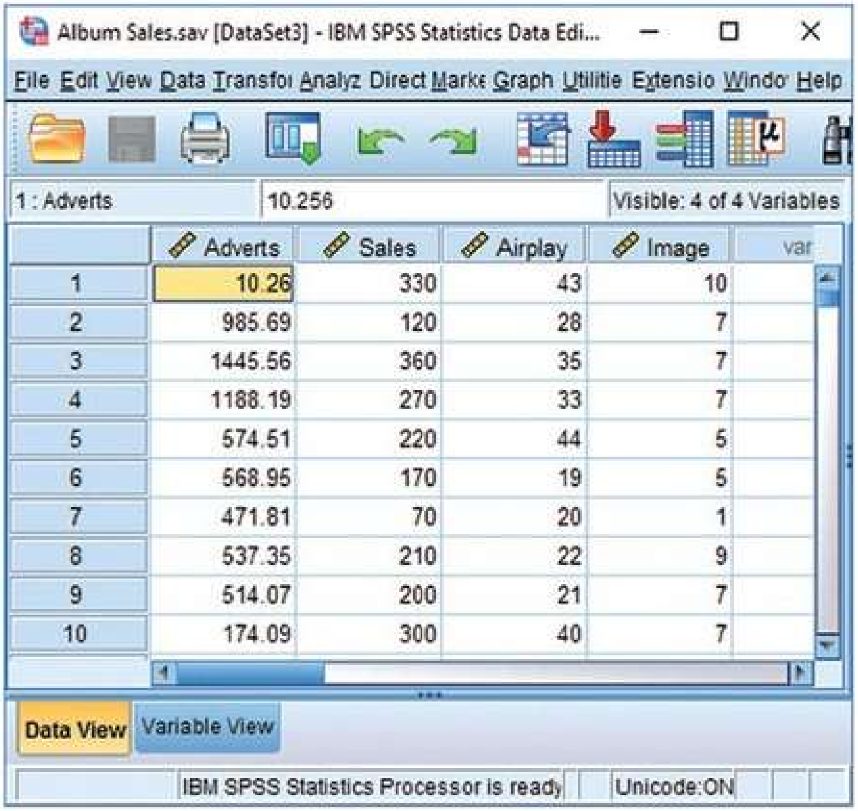


Produce a scatterplot of sales (*y*-axis) against advertising budget (*x*-axis). Include the regression line.

Figure 9.12 shows that a positive relationship exists: the more money spent advertising the album, the more it sells. 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 shows the line of best fit for these data: bearing in mind that the mean would be represented by a flat line at around the 200,000 sales mark, the regression line is noticeably different.

To fit the model, access the main dialog box by selecting *Analyze* ▸ *Regression* ▸ Linear... (Figure 9.13). First, we define the outcome variable (in this example **Sales**). Select **Sales** from the list on the left-hand side, and transfer it to the space labelled *Dependent* by dragging it or clicking ➡. In this model we're going to enter only one predictor (**Adverts**) so select it from the list and click ➡ (or drag it) to transfer it to the box labelled *Independent(s)*. There are a tonne of options available, but we'll explore these when we build up the model in due course. For now, request bootstrapped confidence intervals for the regression coefficients by clicking Bootstrap... (see Section 6.12.3). Select ☑ Perform bootstrapping to activate bootstrapping, and to get a 95% confidence interval select ◉ Bias corrected accelerated (BCa). Click OK in the main dialog box to fit the model.

Figure 9.11 The data editor for fitting a linear model


Album Sales.sav [DataSet3] - IBM SPSS Statistics Data Edi...
File Edit View Data Transfoı Analyz Direct Marke Graph Utilitie Extensio Windo Help
Data View
Variable View

1 : Adverts 10.256 Visible: 4 of 4 Variables

	Adverts	Sales	Airplay	Image	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	
10	174.09	300	40	7	

IBM SPSS Statistics Processor is ready Unicode:ON

Figure 9.12 Scatterplot showing the relationship between album sales and the amount spent promoting the album

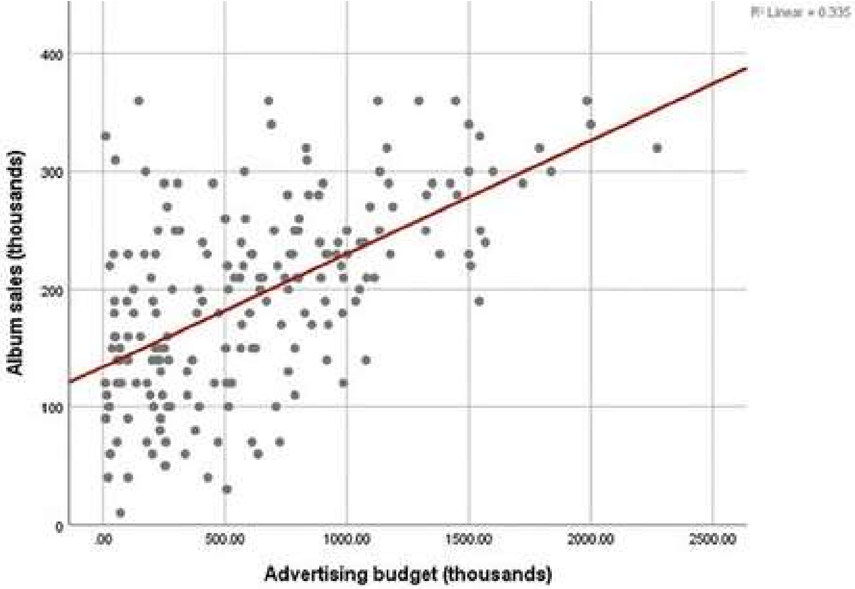
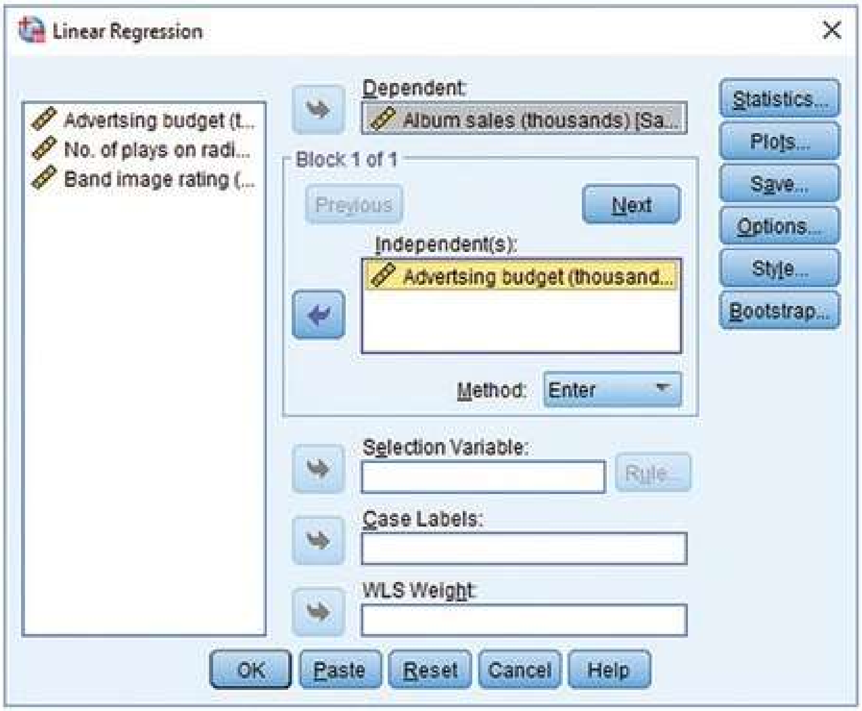


Figure 9.13 Main dialog box for regression



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

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

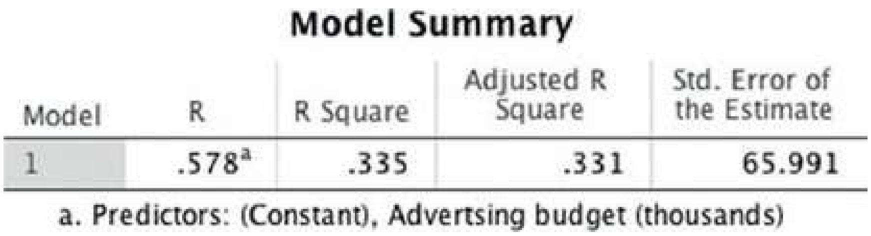
Method: Enter

9.8 Interpreting a linear model with one predictor

9.8.1 Overall fit of the model

The first table is a summary of the model (Output 9.2). This summary table provides the value of *R* and R^2 for the model. For these data, *R* has a value of 0.578 and because there is only one predictor, this value is the correlation between advertising and album sales (you can confirm this by running a correlation using what you learnt in Chapter 8). The value of R^2 is 0.335, which tells us that advertising expenditure can account for 33.5% of the variation in album sales. This means that 66.5% of the variation in album sales remains unaccounted for: there might be other variables that have an influence also.

Output 9.2

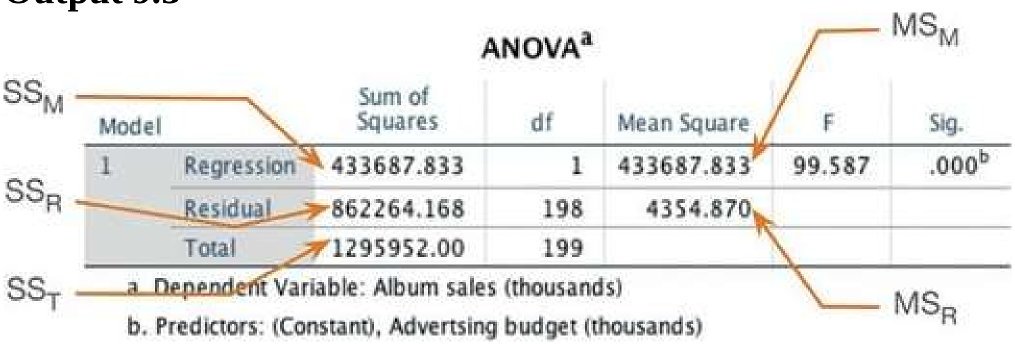


Model Summary

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)

Output 9.3



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[b]
	Residual	862264.168	198	4354.870		
	Total	1295952.00	199			

a. Dependent Variable: Album sales (thousands)

b. Predictors: (Constant), Advertsing budget (thousands)

SS_M
MS_M
SS_R
SS_T
MS_R

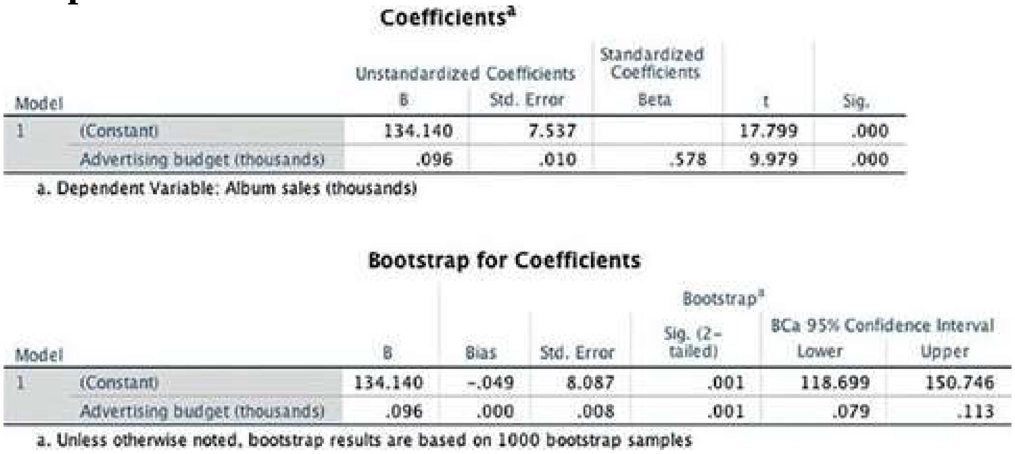
The next part of the output (Output 9.3) reports the various sums of squares described in Figure 9.5, the degrees of freedom associated with each and the resulting mean squares (equation (9.11)). The most important part of the table is the *F*-statistic (equation (9.12)) of 99.59 and its associated significance value of $p < 0.001$ (expressed this way because the value in the column labelled *Sig.* is less than 0.001). This *p*-value tells us that there is less than a 0.1% chance that an *F*-statistic at least this large would happen if the null hypothesis were true. Therefore, we could conclude that our model results in significantly better prediction of album sales than if we used the mean value of album sales. In short, the linear model overall predicts album sales significantly.

9.8.2 Model parameters

Output 9.4 provides estimates of the model parameters (the beta values) and the significance of these values. We saw in equation (9.2) that b_0 was the *Y* intercept, and this value is 134.14 (*B* for the constant in Output 9.4). This value 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 is thousands of albums). We can also read off the value of b_1 from the table, which is 0.096. Although this value is the slope of the line for the model, it is more useful to think of this value as representing *the change in* *the outcome associated with a unit change in the predictor*. In other words, 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. This investment is pretty useless for the record company: it invests £1000 and gets only 96 extra sales! Fortunately, as we already know, advertising accounts for only one-third of album sales.

We saw earlier that if a predictor is having a significant impact on our ability to predict the outcome then its *b* should be different from 0 (and large relative to its standard error). We also saw that the *t*-test and associated *p*-value tell us whether the *b*-value is significantly different from 0. The column *Sig.* contains the exact probability that a value of *t* at least as big as the one in the table would occur if the value of *b* in the population were zero. If this probability is less than 0.05, then people interpret that as the predictor being a ‘significant’ predictor of the outcome (see Chapter 2). For both *t*s, the probabilities are given as 0.000 (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 0.001. In other words, 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 ($p < 0.001$) to predicting album sales.

Output 9.4



Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	134.140	7.537		17.799	.000
	Advertising budget (thousands)	.096	.010	.578	9.979	.000

a. Dependent Variable: Album sales (thousands)

Bootstrap for Coefficients

			Bootstrap[a]				
						BCa 95% Confidence Interval	
Model		B	Bias	Std. Error	Sig. (2-tailed)	Lower	Upper
1	(Constant)	134.140	-.049	8.087	.001	118.699	150.746
	Advertising budget (thousands)	.096	.000	.008	.001	.079	.113

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

If our sample is one of the 95% producing confidence intervals that contain the population value then the bootstrap confidence interval tells us that the population value of *b* for advertising budget is likely to fall between 0.079 and 0.113, and because this interval doesn't include zero we might 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 $p = 0.001$, which is highly significant. Note that the bootstrap process involves re-estimating the standard error (it changes from 0.01 in the original table to a bootstrap estimate of 0.008). This is a very small change. The bootstrap confidence intervals and significance values are useful to report and interpret because they do not rely on assumptions of normality or homoscedasticity.

How is the *t* in Output 9.4 calculated? Use the values in the table to see if you can get the same value as SPSS.

9.8.3 Using the model

We have discovered that we have a model that significantly improves our ability to predict album sales. The next stage is often to use that model to make predictions about the future. First we specify the model by replacing the *b*-values in equation (9.2) with the values from Output 9.4. We can also replace the *X* and *Y* with the variable names:

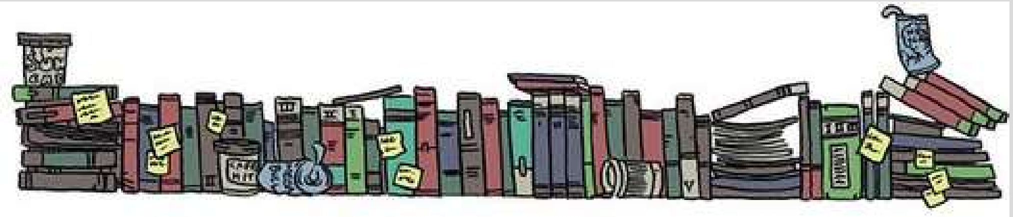
$$\begin{aligned}\text{album sales}_i &= b_0 + b_1\text{advertising budget}_i \\ &= 134.14 + (0.096 \times \text{advertising budget}_i)\end{aligned} \tag{9.16}$$

We can make a prediction about album sales by replacing the advertising budget with a value of interest. For example, if we spend £100,000 on advertising a new album, remembering that our units are already in thousands of pounds, we simply replace the advertising budget with 100. We discover that album sales should be around 144,000 for the first week of sales:

$$\begin{aligned}\text{album sales}_i &= 134.14 + \left(0.096 \times \text{advertising budget}_i\right) \\ &= 134.14 + \left(0.096 \times 100\right) \\ &= 143.74\end{aligned} \tag{9.17}$$

Cramming Sam’s Tips Linear models

- A linear model (regression) is a way of predicting values of one variable from another based on a model that describes a straight line.
- This line is the line that best summarizes the pattern of the data.
- To assess how well the model fits the data use:
  - 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.* < 0.05 in the SPSS output) then the predictor variable significantly predicts the outcome variable.



How many albums would be sold if we spent £666,000 on advertising the latest album by Deafheaven?

9.9 The linear model with two or more predictors (multiple regression)

Imagine that the record company executive wanted to extend the model of albums sales to incorporate other predictors. Before an album is released, the executive notes the amount spent on advertising, the number of times songs from the album are played on a prominent radio station the week before release (**Airplay**), and ratings of the band’s image (**Image**). He or she does this for 200 albums (each by a different band). The credibility of the band’s image was rated by a random sample of the target audience on a scale from 0 (dad dancing at a disco) to 10 (sicker than a dog that’s eaten a bag of onions). The mode rating was used because the executive was interested in what most people thought, not the average opinion.

When we build a model with several predictors, everything we have discussed so far applies. However, there are some additional things to think about. The first is what 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 (and the order in which they are entered). *Do not enter hundreds of predictors, just because you’ve measured them, and expect the resulting model to make sense*. SPSS Statistics will happily generate output based on any garbage you decide to feed it – it will not judge you, but others will. 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 add noise to the model. If predictors are being added that have never been looked at before (in your research context) then select these variables based on their substantive *theoretical* importance. The key point is that the most important thing when building a model is to use your brain – which is slightly worrying if your brain is as small as mine.

[11] Preferably past research that is methodologically and statistically rigorous and yielded reliable, generalizable models.

9.9.1 Methods of entering predictors into the model

Having chosen predictors, you must decide the order to enter them into the model. When predictors are completely uncorrelated the order of variable entry has very little effect on the parameters estimated; however, we rarely have uncorrelated predictors, and so the method of variable entry has consequences and is, therefore, important.

Other things being equal, use **hierarchical regression**, in which you select predictors based on past work and decide in which order to enter them into the model. Generally speaking, you should enter known predictors (from other research) into the model first in order of their importance in predicting the outcome. After having entered known predictors, you can add new predictors into the model simultaneously, in a stepwise manner, or hierarchically (entering the new predictor suspected to be the most important first).

An alternative is forced entry (or *Enter* as it is known in SPSS), in which you force all predictors into the model simultaneously. Like hierarchical, this method relies on good theoretical reasons for including the chosen predictors, but unlike hierarchical, you make 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.

The final option, **stepwise regression**, is generally frowned upon by statisticians. Nevertheless, SPSS Statistics makes it easy to do and actively encourages it in the *Automatic Linear Modeling* 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 900 pages of my drivel unless you wanted to know better, so we'll give stepwise a wide berth. However, you probably ought to know what it does so you can understand why to avoid it. The stepwise method bases decisions about the order in which predictors enter the model 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 model's ability to predict the outcome then it is retained and the computer looks to add a second predictor from the available pool of variables. The next predictor the computer tries will be the one that has the largest semi-partial correlation with the outcome. Remember that the semi-partial correlation quantifies the unique overlap between two variables *X* and *Y*: it 'partials out' or accounts for the relationship that *X* has with other predictors. Therefore, the computer looks for the variable that has the largest *unique* overlap with the outcome. This variable is retained if it significantly improves the fit of the model, otherwise it is rejected and the process stops. If it is retained and there are still potential predictors left out of the model then these are reviewed and the one with the largest semi-partial correlation with the outcome is entered, evaluated and retained if it significantly improves the fit, and so on until there are no more potential predictors or none of the potential predictors significantly improves the model if it is entered.

Let's make this process a bit more concrete. In Section 8.5, we used an example of the relationships between exam performance, exam anxiety and revision time. Imagine our goal is to predict exam performance from the other two variables. Think back to Figure 8.10. If we build the model stepwise, the first step is to see which of exam anxiety and revision time overlaps more with exam performance. The area of overlap between exam performance and exam anxiety is area A + C (19.4%), whereas for revision time it is area B + C (15.8%). Therefore, exam anxiety will enter the model first and is retained only if it significantly improves the model's predictive power. If not, no predictor variables are entered.

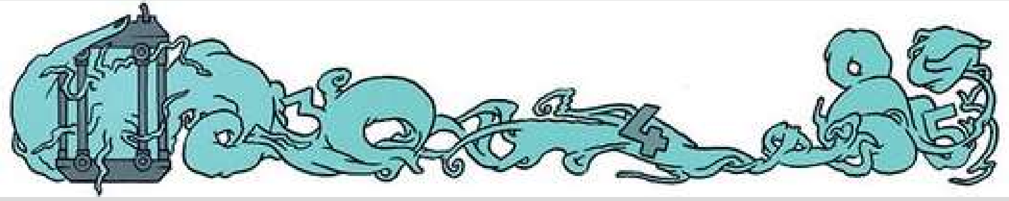
In our exam performance example, there is only one other potential predictor (revision time), so this will be entered next. Remember that its unique overlap with exam performance is area B in Figure 8.10: we ignore the part of its overlap with exam performance that is shared with exam anxiety (area C), because that variable is already in the model. If area B is big enough to improve the fit of the model significantly then revision time will be retained. If not the final model will contain only exam anxiety.

In the case where we had another potential predictor (let's say we measured the difficulty of the exam) and exam anxiety had been entered first, then the unique overlap for revision time and exam performance (area B) would be compared to the equivalent area for exam difficulty. The variable with the bigger area would be entered next, evaluated and retained only if its inclusion improved the fit of the model.

The *stepwise* method in SPSS Statistics 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 constantly reassessed to see whether redundant predictors can be removed. The *backward* method is the opposite of the forward method in that the model initially contains all predictors and the contribution of each is evaluated with the *p*-value of its *t*-test. This significance value is compared against a removal criterion (which can be either an absolute value of the test statistic or a *p*-value). If a predictor meets the removal criterion (i.e., it is not making a statistically significant contribution to the model) it is removed and the model is re-estimated for the remaining predictors. The contribution of the remaining predictors is then reassessed.

Oditi's Lantern *Automatic Linear Modeling*

'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 Modeling*. This feature 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.'



Which of these methods should you use? The short answer is 'not stepwise', because variables are selected based upon mathematical criteria. The issue is that these criteria (e.g., the semi-partial correlation) are at the mercy of sampling variation. That is, a particular variable might have a large semi-partial correlation in your sample but a small one in a different sample. Therefore, models built using stepwise methods are less likely to generalize across samples because the selection of variables in the model is affected by the sampling process. Also, because the criterion for retaining variables is based on statistical significance, your sample size affects the model you get: in large samples significance tests are highly powered, resulting in predictors being retained that make trivial contributions to predicting the outcome, and in small samples where power is low, predictors that make a large contribution may get overlooked. Consequently, there is the danger of overfitting (having too many variables in the model that essentially make little contribution to predicting the outcome) and underfitting (leaving out important predictors) the model. Stepwise methods also take important methodological decisions out of the hands of the researcher.

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) illustrates this problem by imagining getting dressed using a stepwise method. You wake up one morning and 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 keep 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 put your underwear on you but it won’t fit over your jeans. It decides they are ‘a poor fit’ and discards them. It tries a jacket – that fits, but your T-shirt doesn’t go over the top and is discarded. You end up leaving the house in jeans and a jacket with nothing underneath. You are very cold. Later in the day during a university seminar you stand up and your trousers fall down (because your body has shrunk from the cold), exposing you to your year group. It’s a mess. The problem is that the underwear was a poor fit only because when you tried to put it 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 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 9.4.2).

9.9.2 Comparing models

Hierarchical and (although obviously you'd never use them) stepwise methods involve adding predictors to the model in stages, and it is useful to assess the improvement to the model at each stage. Given that larger values of R^2 indicate better fit, a simple way to quantify the improvement when predictors are added is to compare the R^2 for the new model to that for the old model. We can assess the significance of the change in R^2 using equation (9.13), but because we're looking at the change in models we use the change in R^2 (R^2_{change}) and the change in the number of predictors (k_{change}), as well as the R^2 (R^2_{new}) and number of predictors (k_{new}) in the new model:

$$F_{\text{change}} = \frac{\left(N - k_{\text{new}} - 1\right) R^2_{\text{change}}}{k_{\text{change}}\left(1 - R^2_{\text{new}}\right)} \qquad (9.18)$$

We can compare models using this *F*-statistic. The problem with R^2 is that when you add more variables to the model, it always goes 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 that 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 *Automatic Linear Model* function in SPSS Statistics 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 an AIC value of 10 is small, or that a value 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, the fit of your model is improving.

[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.

9.9.3 Multicollinearity

A final consideration for models with more than one predictor 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. Big standard errors for *b* coefficients mean more variability in these *b*s across samples, and a greater chance of (1) predictor equations that are unstable across samples too; and (2) *b* coefficients in the sample that are unrepresentative of those in the population. Crudely put, multicollinearity leads to untrustworthy *b*-values. Don’t lend them money and don’t let them go for dinner with your boy-or girlfriend.
- **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 with $R = 0.80$ and a second predictor variable is 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 (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 (*R* might increase from 0.80 to 0.82). 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. The second predictor might account for only a little of the variance in the outcome, but the variance it does account for is different from that of the other predictor (and so when both predictors are included, *R* is substantially larger, say 0.95).
- **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? We can't – the model could include either one, interchangeably.

A 'ball park' method of identifying multicollinearity (that will miss subtler forms) is to scan the correlation matrix for predictor variables that correlate very highly (values of *r* above 0.80 or 0.90). SPSS Statistics can compute the **variance inflation factor (VIF)**, which indicates whether a predictor has a strong linear relationship with the other predictor(s), and the **tolerance** statistic, which is its reciprocal (1/VIF). Some general guidelines have been suggested for interpreting the VIF:

- If the largest VIF is greater than 10 (or the tolerance is below 0.1) then this indicates a serious problem (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 0.2 indicates a potential problem (Menard, 1995).

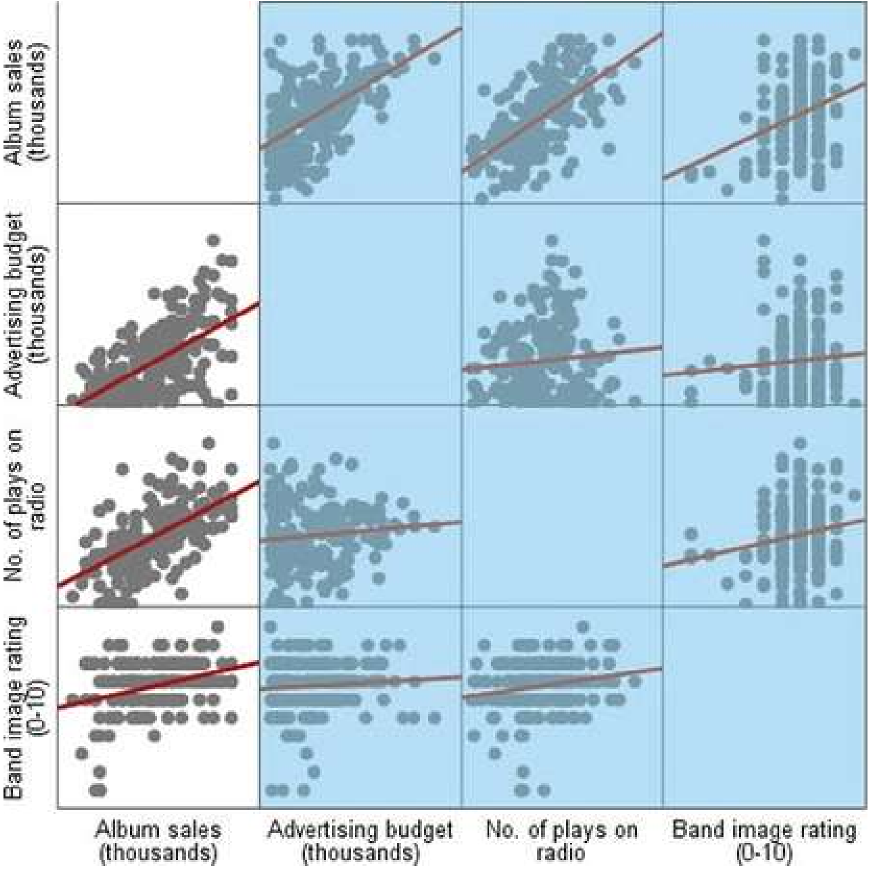
Other measures that are useful in discovering whether predictors are dependent are the *eigenvalues of the scaled, uncentred cross-products matrix,* the *condition indexes* and the *variance proportions*. These statistics will be covered as part of the interpretation of SPSS output (see Section 9.11.5). If none of this made sense, Hutcheson and Sofroniou (1999) explain multicollinearity very clearly.

9.10 Using SPSS Statistics to fit a linear model with

several predictors

Remember the general procedure in Figure 9.10. First, we could look at scatterplots of the relationships between the outcome variable and the predictors. Figure 9.14 shows a matrix of scatterplots for our album sales data, but I have shaded all the scatterplots except the three related to the outcome, album sales. Although the data are messy, the three predictors have reasonably linear relationships with the album sales and there are no obvious outliers (except maybe in the bottom left of the scatterplot with band image).

Figure 9.14 Matrix scatterplot of the relationships between advertising budget, airplay, image rating and album sales



SELF TEST

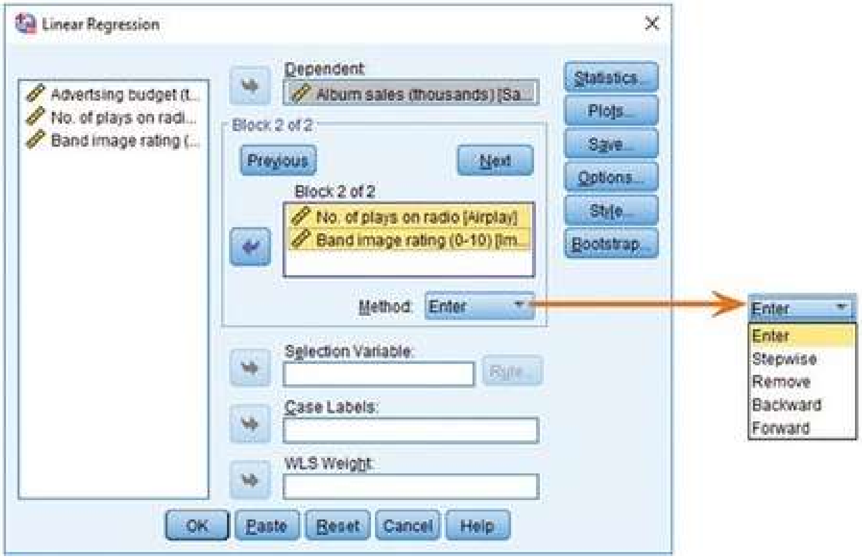
Produce a matrix scatterplot of **Sales, Adverts, Airplay** and **Image** including the regression line.

9.10.1 Main options

Past research shows that advertising budget is a significant predictor of album sales, and so we should include this variable in the model first, entering the new variables (**Airplay** and **Image**) afterwards. This method is hierarchical (we decide the order that variables are entered based on past research). To do a hierarchical regression we enter the predictors in blocks, with each block representing one step in the hierarchy. Access the main *Linear Regression* dialog box by selecting *Analyze* ▸ *Regression* ▸ Linear.... We encountered this dialog box when we looked at a model with only one predictor (Figure 9.13). To set up the first block we do what we did before: drag **Sales** to the box labelled *Dependent* (or click ➡). We also need to specify the predictor variable for the first block, which we decided should be advertising budget. Drag this variable from the left-hand list to the box labelled *Independent(s)* (or click ➡). Underneath the *Independent(s)* box is a drop-down menu for specifying the *Method* of variable entry (see Section 9.9.1). You can select a different method for each block by clicking on Enter ▾. The default option is forced entry, and this is the option we want, but if you were carrying out more exploratory work, you might use a different method.

Having specified the first block, we can specify a second by clicking Next. This process clears the *Independent(s)* box so that you can enter the new predictors (note that it now reads *Block 2 of 2* above this box to indicate 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, so select **Airplay** and **Image** and drag them to the *Independent(s)* box (or click ➡). The dialog box should look like Figure 9.15. To move between blocks use the Previous and Next buttons (e.g., to move back to block 1, click Previous).

Figure 9.15 Main dialog box for block 2 of the multiple regression



Linear Regression

Advertsing budget (t...
No. of plays on radi...
Band image rating (...

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

Block 2 of 2

Block 2 of 2:
No. of plays on radio [Airplay]
Band image rating (0-10) [Im...

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

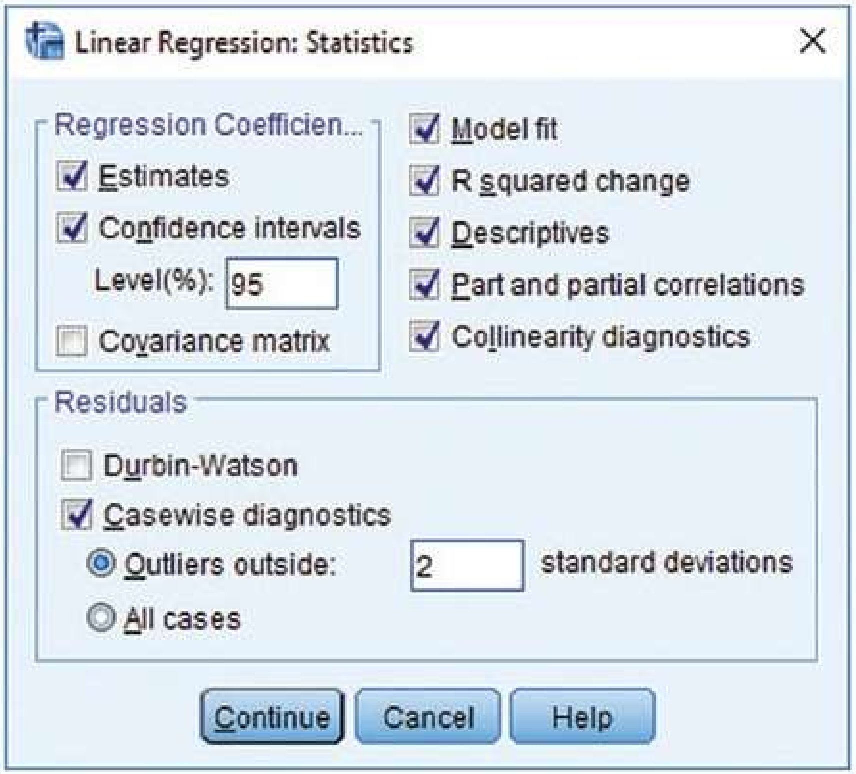
It is possible to select different methods of variable entry for different blocks. For example, having specified forced entry for the first block, we could now specify a stepwise method for the second. Given that we have no previous research regarding the effects of image and airplay on album sales,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

9.10.2 Statistics

In the main *Regression* dialog box click Statistics... to open the dialog box in Figure 9.16. Below is a run-down of the options available. Select the options you require and click Continue to return to the main dialog box.

- *Estimates*: This option is selected by default because it gives us the estimated *b*-values for the model as well as the associated *t*-test and *p*-value (see Section 9.2.5).
- *Confidence intervals*: This option produces confidence intervals for each *b*-value in the model. Remember that if the model assumptions are not met these confidence intervals will be inaccurate and bootstrap confidence intervals should be used instead.
- *Covariance matrix*: This option produces a matrix of the covariances, correlation coefficients and variances between the *b*-values for each variable in the model. A variance–covariance matrix displays variances along the diagonal and covariances as off-diagonal elements. Correlations are produced in a separate matrix.
- *Model fit*: This option produces the *F*-test, *R*, R^2 and the adjusted R^2 (described in Sections 9.2.4. and 9.4.2).
- *R squared change*: This option displays the change in R^2 resulting from the inclusion of a new predictor (or block of predictors) – see Section 9.9.2.
- *Descriptives*: This option displays a table of the mean, standard deviation and number of observations of the variables included in the model. A correlation matrix is produced too, which can be helpful for spotting multicollinearity.

Figure 9.16 *Statistics* dialog box for regression analysis



Regression Coefficien...

☑ 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

- *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 semi-partial (part) and partial correlation between each predictor and the outcome (see Sections 8.5 and 9.9.1).
- *Collinearity diagnostics*: This option produces collinearity statistics such as the VIF, tolerance, eigenvalues of the scaled, uncentred cross-products matrix, condition indexes and variance proportions (see Section 9.9.3).
- *Durbin-Watson*: This option produces the Durbin–Watson test statistic, which tests the assumption of independent errors when cases have some meaningful sequence. In this case it isn't useful because our cases do not have a meaningful order.
- *Casewise diagnostics*: This option produces a table that lists the observed value of the outcome, the predicted value of the outcome, the difference between these values (the residual) and this difference standardized. You can choose to have this information for all cases, but that will result in a big table in large samples. The alternative option is to list only cases for which the standardized residual is greater than 3 (when the ± sign is ignored). I usually change this to 2 (so that I don't miss cases with standardized residuals not quite reaching the threshold of 3) 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 is also produced (see Section 9.10.4).

9.10.3 Regression plots

Once you are back in the main dialog box, click Plots... to activate the dialog box in Figure 9.17, which we can use to test some assumptions of the model. Most of these plots involve various *residual* values, which were described in Section 9.3. The left-hand side lists several variables:

- **DEPENDNT**: the outcome variable.
- ***ZPRED**: the standardized predicted values of the outcome 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 values of the outcome and those predicted by the model.
- ***DRESID**: the deleted residuals described in Section 9.3.2.
- ***ADJPRED**: the adjusted predicted values described in Section 9.3.2.
- ***SRESID**: the Studentized residual described in Section 9.3.1.
- ***SDRESID**: the Studentized deleted residual described in Section 9.3.2.

In Section 6.11.1 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 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 drag a variable from the list to the space labelled either *X* or *Y* (which refer to the axes), or select the variable and click ➡. When you have selected two variables for the first plot (as in Figure 9.17) you can specify a new plot (up to nine different plots) by clicking on Next. This process clears the dialog box and you can specify a second plot. Click Next or Previous to move between plots you have specified.

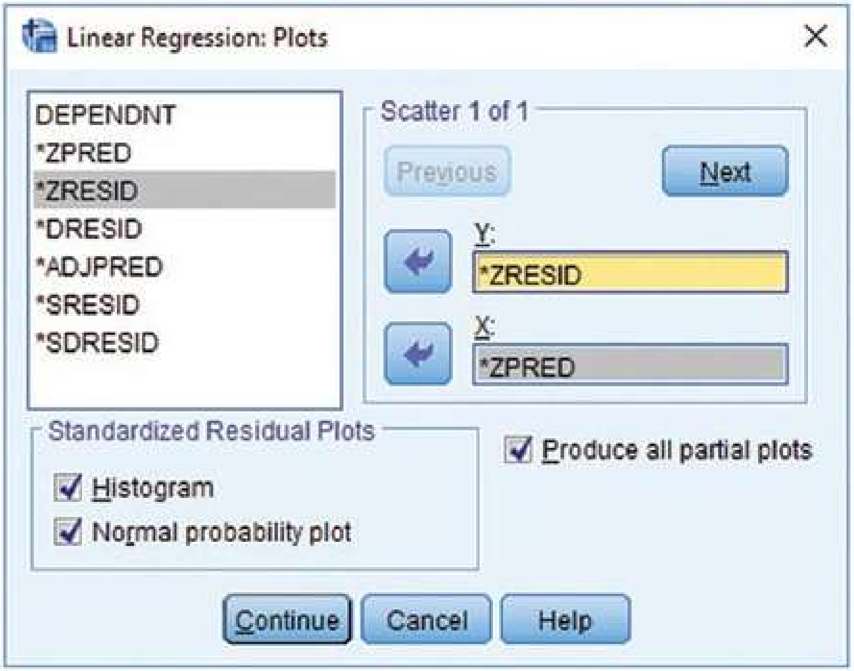
Ticking the box labelled *Produce all partial plots* will produce scatterplots of the residuals of the outcome variable and each of the predictors when both variables are regressed separately on the remaining predictors. Regardless of whether the previous sentence made any sense to you, these plots have 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 coefficient, *b*. Second, non-linear relationships between a predictor and the outcome variable are much more evident on these plots. Finally, they are useful for spotting collinearity.

There are two other tick-boxes labelled *Standardized Residual Plots*. One produces a histogram of the standardized residuals and the other produces a normal probability plot, both of which are useful for checking for normality of errors). Click Continue to return to the main dialog box.

9.10.4 Saving regression diagnostics 🟩🟩⬛⬛

Section 9.3 described numerous variables that we can use to diagnose outliers and influential cases. We can save these diagnostic variables for our model in the data editor (SPSS calculates them and places the values in new columns in the data editor) by clicking Save... to access the dialog box in Figure 9.18. Most of the available options were explained in Section 9.3, and Figure 9.18 shows what I consider to be a reasonable set of diagnostic statistics. Standardized (and Studentized) versions of these diagnostics are generally easier to interpret, and so I tend to select them in preference to the unstandardized versions. Once the model has been estimated, SPSS creates a column in your data editor for each statistic requested; it uses a standard set of variable names to describe each one. After the name, there will be a number that refers to the model from which they were generated. For example, for the first model fitted to a data set the variable names will be followed by a 1, if you estimate a second model it will create a new set of variables with names followed by a 2, and so on. For reference, the names used by SPSS are listed below. Selected the diagnostics you require and click Continue to return to the main dialog box.

Figure 9.17 The *Plots* dialog box



Linear Regression: Plots

DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID

Scatter 1 of 1

Y: *ZRESID

X: *ZPRED

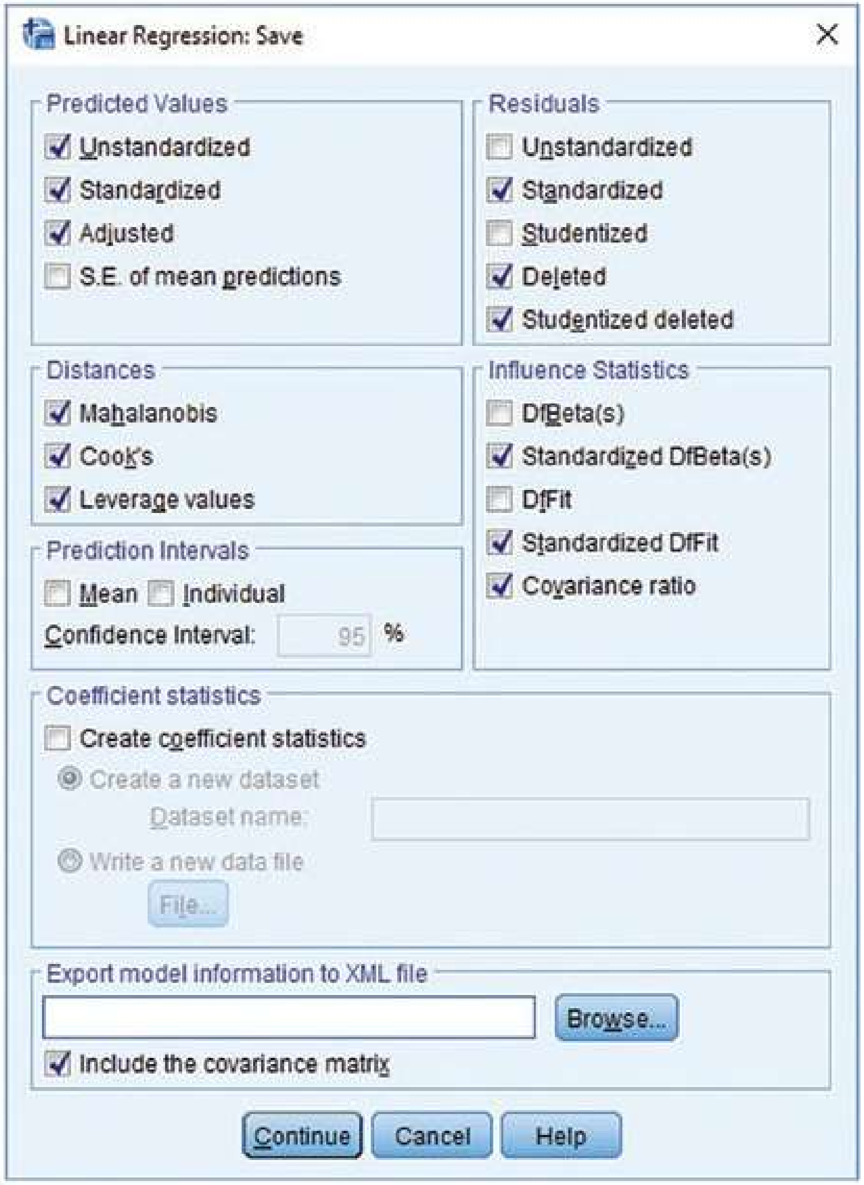
Standardized Residual Plots

☑ Histogram

☑ Normal probability plot

☑ Produce all partial plots

Figure 9.18 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

Export model information to XML file
☑ Include the covariance matrix

- **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.

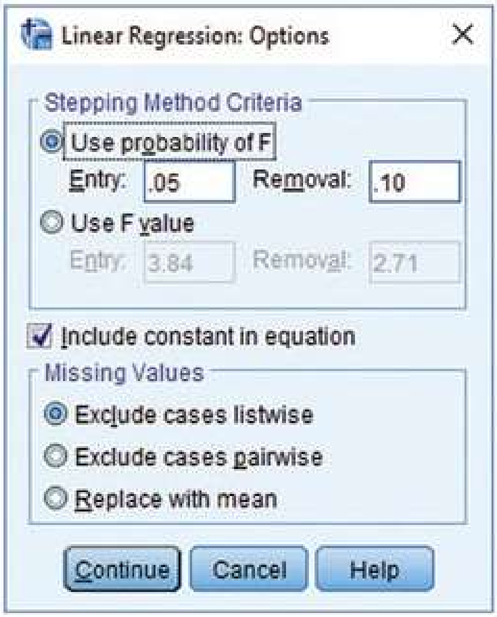
9.10.5 Further options

Clicking Options... activates the dialog box in Figure 9.19. 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 0.05 probability for entry alone. However, you can make this criterion more stringent (0.01). 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 6.1 for a description). Just a hint, but leave the default of listwise alone because using pairwise can lead to absurdities such as R^2 that is negative or greater than 1.0.

9.11 Interpreting a linear model with several predictors

Having selected the relevant options and returned to the main dialog box, click OK and watch in awe as SPSS Statistics spews forth quite terrifying amounts of output in the viewer window.

Figure 9.19 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

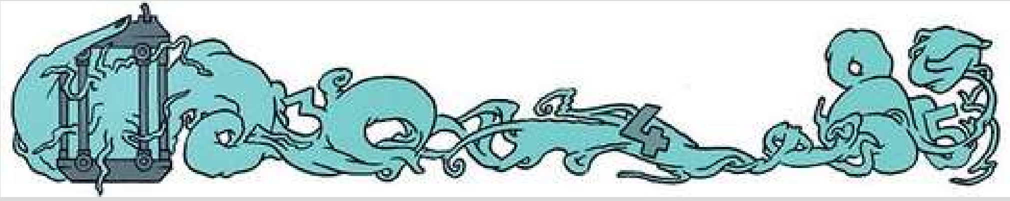
● Exclude cases listwise

○ Exclude cases pairwise

○ Replace with mean

Oditi’s Lantern The linear model

‘I, Oditi, wish to predict when I can take over the world, and rule you pathetic mortals with a 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.’



9.11.1 Descriptives

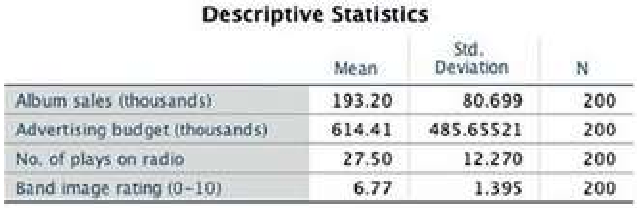
The output described in this section is produced using the options in Figure 9.16. If you selected the *Descriptives* option, you’ll get Output 9.5, which tells us the mean and standard deviation of each variable in our model. This table is a useful summary of the variables. You’ll also see a correlation matrix containing the Pearson correlation coefficient between every pair of variables, the one-tailed significance of each correlation, and the number of cases contributing to each correlation. Along the diagonal of the matrix the values for the correlation coefficients are all 1.00 (a perfect positive correlation) because they are the correlation of each variable with itself.

We can use the correlation matrix to get a sense 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 > 0.9) between predictors. If we look only at the predictors (ignore album sales) then the highest correlation is between the ratings of the band’s image and the amount of airplay, which is significant at the 0.01 level (r = 0.182, p = 0.005). Despite the significance, the coefficient itself is small and so there is no collinearity to worry about. If we look at the outcome variable, then it’s apparent that of the predictors airplay correlates best with the outcome (r = 0.599, p < 0.001).

9.11.2 Summary of the model 🟩🟩⬛⬛

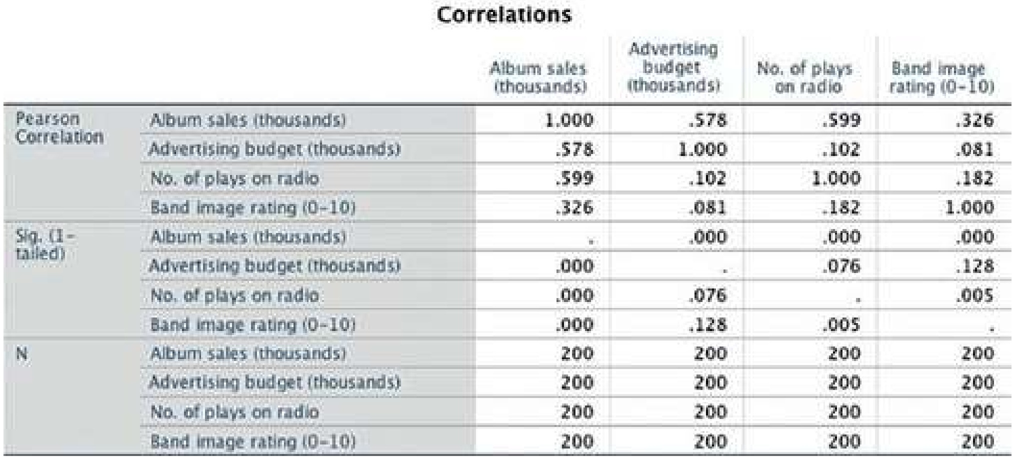
Output 9.6 describes the overall fit of the model. There are two models in the table because we chose a hierarchical method with two blocks and the summary statistics are repeated for each model/block. Model 1 refers to the first stage in the hierarchy when only advertising budget is used as a predictor. Model 2 refers to when all three predictors are used. We can tell this from the footnotes under the table. If you selected the *R squared change* and *Durbin-Watson* options, then these values are included also (we didn’t select Durbin–Watson so it is missing from Output 9.6).

Output 9.5



Descriptive Statistics

	Mean	Std. Deviation	N
Album sales (thousands)	193.20	80.699	200
Advertising budget (thousands)	614.41	485.65521	200
No. of plays on radio	27.50	12.270	200
Band image rating (0-10)	6.77	1.395	200



Correlations

		Album sales (thousands)	Advertising budget (thousands)	No. of plays on radio	Band image rating (0-10)
Pearson Correlation	Album sales (thousands)	1.000	.578	.599	.326
	Advertising budget (thousands)	.578	1.000	.102	.081
	No. of plays on radio	.599	.102	1.000	.182
	Band image rating (0-10)	.326	.081	.182	1.000
Sig. (1-tailed)	Album sales (thousands)	.	.000	.000	.000
	Advertising budget (thousands)	.000	.	.076	.128
	No. of plays on radio	.000	.076	.	.005
	Band image rating (0-10)	.000	.128	.005	.
N	Album sales (thousands)	200	200	200	200
	Advertising budget (thousands)	200	200	200	200
	No. of plays on radio	200	200	200	200
	Band image rating (0-10)	200	200	200	200

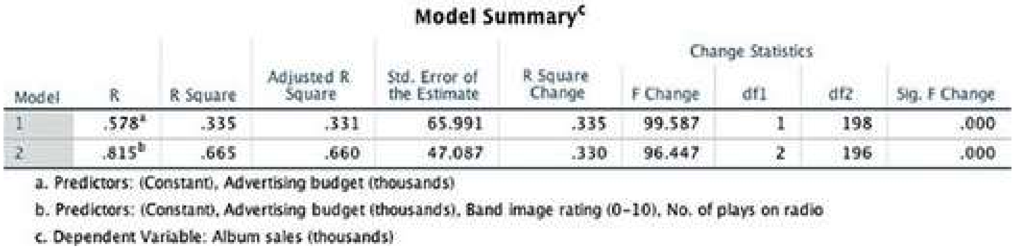
Cramming Sam’s Tips Descriptive statistics

- Use the descriptive statistics to check the correlation matrix for multicollinearity; that is, predictors that correlate too highly with each other, $r > 0.9$.

The column labelled *R* contains 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 (0.578). In fact, all of the statistics for model 1 are the same as the simple regression model earlier (see Section 9.8). 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 0.335, 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 0.665 or 66.5% of the variance in album sales. If advertising accounts for 33.5%, then image and airplay must account for an additional 33%.[13]

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

Output 9.6



Model Summary[c]

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	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

a. Predictors: (Constant), Advertising budget (thousands)

b. Predictors: (Constant), Advertising budget (thousands), Band image rating (0–10), No. of plays on radio

c. Dependent Variable: Album sales (thousands)

The adjusted R^2 gives us some idea of how well our model generalizes, and ideally we'd 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 (it is 0.665 − 0.660 = 0.005 or about 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 (equation (9.15)) you'll get an adjusted value of 0.653 (Jane Superbrain Box 9.2), which is very close to the observed value of R^2 (0.665), indicating that the cross-validity of this model is very good.

The change statistics are provided only if requested and they tell us whether the change in R^2 is significant (i.e., how much does the model fit improve as predictors are added?). The change is reported for each block of the hierarchy: for model 1, R^2 changes from 0 to 0.335 and gives rise to an F-statistic of 99.59, which is significant with a probability less than 0.001. In model 2, in which image and airplay have been added as predictors, R^2 increases by 0.330, making the R^2 of the new model 0.665 with a significant ($p < 0.001$) F-statistic of 96.44

(Jane Superbrain Box 9.2).

Output 9.7 shows the *F*-test of whether the model is significantly better at predicting the outcome than using the mean outcome (i.e., no predictors). The *F*-statistic 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 9.2.4). This table again reports the information for each model separately. The table contains the sum of squares for the model (the value of SS_M from Section 9.2.4), the residual sum of squares (the value of SS_R from Section 9.2.4) and their respective degrees of freedom. For SS_M the *df* are the number of predictors (1 for the first model and 3 for the second). For SS_R the *df* are the number of observations (200) minus the number of coefficients in the regression model. The first model has two coefficients (one each for the predictor and constant) whereas the second has four (the constant plus one for each of the three predictors). Therefore, model 1 has 198 residual degrees of freedom whereas model 2 has 196. Remember, the mean sum of squares (MS) is the SS divided by the *df* and the *F*-statistic is average improvement in prediction by the model (MS_M) divided by the average error in prediction (MS_R). The *p*-value tells us the probability of getting an *F* at least as large as the one we have if the null hypothesis were true (if we used the outcome mean to predict album sales). The *F*-statistic is 99.59, $p < 0.001$ for the initial model and 129.498, $p < 0.001$ for the second. 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.

Jane Superbrain 9.2 Maths frenzy

We can 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 (9.13):[14]

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

In model 2 in Output 9.6, two predictors have been added (image and airplay), so the new model has 3 predictors (k_{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 0.330 (R^2_{change}), making the R^2 of the new model 0.665 (R^2_{new}).[15] The F-statistic for this change comes from equation (9.18):

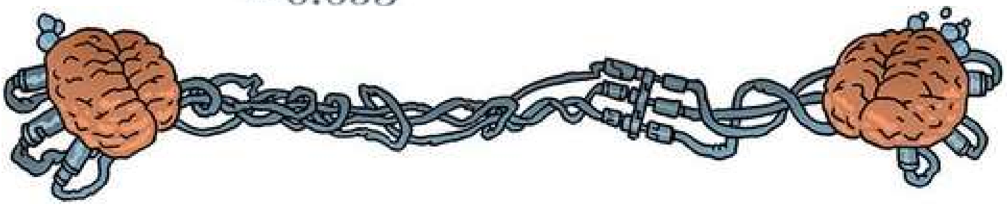
[14] To get the same values as SPSS we have to use the exact value of R^2, which is 0.3346480676231 (if you don’t believe me, double-click in the table in the SPSS output that reports this value, then double-click the cell of the table containing the value of R^2 and you’ll see that what was 0.335 now has a lot more decimal places).

[15] The more precise value is 0.664668.

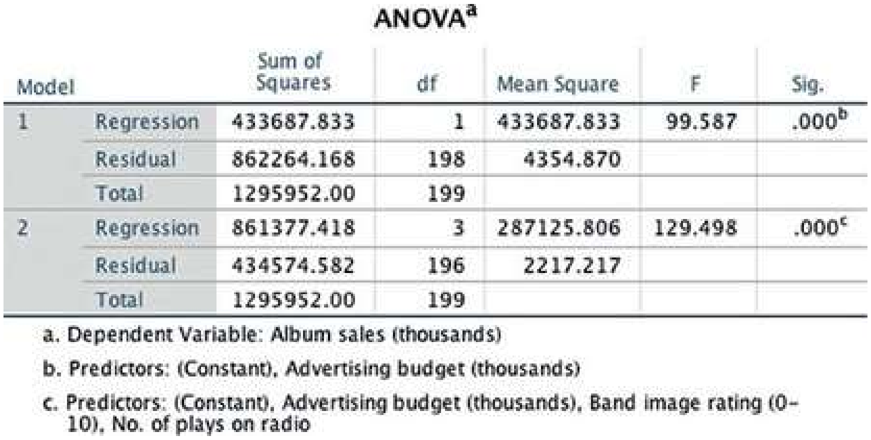
$$F_{\text{change}} = \frac{(N-3-1)0.33}{2(1-0.664668)} = 96.44$$

We can apply Stein’s formula (equation (9.15)) to R^2 to get an idea of its likely value in different samples. We replace n with the sample size (200) and k with the number of predictors (3):

$$\begin{aligned}\text{adjusted } R^2 &= 1-\left[\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\right](1-R^2)\\ &= 1-\left[(1.015)(1.015)(1.005)\right](0.335)\\ &= 1-0.347\\ &= 0.653\end{aligned}$$



Output 9.7



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[b]
	Residual	862264.168	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), Advertising budget (thousands)

c. Predictors: (Constant), Advertising budget (thousands), Band image rating (0–10), No. of plays on radio

Cramming Sam's Tips The model summary

- The fit of the linear model can be assessed using the *Model Summary* and *ANOVA* tables from SPSS.
- R^2 tells you the proportion of variance explained by the model.
- If you have done a hierarchical regression, assess the improvement of the model at each stage by looking at the change in R^2 and whether it is significant (values less than 0.05 in the column labelled *Sig. F Change*).
- The *F*-test tells us whether the model is a significant fit to the data overall (look for values less than 0.05 in the column labelled *Sig.*).

9.11.3 Model parameters

Output 9.8 shows the model parameters for both steps in the hierarchy. The first step in our hierarchy was to include advertising budget, and so the parameters for this first model are identical to those we obtained earlier in this chapter in Output 9.4. Therefore, we will focus on the parameters for the final model (in which all predictors were included). The format of the table of coefficients depends on the options selected in Figure 9.16; for example, the confidence intervals *b*, collinearity diagnostics and the part and partial correlations will be present only if you checked those options.

Earlier in the chapter we saw that a linear model with several predictors takes the form of equation (9.8), which contains several unknown parameters (the *b*-values). The first column in Output 9.8 contains estimates for these *b*-values, which indicate the individual contribution of each predictor to the model. By replacing the *X*s in equation (9.8) with variables names and taking the *b*-values from Output 9.8 we can define our specific model as:

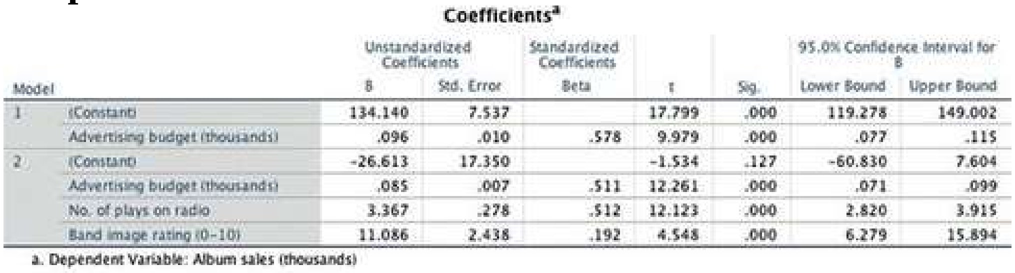
$$\begin{aligned}\text{sales}_i &= b_0 + b_1\text{advertising}_i + b_2\text{airplay}_i + b_3\text{image}_i \\ &= -26.61 + (0.085\ \text{advertising}_i) + (3.367\ \text{airplay}_i) + (11.086\ \text{image}_i)\end{aligned} \tag{9.19}$$

The *b*-values quantify the relationship between album sales and each predictor. The *direction* of the coefficient – positive or negative – corresponds to whether the relationship with the outcome is positive or negative. All three predictors have positive *b*-values, indicating positive relationships. So, as advertising budget, plays on the radio, and image rating increase so do album sales. The *size* of the *b* indicates the degree to which each predictor affects the outcome *if the effects of all other predictors are held constant*:

- **Advertising budget**: *b* = 0.085 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 band image and airplay are held constant.
- **Airplay**: *b* = 3.367 indicates that as the number of plays on radio in the week before release increases by one, album sales increase by 3.367 units. 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 the band’s image and advertising budget are held constant.
- **Image**: *b* = 11.086 indicates that if a band can increase their image rating by 1 unit they can expect additional album sales of 11.086 units. Every unit

increase in the band's image rating is associated with an extra 11.086 thousand albums (11,086 albums) being sold. This interpretation is true only if the effects of airplay and advertising are held constant.

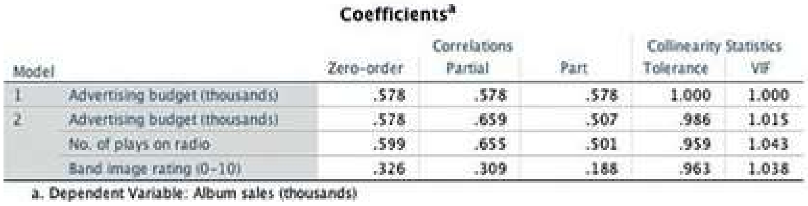
Output 9.8[16]



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	134.140	7.537		17.799	.000	119.278	149.002
	Advertising budget (thousands)	.096	.010	.578	9.979	.000	.077	.115
2	(Constant)	-26.613	17.350		-1.534	.127	-60.830	7.604
	Advertising budget (thousands)	.085	.007	.511	12.261	.000	.071	.099
	No. of plays on radio	3.367	.278	.512	12.123	.000	2.820	3.915
	Band image rating (0-10)	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	Advertising budget (thousands)	.578	.578	.578	1.000	1.000
2	Advertising budget (thousands)	.578	.659	.507	.986	1.015
	No. of plays on radio	.599	.655	.501	.959	1.043
	Band image rating (0-10)	.326	.309	.188	.963	1.038

a. Dependent Variable: Album sales (thousands)

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

Each of the beta values has an associated standard error indicating to what extent these values vary across different samples. The standard errors are also used to compute a *t*-statistic that tests whether the *b*-value is significantly different from 0 (Section 9.2.5). Remember that if a predictor's *b* is zero then its relationship to the outcome is zero also. By testing whether an observed *b* is significantly different from zero, we're testing whether the relationship between the predictor and outcome is different from zero. The *p*-value associated with a *b*'s *t*-statistic (in the column *Sig.*) is the probability of getting a *t* at least as big as the one we have if the population value of *b* was zero (i.e., if there was no relationship between that predictor and the outcome).

For this model, the advertising budget, $t(196) = 12.26$, $p < 0.001$, the amount of radio play prior to release, $t(196) = 12.12$, $p < 0.001$ and band image, $t(196) = 4.55$, $p < 0.001$, are all significant predictors of album sales.[17] Remember that these significance tests are accurate only if the assumptions discussed in Chapter 6 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 band's image had less impact.

[17] For all of these predictors I wrote $t(196)$. The number in brackets is the degrees of freedom. We saw in Section 9.2.5 that the degrees of freedom are $N - k - 1$, where N is the total sample size (in this case 200) and k is the number of predictors (in this case 3). For these data we get 200 – 3 – 1 = 196.

The standardized versions of the *b*-values are sometimes easier to interpret (because they are not dependent on the units of measurement of the variables). The standardized beta values (in the column labelled *Beta*, β_i) tell us the number of standard deviations that the outcome changes when the predictor changes by one standard deviation. Because the standardized beta values are measured in standard deviation units they are directly comparable: the values for airplay and advertising budget are virtually identical (0.512 and 0.511, respectively), suggesting that both variables have a comparably large effect, whereas image (standardized beta of 0.192) has a relatively smaller effect (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 the variables, and these values can be found in Output 9.5.

- **Advertising budget**: Standardized $\beta = 0.511$ 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, 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 the band’s image and airplay are held constant.
- **Airplay**: Standardized $\beta = 0.512$ indicates that as the number of plays on radio in the week before release increases by 1 standard deviation (12.27), album sales increase by 0.512 standard deviations. The standard deviation for album sales is 80,699, so this is a change of 41,320 sales (0.512 × 80,699). Basically, if the station 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 the band’s image and advertising are held constant.
- **Image**: Standardized $\beta = 0.192$ indicates that a band rated one standard deviation (1.40 units) higher on the image scale can expect additional album sales of 0.192 standard deviations units. This is a change of 15,490 sales (0.192 × 80,699). A band with an image rating 1.40 higher than another band can expect 15,490 additional sales. This interpretation is true only if the effects of airplay and advertising are held constant.

SELF TEST

Think back to what the confidence interval of the mean represented (Section 2.8). Can you work out what the confidence intervals for *b* represent?

Output 9.8 also contains the confidence intervals for the *b*s (again these are accurate only if the assumptions discussed in Chapter 6 are met). A bit of revision. Imagine that we collected 100 samples of data measuring the same variables as our current model. For each sample we estimate the same model that we have in this chapter, including confidence intervals for the unstandardized beta values. These boundaries are constructed such that in 95% of samples they contain the population value of *b* (see Section 2.8). Therefore, 95 of our 100 samples will yield confidence intervals for *b* that contain the population value. The trouble is that we don't know if our sample is one of the 95% with confidence intervals containing the population values or one of the 5% that misses.

The typical pragmatic solution to this problem is to assume that your sample is one of the 95% that hits the population value. If you assume this, then you can reasonably interpret the confidence interval as providing information about the population value of *b*. A narrow confidence interval suggests that all samples would yield estimates of *b* that are fairly close to the population value, whereas wide intervals suggest a lot of uncertainty about what the population value of *b* might be. If the interval contains zero then it suggests that the population value of *b* might be zero – in other words, no relationship between that predictor and the outcome – and could be positive but might be negative. All of these statements are reasonable if you're prepared to believe that your sample is one of the 95% for which the intervals contain the population value. Your belief will be wrong 5% of the time, though.

In our model of album sales, 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 the band's image is wider (but still does not cross zero), indicating that the parameter for this variable is less representative, but nevertheless significant.

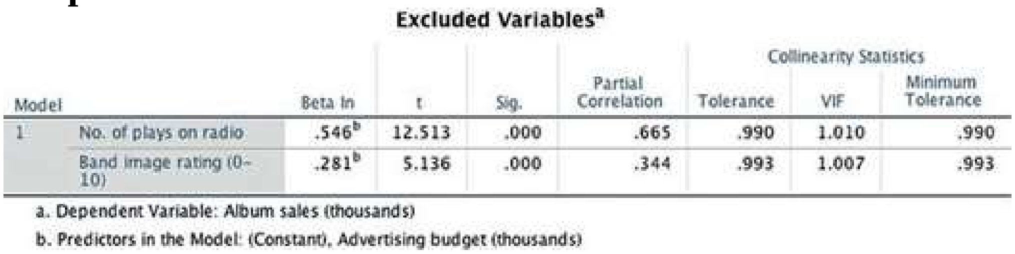
If you asked for part and partial correlations, then they appear in separate columns of the table. The zero-order correlations are the Pearson's correlation coefficients and correspond to the values in Output 9.5. Semi-partial (part) and partial correlations were described in Section 8.5; in effect, the part correlations quantify the unique relationship that each predictor has with the outcome. If you opted 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 that of image rating) and then finally the band's image rating – try running a forward stepwise regression on these data to see if I'm right. Finally, Output 9.8 contains collinearity statistics, but we'll discuss these in Section 9.11.5.

9.11.4 Excluded variables

At each stage of fitting a linear model a summary is provided of predictors that are not yet in the model.

We had a two-block hierarchy with one predictor entered (and two excluded) in block 1, and three predictors entered (and none excluded) in block 2. Output 9.9 details excluded variables only for the first block of our hierarchy, because in the second block no predictors were excluded. The table includes an estimate of the *b*-value and associated *t*-statistic for each predictor *if* it entered the model at this point. Using a stepwise method, the predictor with the highest *t*-statistic will enter the model next, and predictors will continue to be entered until there are none left with *t*-statistics that have significance values less than 0.05. The partial correlation also indicates what contribution (if any) an excluded predictor would make if it entered the model.

Output 9.9



Excluded Variables[a]

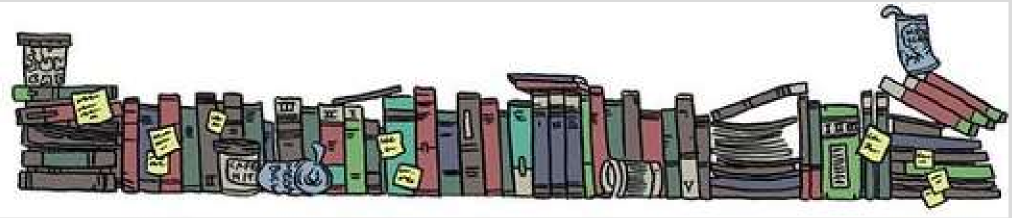
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	No. of plays on radio	.546[b]	12.513	.000	.665	.990	1.010	.990
	Band image rating (0-10)	.281[b]	5.136	.000	.344	.993	1.007	.993

a. Dependent Variable: Album sales (thousands)

b. Predictors in the Model: (Constant), Advertising budget (thousands)

Cramming Sam's Tips Coefficients

- The individual contribution of variables to the regression model can be found in the *Coefficients* table. If you have done a hierarchical regression then look at the values for the final model.
- You can see whether each predictor variable has made a significant contribution to predicting the outcome by looking at the column labelled *Sig.* (values less than 0.05 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, so make a note of them.



9.11.5 Assessing multicollinearity

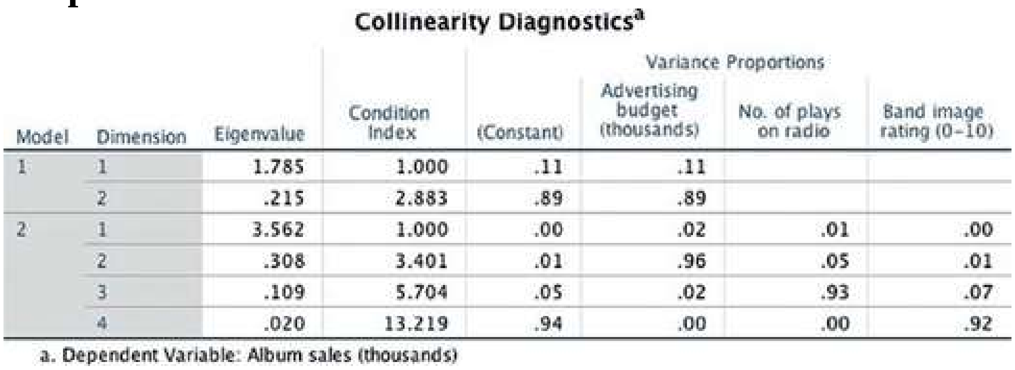
I promised to come back to the measures of collinearity in Output 9.8, so here we go. The output contains the VIF and tolerance statistics (with tolerance being 1 divided by the VIF), and we need to apply the guidelines from Section 9.9.3. The VIF values are well below 10 and the tolerance statistics are well above 0.2. The average VIF, obtained by adding the VIF values for each predictor and dividing by the number of predictors (*k*), is also very close to 1:

$$\overline{\text{VIF}} = \frac{\sum_{i=1}^{k}\text{VIF}_i}{k} = \frac{1.015+1.043+1.038}{3} = 1.032 \tag{9.20}$$

It seems unlikely, therefore, that we need to worry about collinearity among predictors.

The other information we get about collinearity is a table of eigenvalues of the scaled, uncentred cross-products matrix, condition indexes and variance proportions. I discuss collinearity and variance proportions at length in Section 20.8.2, so here I'll just give you the headline: look for large variance proportions on the same *small* eigenvalues (Jane Superbrain Box 9.3). Therefore, in Output 9.10 inspect the bottom few rows of the table (these are the small eigenvalues) and look for variables that *both* have high variance proportions for that eigenvalue. The variance proportions vary between 0 and 1, and you'd like to see each predictor having a high proportion on a different eigenvalue to other predictors (in other words, the large proportions are distributed across different eigenvalues). For our model, each predictor has most of its variance loading onto a different dimension than other predictors (advertising has 96% of variance on dimension 2, airplay has 93% of variance on dimension 3 and image rating has 92% of variance on dimension 4). These data represent no multicollinearity. For an example of when collinearity exists in the data and some suggestions about what can be done, see Chapters 20 (Section 20.8.2) and 18 (Section 18.3.3).

Output 9.10



Collinearity Diagnostics[a]

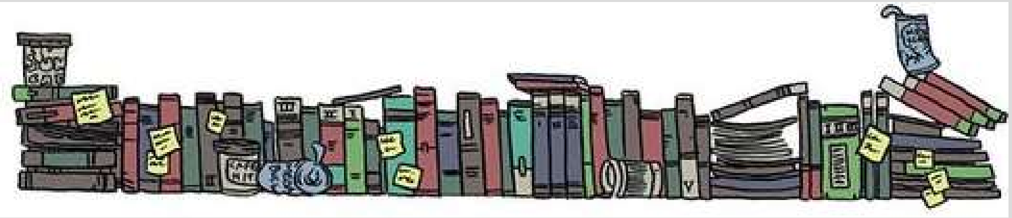
				Variance Proportions			
Model	Dimension	Eigenvalue	Condition Index	(Constant)	Advertising budget (thousands)	No. of plays on radio	Band image rating (0–10)
1	1	1.785	1.000	.11	.11		
	2	.215	2.883	.89	.89		
2	1	3.562	1.000	.00	.02	.01	.00
	2	.308	3.401	.01	.96	.05	.01
	3	.109	5.704	.05	.02	.93	.07
	4	.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*.
- If these values are less than 10 then that indicates 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.



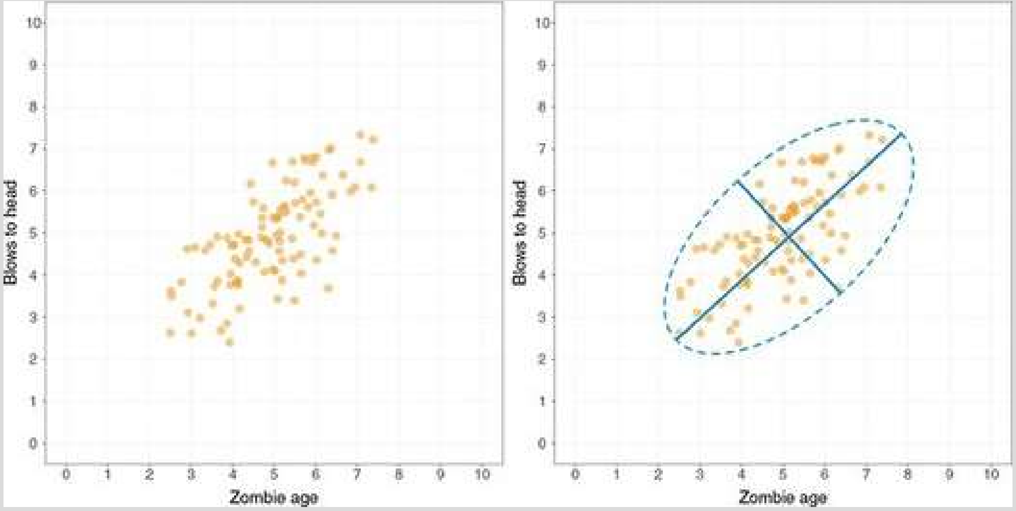
Jane Superbrain 9.3 What are eigenvectors and eigenvalues?

The definitions and mathematics of eigenvalues and eigenvectors are complicated and most of us need not worry about them (although they do crop up again in Chapters 17 and 18). Although the mathematics is hard, we can get a sense of what they represent visually. Imagine we have two variables: the age of a zombie (how long it has been a zombie), and how many blows to the head it takes to kill it.[18] These two variables are normally distributed and can be considered together as a bivariate normal distribution. If these variables are correlated their scatterplot forms an ellipse: if we draw a dashed line around the outer values of the scatterplot we get an oval shape (Figure 9.20). Imagine two lines to measure the length and height of this ellipse: these represent the *eigenvectors* of the correlation matrix for these two variables (a vector is a set of numbers that tells us the location of a line in geometric space). Note that the two straight lines in Figure 9.20 are at 90 degrees to each other, which means that they are independent of one another. So, with two variables, think of eigenvectors as 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., force of blow) our scatterplot gets a third dimension (depth), the ellipse turns into something shaped like a rugby ball (or American football), and we get an extra eigenvector to measure the extra dimension. If we add a fourth variable, a similar logic applies (although it’s harder to visualize).

Each eigenvector has an *eigenvalue* that tells us its length (i.e., the distance from one end of the eigenvector to the other). By looking at the eigenvalues for a data set, we know the dimensions of the ellipse (length and height) or rugby ball (length, height, depth); more generally, we know the dimensions of the data. Therefore, the eigenvalues quantify how evenly (or otherwise) the variances of the matrix are distributed.

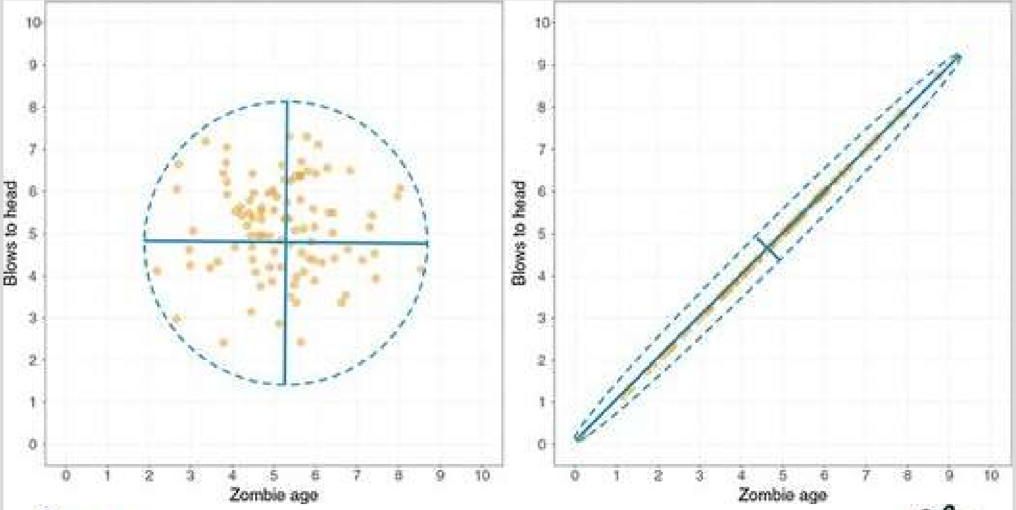
[18] Assuming you can ever kill a zombie, that is.

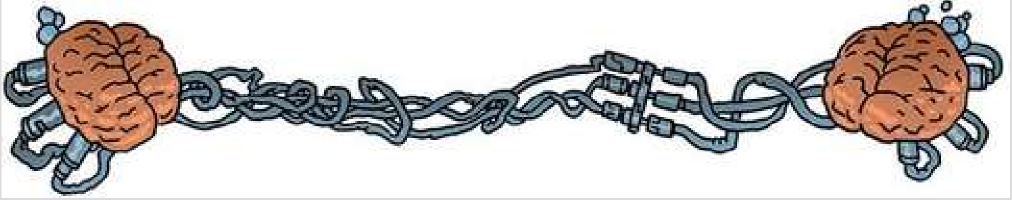
Figure 9.20 A scatterplot of two correlated variables forms an ellipse



In the case of two variables, the *condition* of the data is related to the ratio of the larger eigenvalue to the smaller. Figure 9.21 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 data cloud 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 they'll be the same length, which means they'll have the same eigenvalues. Consequently, when we divide the largest eigenvalue by the smallest we'll get a value of 1. When the variables are perfectly correlated (i.e., there is perfect collinearity) the data cloud (and ellipse around it) will collapse to a straight line. The height of the ellipse will be very small indeed (it will approach zero). Therefore, the largest eigenvalue divided by the smallest will tend to infinity (because the smallest eigenvalue is close to zero). An infinite condition index is a sign of deep trouble.

Figure 9.21 Perfectly uncorrelated (left) and correlated (right) variables

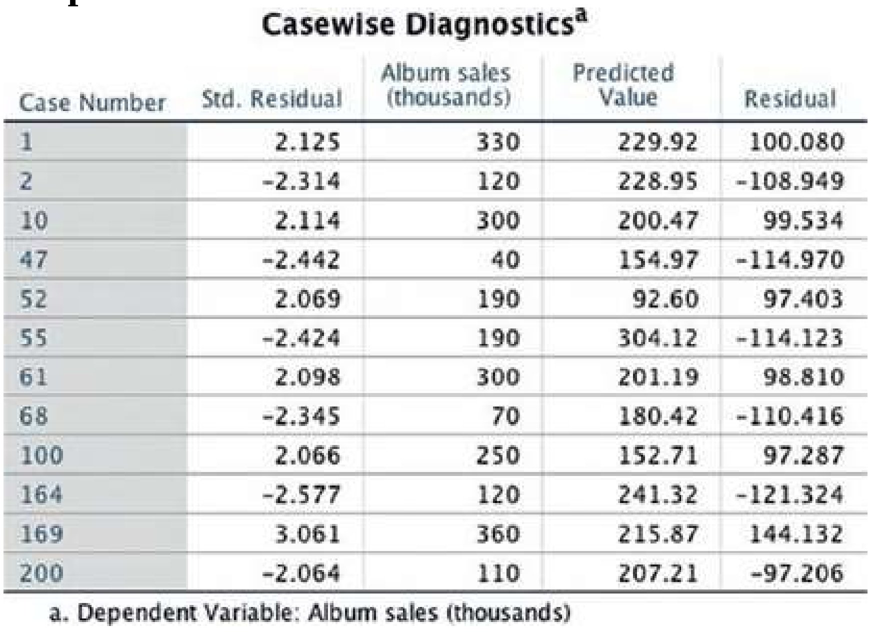




9.11.6 Bias in the model: casewise diagnostics

The final stage of the general procedure outlined in Figure 9.10 is to check the residuals for evidence of bias. The first step is to examine the casewise diagnostics. Output 9.11 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 9.16). In an ordinary sample we would expect 95% of cases to have standardized residuals within about ±2 (Jane Superbrain Box 6.4). We have a sample of 200, therefore it is reasonable to expect about 10 cases (5%) to have standardized residuals outside these limits. Output 9.11 shows that we have 12 cases (6%) that are outside of the limits: pretty close to what we would expect. In addition, 99% of cases should lie within ±2.5 and only 1% of cases should lie outside these limits. We have two cases that lie outside of the limits (cases 164 and 169), which is 1% and what we would expect. These diagnostics give us no cause for concern, except that case 169 has a standardized residual greater than 3, which is probably large enough for us to investigate this case further.

Output 9.11



Casewise Diagnostics[a]

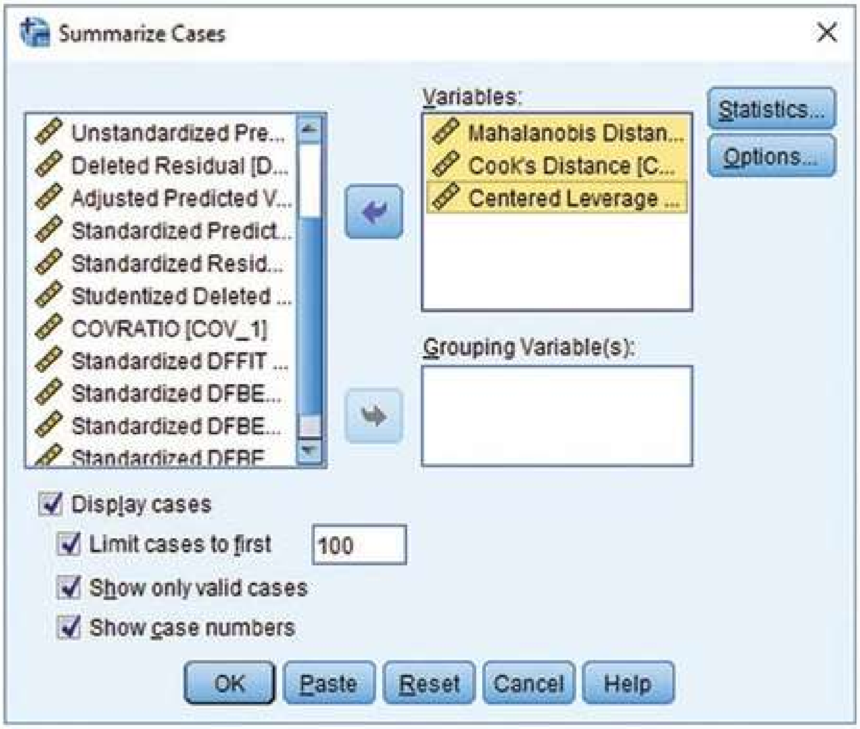
Case Number	Std. Residual	Album sales (thousands)	Predicted Value	Residual
1	2.125	330	229.92	100.080
2	-2.314	120	228.95	-108.949
10	2.114	300	200.47	99.534
47	-2.442	40	154.97	-114.970
52	2.069	190	92.60	97.403
55	-2.424	190	304.12	-114.123
61	2.098	300	201.19	98.810
68	-2.345	70	180.42	-110.416
100	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)

In Section 9.10.4 we opted to save various diagnostic statistics. You should find that the data editor contains columns for these variables. You can check these values in the data editor, or list values in the viewer window. To create a table of values in the viewer, select *Analyze* ▸ *Reports* ▸ Case Summaries... to access the dialog box in Figure 9.22. Select and drag the variables that you want to list into the box labelled *Variables* (or click ➡). By default, the output is limited to the first 100 cases, but if you want to list all cases deselect this option (also see SPSS Tip 9.1). It is also useful to select *Show case numbers* to enable you to identify the case numbers of any problematic cases.

To save space, Output 9.12 shows the influence statistics for 12 cases that I selected. None of them has a Cook's distance greater than 1 (even case 169 is well below this criterion), and so no case appears to have an undue influence on the model. The average leverage can be calculated as $(k + 1)/n = 4/200 = 0.02$, and we should look for values either twice (0.04) or three times (0.06) this value (see Section 9.3.2). All cases are within the boundary of three times the average, and only case 1 is close to two times the average. For the Mahalanobis distances we saw earlier in the chapter that with a sample of 100 and three predictors, values greater than 15 are problematic. Also, with 3 predictors values greater than 7.81 are significant ($p < 0.05$). None of our cases comes close to exceeding the criterion of 15 although case 1 would be deemed 'significant'.

Figure 9.22 The *Summarize Cases* dialog box



Variables:
- Mahalanobis Distan...
- Cook's Distance [C...
- Centered Leverage ...

Grouping Variable(s):

☑ Display cases

☑ Limit cases to first 100

☑ Show only valid cases

☑ Show case numbers

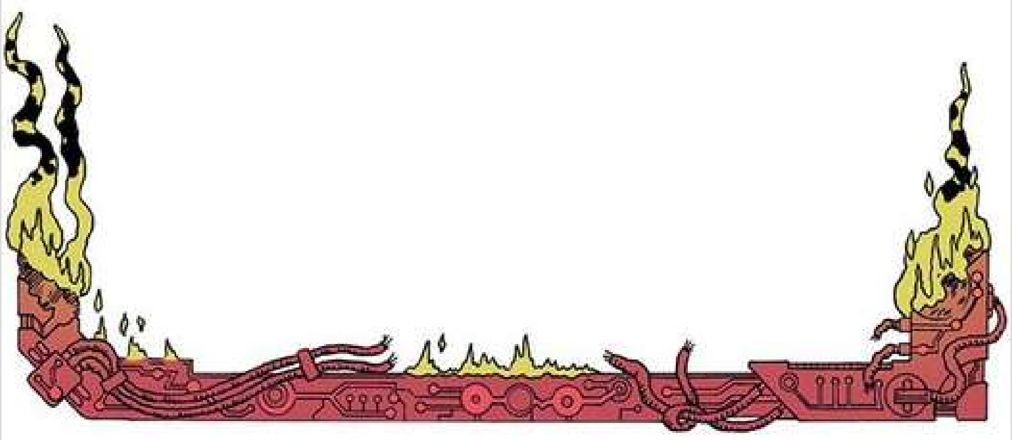
SPSS Tip 9.1 Selecting cases 🟩🟩🟩⬛

In large data sets, a useful strategy when summarizing cases is to use the *select cases* function (see Section 6.12.2) and set conditions that 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.
```

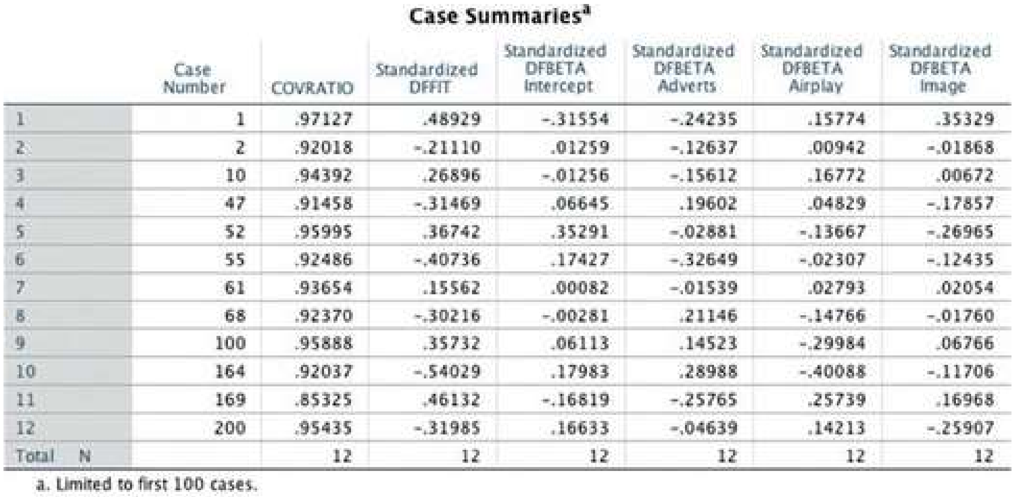
```
      EXECUTE.
```

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* 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).



The DFBeta statistics tell us how much influence each case has on the model parameters. An absolute value greater than 1 is a problem, and all cases in Output 9.12 have values within ±1, which is good news.

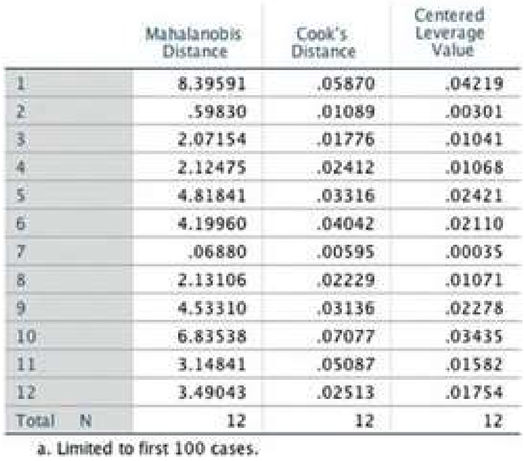
Output 9.12



Case Summaries[a]

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

a. Limited to first 100 cases.



		Mahalanobis Distance	Cook's Distance	Centered Leverage Value
1		8.39591	.05870	.04219
2		.59830	.01089	.00301
3		2.07154	.01776	.01041
4		2.12475	.02412	.01068
5		4.81841	.03316	.02421
6		4.19960	.04042	.02110
7		.06880	.00595	.00035
8		2.13106	.02229	.01071
9		4.53310	.03136	.02278
10		6.83538	.07077	.03435
11		3.14841	.05087	.01582
12		3.49043	.02513	.01754
Total	N	12	12	12

a. Limited to first 100 cases.

For the covariance ratio we need to use the following criteria (Section 9.3.2):

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

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 will have requested other diagnostic statistics and you can apply what we learnt earlier in the chapter when glancing over them.

From this minimal set of diagnostics there's nothing to suggest that there are influential cases (although we'd need to look at all 200 cases to confirm this conclusion); we appear to have a fairly reliable model that has not been unduly influenced by any subset of cases.

Cramming Sam’s Tips Residuals

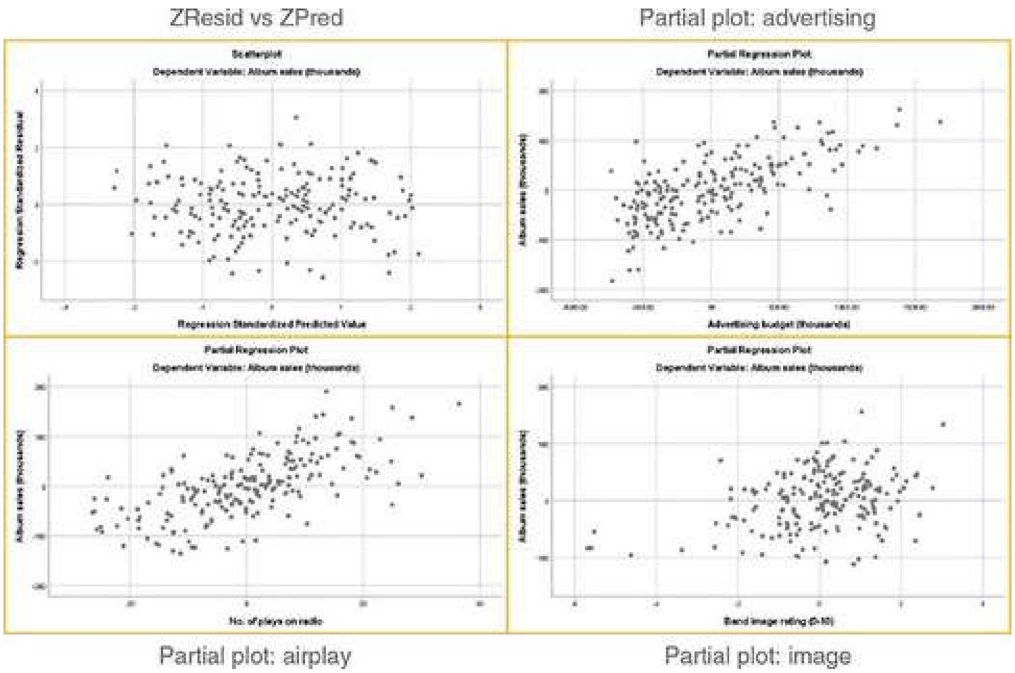
- Look for cases that might be influencing the model.
- Look at standardized residuals and check that no more than 5% of cases have absolute values above 2, and that no more than about 1% have absolute values above 2.5. Any case with a value above about 3 could be an outlier.
- Look in the data editor for the values of Cook’s distance: any value above 1 indicates a case that might be influencing the model.
- Calculate the average leverage and 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 above15 in smaller samples (100). However, Barnett and Lewis (1978) should be consulted for more refined guidelines.
- Look for absolute values of DFBeta greater than 1.
- Calculate the upper and lower limit of acceptable values for the covariance ratio, CVR. Cases that have a CVR that fall outside these limits may be problematic.

9.11.7 Bias in the model: assumptions

The general procedure outlined in Figure 9.10 suggests that having fitted a model, we need to look for evidence of bias, and the second stage of this process is to check the assumptions described in Chapter 6. We saw in Section 6.11.1 that we can look for heteroscedasticity and non-linearity using a plot of standardized predicted values against standardized residuals. We asked for this plot in Section 9.10.3. If everything is OK then this graph should look like a random array of dots. Figure 9.23 (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 it with the examples in Figure 6.19.

Figure 9.23 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 *b* coefficient. Non-linear relationships and heteroscedasticity can be detected using these plots as well. For advertising budget (Figure 9.23, top right) the partial plot shows the strong positive relationship to album sales. There are no obvious outliers and the cloud of dots is evenly spaced out around the line, indicating homoscedasticity. The plot for airplay (Figure 9.23, bottom left) also shows a strong positive relationship to album sales, there are no obvious outliers, and the cloud of dots is evenly spaced around the line, again indicating homoscedasticity. For image (Figure 9.23, bottom right) the plot again shows a positive relationship to album sales, but the dots show funnelling, indicating greater spread for bands with a high image rating. There are no obvious outliers on this plot, but the funnel-shaped cloud indicates a violation of the assumption of homoscedasticity.

Figure 9.23 Plot of standardized predicted values against standardized residuals (top left), and partial plots of album sales against advertising (top right), airplay (bottom left) and image of the band (bottom right)


ZResid vs ZPred

Partial plot: advertising

Partial plot: airplay

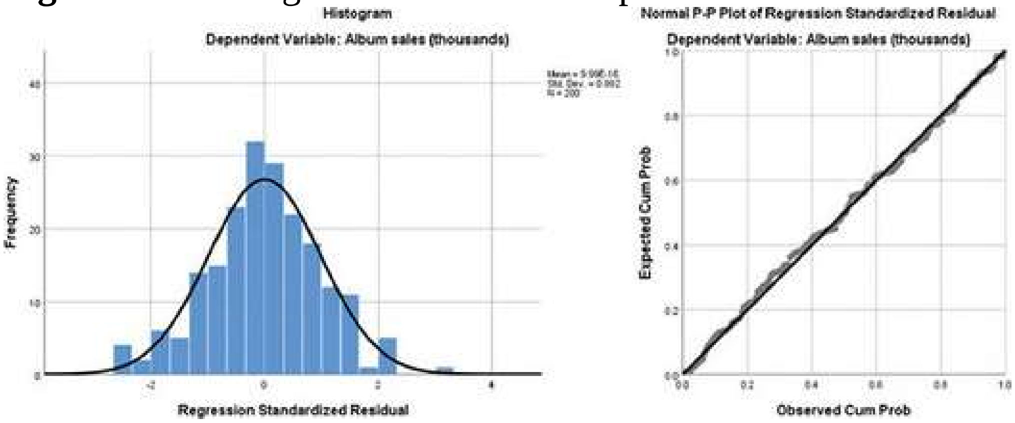
Partial plot: image

To test the normality of residuals, we look at the histogram and normal probability plot selected in Figure 9.17 and shown in Figure 9.24. Compare these plots to examples of non-normality in Section 6.10.1. For the album sales data, the distribution is very normal: the histogram is symmetrical and approximately bell-shaped. In the P-P plot the dots lie almost exactly along the diagonal, which we know indicates a normal distribution (see Section 6.10.1); hence this plot also suggests that the residuals are normally distributed.

9.12 Robust regression

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 image ratings violated the assumption of homoscedasticity. Therefore, we could conclude that in our sample advertising budget and airplay are equally important in predicting album sales. The image of the band is a significant predictor of album sales, but is less important than the other 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. You won’t always (ever?) have such nice data: there will be times when you uncover problems that cast a dark shadow of evil over your model. It will invalidate significance tests, confidence intervals and generalization of the model (use Chapter 6 to remind yourself of the implications of violating model assumptions).

Figure 9.24 Histogram and normal P-P plot for the residuals from our model



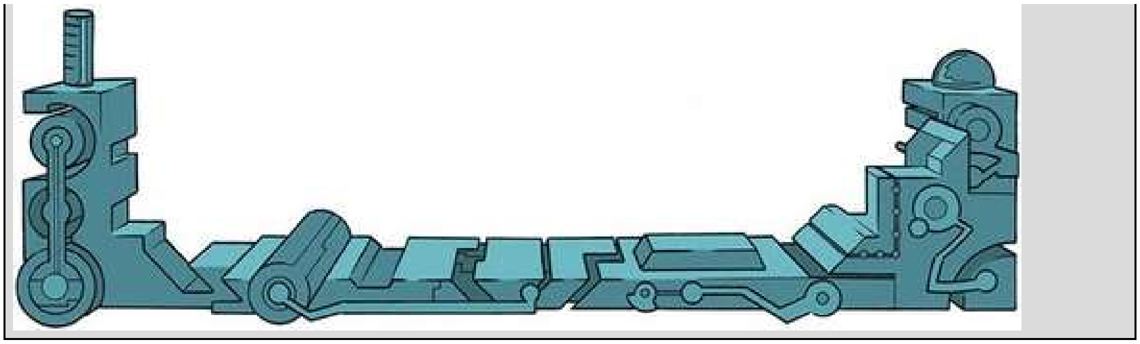
Cramming Sam’s Tips Model assumptions

- 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 get more or less spread out over the graph (look like a funnel) then the assumption of homogeneity of variance is probably unrealistic. If the dots have a pattern to them (i.e., a curved shape) then the assumption of linearity is probably not true. If the dots seem to have a pattern and are more spread out at some points on the plot than others then this could reflect 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 the histogram and P-P plot. If the histogram looks like a normal distribution (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.

Labcoat Leni’s Real Research 9.1 I want to be loved (on Facebook)

Ong, E. Y. L., *et al.* (2011). *Personality and Individual Differences*, *50*(2), 180–185.

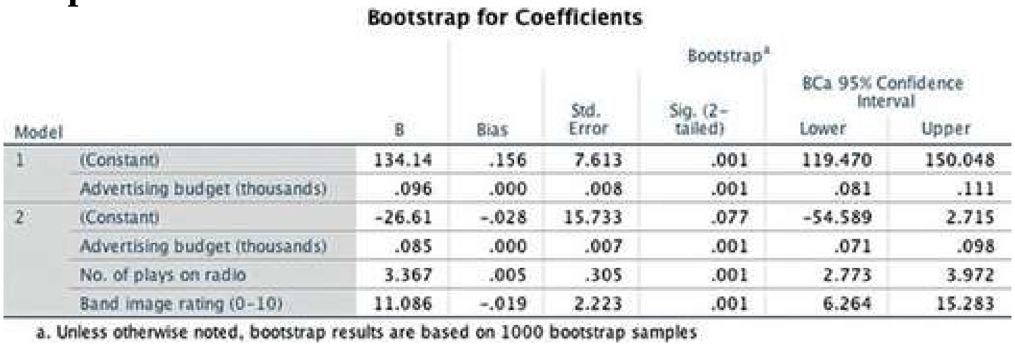
Social media websites such as Facebook offer an unusual opportunity to carefully manage your self-presentation to others (i.e., you can appear rad when in fact you write statistics books, appear attractive when you have huge pustules all over your face, fashionable when you wear 1980s heavy metal band T-shirts, and so on). Ong *et al.* (2011) 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**). Ong *et al.* hypothesized that narcissism would predict 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. Using **Ong *et al.* (2011).sav**, Labcoat Leni wants you to replicate the 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).



Luckily, a lot of the problems can be overcome. If confidence intervals and significance tests of the model parameters are in doubt then use bootstrapping to generate confidence intervals and *p*-values. If homogeneity of variance is the issue then estimate the model with standard errors designed for heteroscedastic residuals (Hayes & Cai, 2007) – you can do this using the PROCESS tool described in Chapter 11. Finally, if the model parameters themselves are in doubt, estimate them using robust regression.

To get robust confidence intervals and significance tests of the model parameters re-estimate your model, selecting the same options as before but clicking Bootstrap... in the main dialog box (Figure 9.13) to access the dialog box explained in Section 6.12.3. 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 (BCa) confidence interval. Bootstrapping won't work if you have set options to save diagnostics, so click Save... to open the dialog box in Figure 9.18 and *deselect everything*. Back in the main dialog box click OK to estimate the model.

Output 9.13



Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	134.14	.156	7.613	.001	119.470	150.048
	Advertising budget (thousands)	.096	.000	.008	.001	.081	.111
2	(Constant)	-26.61	-.028	15.733	.077	-54.589	2.715
	Advertising budget (thousands)	.085	.000	.007	.001	.071	.098
	No. of plays on radio	3.367	.005	.305	.001	2.773	3.972
	Band image rating (0-10)	11.086	-.019	2.223	.001	6.264	15.283

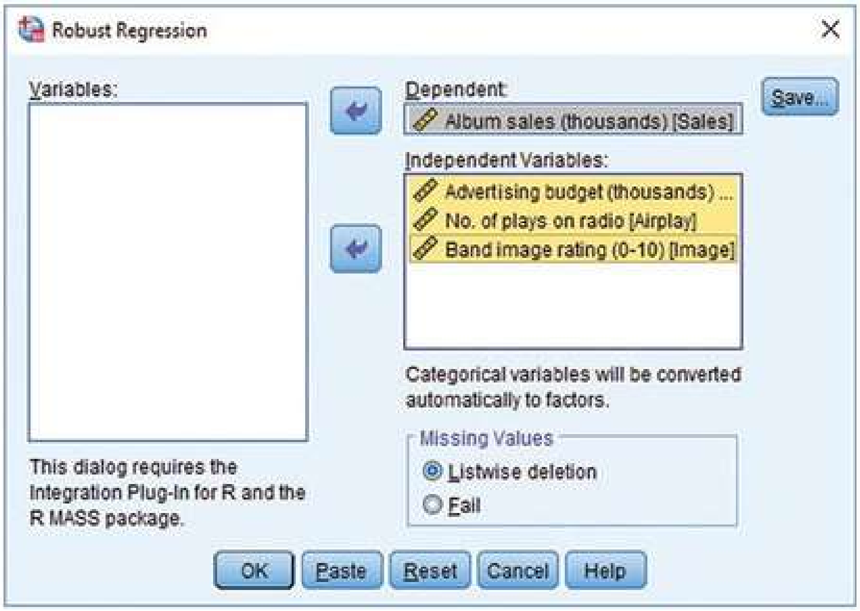
a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

The output will contain a table of bootstrap confidence intervals for each predictor and their significance value (Output 9.13).[19] These tell us that advertising, $b = 0.09$ [0.07, 0.10], $p = 0.001$, airplay, $b = 3.37$ [2.77, 3.97], $p = 0.001$, and the band's image, $b = 11.09$ [6.26, 15.28], $p = 0.001$, all significantly predict album sales. These bootstrap confidence intervals and significance values do not rely on assumptions of normality or homoscedasticity, so they give us an accurate estimate of the population value of *b* for each predictor (assuming our sample is one of the 95% with confidence intervals that contain the population value).

To estimate the *b*s themselves using a robust method we can use the R plugin. If you have installed this plugin (Section 4.13.2) then you can access a dialog box (Figure 9.25) to run robust regression using R by selecting *Analyze* ▸ *Regression* ▸ Robust Regression. If you haven't installed the plugin then this menu won't be there! Drag the outcome (album sales) to the box labelled *Dependent* and any predictors in the final model (in this case advertising budget, airplay and image rating) to the box labelled *Independent Variables*. Click OK to estimate the model.

Figure 9.25 Dialog box for robust regression

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



Dependent: Album sales (thousands) [Sales]

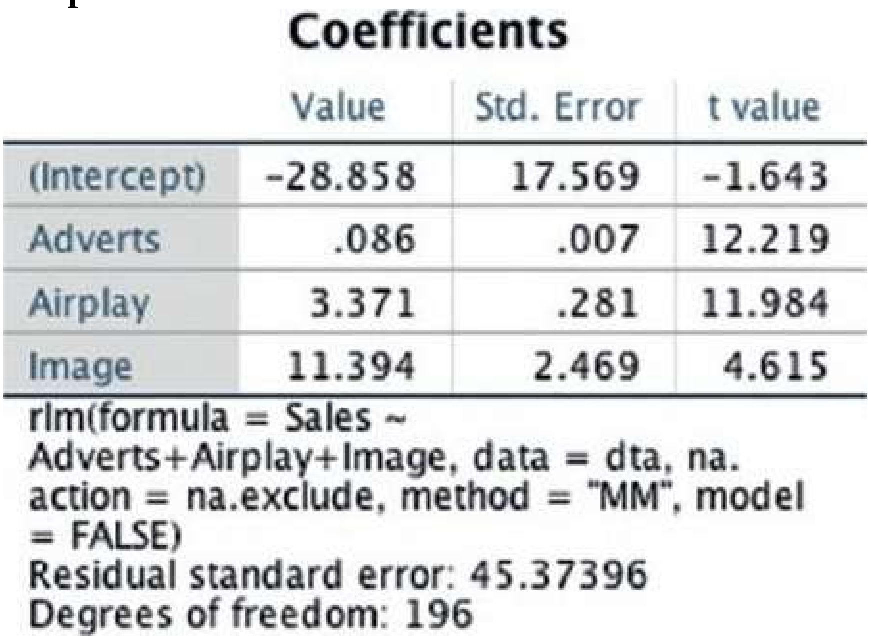
Independent Variables:
Advertising budget (thousands) ...
No. of plays on radio [Airplay]
Band image rating (0-10) [Image]

Categorical variables will be converted automatically to factors.

Missing Values
● Listwise deletion
○ Fail

This dialog requires the Integration Plug-In for R and the R MASS package.

Output 9.14



Coefficients

	Value	Std. Error	t value
(Intercept)	-28.858	17.569	-1.643
Adverts	.086	.007	12.219
Airplay	3.371	.281	11.984
Image	11.394	2.469	4.615

```
rlm(formula = Sales ~
Adverts+Airplay+Image, data = dta, na.
action = na.exclude, method = "MM", model
= FALSE)
Residual standard error: 45.37396
Degrees of freedom: 196
```

Output 9.14 shows the resulting robust *b*-values, their robust standard errors and *t*-statistics. Compare these with the non-robust versions in Output 9.8. The values are not much different (mainly because our original model didn’t seem to violate its assumptions); for example, the *b* for image rating has changed from 11.09 (Output 9.8.) to 11.39 (Output 9.14), the associated standard error was 2.44 and the robust version is 2.47, and the associated *t*-statistic has changed from 4.55 to 4.62. Essentially our interpretation of the model won’t have changed, but this is still a useful sensitivity analysis in that if robust estimates are giving us basically the same results as non-robust estimates then we know that the non-robust estimates have not been unduly biased by properties of the data. So, this is always a useful double check, and if the robust estimates are hugely different from the original estimates then you can use and report the robust versions.

9.13 Bayesian regression 🟩🟩🟩⬜

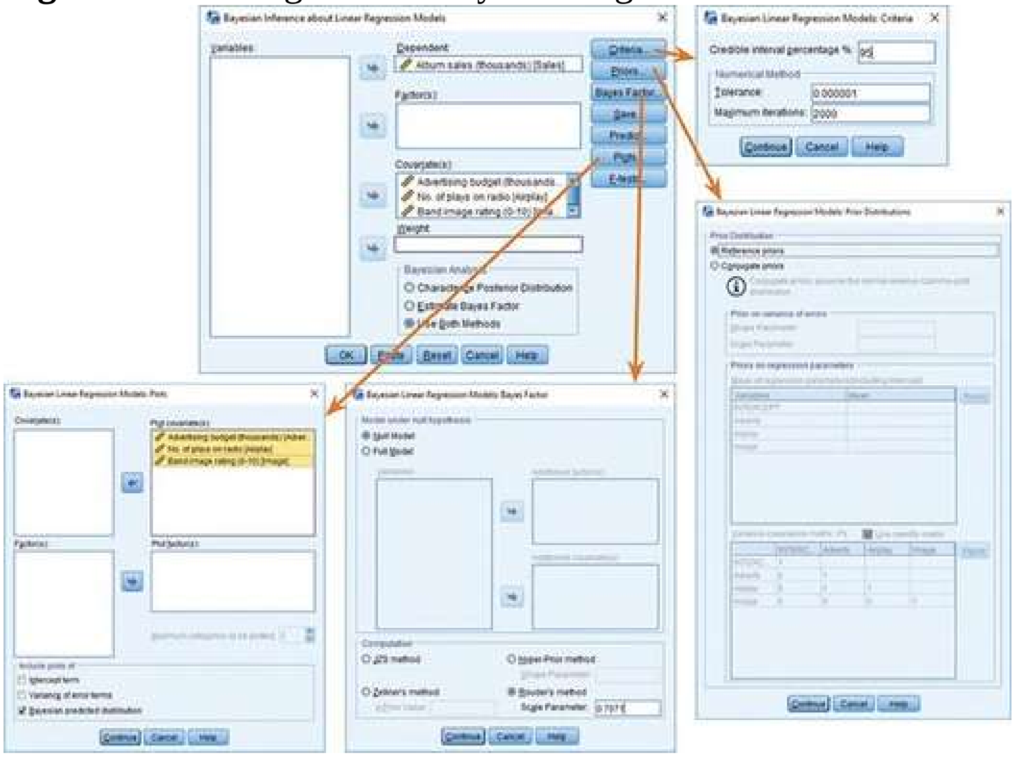
In Section 3.8.4 we looked at Bayesian approaches. To access a dialog box (Figure 9.26) to fit a Bayesian linear model select *Analyze* ▸ *Bayesian Statistics* ▸ *Linear Regression*. You can fit the model either using default priors (so called ◉ Reference priors), which set distributions that represent very diffuse prior beliefs, or conjugate priors, which allow you to specify more specific priors. One of the key strengths of Bayesian statistics (in my opinion) is that you can set evidence-based priors that you update with the data that you collect. However, this is not a trivial undertaking, and it requires a deeper understanding of the models being fit than we have covered. So, to help you to dip your toe in the water of Bayesian statistics we will stick to using the reference priors built into SPSS Statistics. The benefit of reference priors is that they enable you to get going with Bayesian models without drowning in a lot of quite technical material, but the cost is that you are building only very uninformative prior information into your models.[20]

[20] Another downside of this convenience is that I find it hard to know what these priors actually represent (especially in the case of regression).

In the main dialog box (Figure 9.26) drag **Sales** to the box labelled *Dependent* (or click ➡) and drag **Adverts, Airplay** and **Image** to the

Covariate(s) box (or click ➡). If your model has categorical predictors (which we'll look at in the following chapter) drag them to the *Factor(s)* box. If you want to both compute Bayes factors and estimate the model parameters then select ◉ Use Both Methods.

Figure 9.26 Dialog box for Bayesian regression



Dependent: Album sales (thousands) [Sales]

Covariate(s): Advertising budget (thousands... ; No. of plays on radio [Airplay] ; Band image rating (0-10) [Ima...

Bayesian Analysis: ○ Characterize Posterior Distribution ○ Estimate Bayes Factor ● Use Both Methods

Credible interval percentage %: 95

Tolerance: 0.000001

Maximum iterations: 2000

● Reference priors

○ Conjugate priors

Plot covariate(s): Advertising budget (thousands) [Adver... ; No. of plays on radio [Airplay] ; Band image rating (0-10) [Image]

□ Intercept term

□ Variance of error terms

☑ Bayesian predicted distribution

● Null Model

○ Full Model

○ JZS method

○ Zellner's method

○ Hyper-Prior method

● Rouder's method

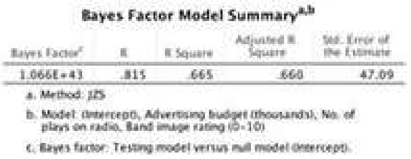
Scale Parameter: 0.7071

If you want a credible interval other than 95% then click Criteria... and change the 95 to the value you want. Click Priors... to set your priors, although we'll stick with ◉ Reference priors. Click Bayes Factor... if you want to get a Bayes factor for your model. By default, the full model will be compared to the null model and there are four methods to compute them. I have selected ◉ JZS method (Jeffreys, 1961; Zellner & Siow, 1980). Click Priors... to inspect the prior and posterior distributions for each predictor. Drag all predictors to the box labelled *Plot covariate(s)* (or click ➡) and select ☑ Bayesian predicted distribution. In the main dialog box click OK to fit the model.

Output 9.15 (left) shows the Bayes factor for the full model compared to the null model, which I assume is the model including only the intercept. The right side of the output shows the parameter estimates based on Bayesian estimation. The Bayes factor is 1.066×10^{43} (that's what the E+43 means). In other words it is massive. In short, the probability of the data given the model including all three predictors is 1.07×10^{43} greater than the probability of the data given the model with only the intercept. We should shift our belief in the model (relative to the null model) by a factor of 1.07×10^{43}! This is very strong evidence for the model.

The Bayesian estimate of *b* can be found in the columns labelled *Posterior Mode* and *Posterior Mean*. In fact the columns contain identical values, but they won't always. The reason for the two columns is that we use the peak of the posterior distribution as our estimate and that peak can be defined by either the mode of the posterior or its mean. The values are 0.085 for advertising budget, 3.367 for airplay and 11.086 for image, compared to the values of 0.085, 3.37 and 11.09 (Output 9.8). from the non-Bayesian model. They are basically the same, which is not all that surprising because we started off with very diffuse priors (and so these priors will have had very little influence over the estimates – think back to Section 3.8). We can see this fact in Output 9.16, which shows the prior distribution for the *b* for advertising budget as a red line (in your output you will see similar plots for the other two predictors): the line is completely flat, representing a completely open and diffuse belief about the model parameters. The green line is the posterior distribution, which is quantified in Output 9.15 (right).

Output 9.15



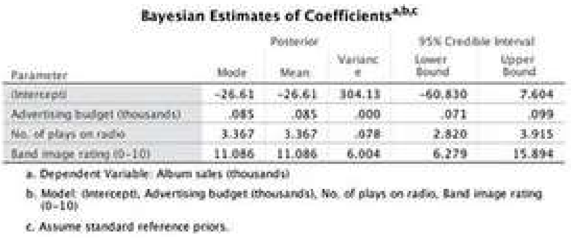
Bayes Factor Model Summary[a,b]

Bayes Factor[c]	R	R Square	Adjusted R Square	Std. Error of the Estimate
1.066E+43	.815	.665	.660	47.09

a. Method: JZS

b. Model: (Intercept), Advertising budget (thousands), No. of plays on radio, Band image rating (0-10)

c. Bayes factor: Testing model versus null model (Intercept).



Bayesian Estimates of Coefficients[a,b,c]

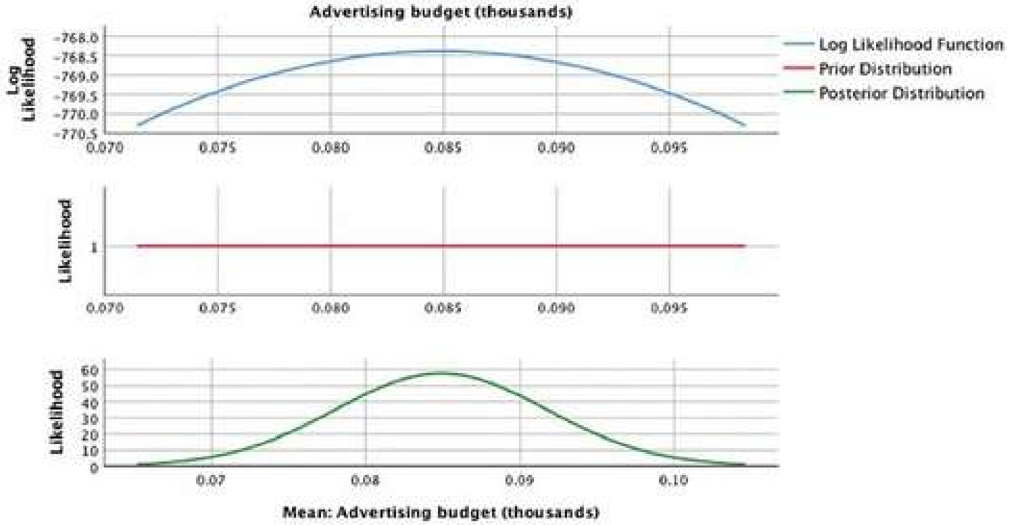
Parameter	Posterior			95% Credible Interval	
	Mode	Mean	Variance	Lower Bound	Upper Bound
(Intercept)	-26.61	-26.61	304.13	-60.830	7.604
Advertising budget (thousands)	.085	.085	.000	.071	.099
No. of plays on radio	3.367	3.367	.078	2.820	3.915
Band image rating (0-10)	11.086	11.086	6.004	6.279	15.894

a. Dependent Variable: Album sales (thousands)

b. Model: (Intercept), Advertising budget (thousands), No. of plays on radio, Band image rating (0-10)

c. Assume standard reference priors.

Output 9.16



Perhaps the most useful parts of Output 9.15 are the 95% credible intervals for the model parameters. Unlike confidence intervals, credible intervals contain the population value with a probability of 0.95 (95%). For advertising budget, therefore, there is a 95% probability that the population value of *b* lies between 0.071 and 0.099, for airplay the population value is plausibly between 2.820 and 3.915, and for image it plausibly lies between 6.279 and 15.894. These intervals are constructed assuming that an effect exists, so you cannot use them to test hypotheses, only to establish plausible population values of the *b*s in the model.

9.14 Reporting linear models

If your model has several predictors than you can't really beat a summary table as a concise way to report your model. As a bare minimum report the betas along with their standard errors and confidence interval (or credible interval if you've gone Bayesian). If you haven't gone Bayesian, report the significance value and perhaps the standardized beta. Include some general fit statistics about the model such as R^2 or the Bayes factor. 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. For the example in this chapter we might produce a table like that in Table 9.2.

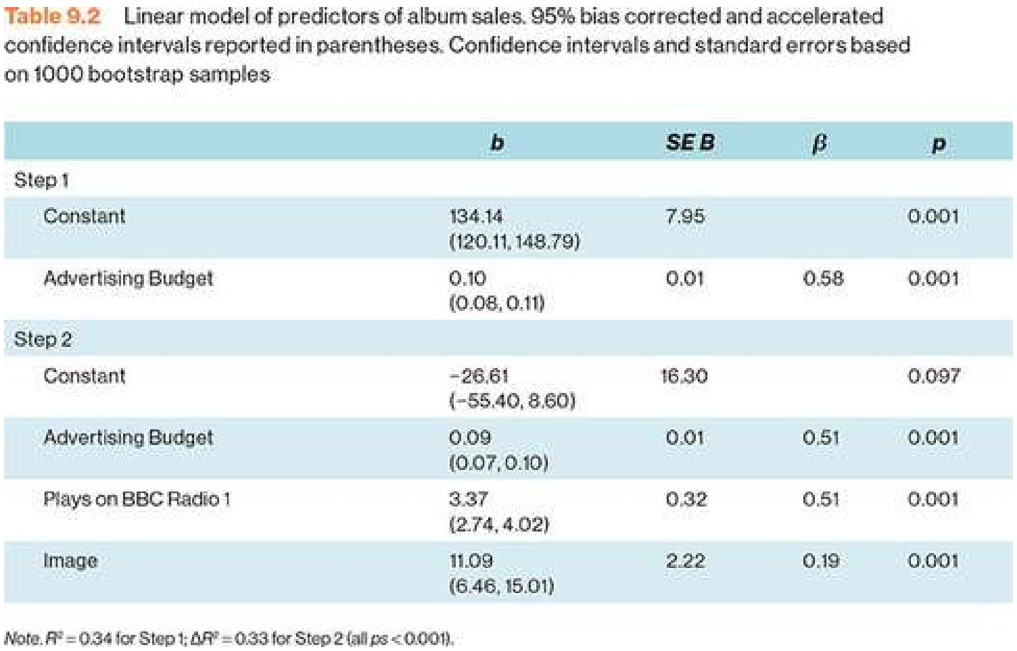


Table 9.2 Linear model of predictors of album sales. 95% bias corrected and accelerated confidence intervals reported in parentheses. Confidence intervals and standard errors based on 1000 bootstrap samples

	b	*SE B*	*β*	*p*
Step 1				
Constant	134.14 (120.11, 148.79)	7.95		0.001
Advertising Budget	0.10 (0.08, 0.11)	0.01	0.58	0.001
Step 2				
Constant	−26.61 (−55.40, 8.60)	16.30		0.097
Advertising Budget	0.09 (0.07, 0.10)	0.01	0.51	0.001
Plays on BBC Radio 1	3.37 (2.74, 4.02)	0.32	0.51	0.001
Image	11.09 (6.46, 15.01)	2.22	0.19	0.001

Note. $R^2 = 0.34$ for Step 1; $\Delta R^2 = 0.33$ for Step 2 (all *ps* < 0.001).

Note. $R^2 = 0.34$ for Step 1; $\Delta R^2 = 0.33$ for Step 2 (all *ps* < 0.001).

Labcoat Leni’s Real Research 9.2 Why do you like your lecturers?

Chamorro-Premuzic, T., *et al.* (2008). *Personality and Individual Differences*, *44*, 965–976.

In the previous chapter we encountered a study by Chamorro-Premuzic *et al.* that linked students’ personality traits with those they want to see in lecturers (see Labcoat Leni’s Real

Research 8.1 for a full description). In that chapter we correlated these scores, but now Labcoat Leni wants you to carry out five multiple regression analyses: the outcome variables across the five models are 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). The data are in the file **Chamorro-Premuzic.sav**.

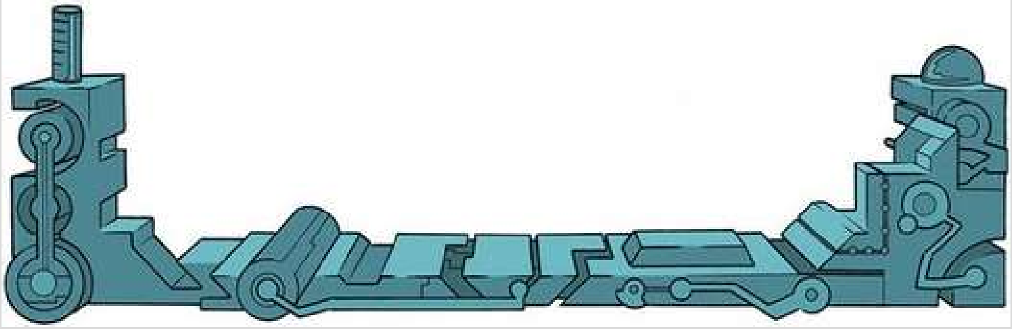
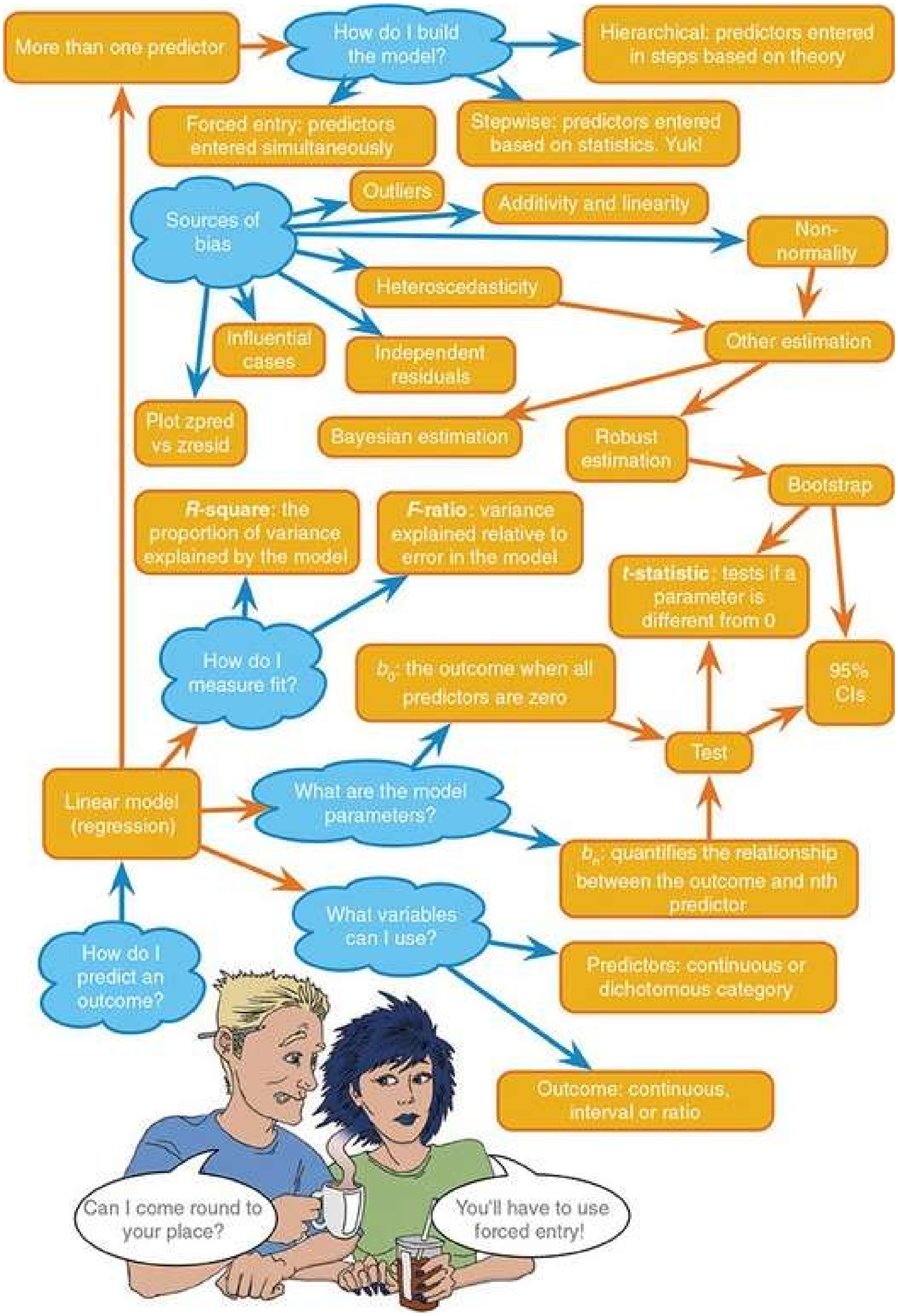


Figure 9.27 What Brian learnt from this chapter


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
Additivity and linearity
Non-normality
Heteroscedasticity
Other estimation
Influential cases
Independent residuals
Plot zpred vs zresid
Bayesian estimation
Robust estimation
Bootstrap
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 from 0
How do I measure fit?
b_0: the outcome when all predictors are zero
95% CIs
Test
Linear model (regression)
What are the model parameters?
b_n: quantifies the relationship between the outcome and nth predictor
What variables can I use?
Predictors: continuous or dichotomous category
How do I predict an outcome?
Outcome: continuous, interval or ratio
Can I come round to your place?
You'll have to use forced entry!

Things to note are as follows: (1) I’ve rounded off to 2 decimal places throughout because this is a reasonable level of precision given the variables measured; (2) if you are following APA format (which I’m not), do not put zeros before the decimal point for the standardized betas, R^2 and *p*-values (because these values shouldn’t exceed 1); (3) I’ve reported exact *p*-values, which is good practice; (4) the R^2 for the initial model and the change in R^2 (denoted by ΔR^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, which is important for readers to know.

9.15 Brian’s attempt to woo Jane

Jane put the fork down next to the jar and suppressed her reflex to gag. When she’d started at this university she’d had utter conviction in testing her flatworm theory. She would be her own single-case design. She knew that experimenting on herself would confound everything, but she wanted some evidence to firm up her beliefs. If it didn’t work on her then she could move on, but if she found evidence for some effect then that was starting point for better research. She felt conflicted, though. Was it the experiments making her mind so unfocused, or was it the interest from campus guy? She hadn’t come here looking for a relationship, she hadn’t expected it, and it wasn’t in the plan. Usually she was so good at ignoring other people, but his kindness was slowly corroding her shell. As she got up and replaced the jar on the shelf she told herself that the nonsense with campus guy had to stop. She needed to draw a line.

9.16 What next?

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’ll ever need to know about statistics. Seriously – you’ll discover in the coming chapters that everything else we discuss is 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 a linear model to tell me which variables are important in predicting talent competition success. Unfortunately I didn't have one, but I did have my dad (and he's better than a linear model). 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 fake-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 mirrors. The only things needed to complete the job were a platinum-selling album and a heroin addiction. However, before I could get those my parents and teachers were about to impress reality upon my young mind …

9.17 Key terms that I've discovered

Adjusted predicted value
Adjusted R^2
Akaike information criterion (AIC)
Autocorrelation
b_i
β_i
Cook's distance
Covariance ratio (CVR)
Cross-validation
Deleted residual
DFBeta
DFFit
Durbin–Watson test
F-statistic
Generalization
Goodness of fit
Hat values
Heteroscedasticity
Hierarchical regression
Homoscedasticity
Independent errors
Leverage
Mahalanobis distance
Mean squares

Model sum of squares
Multicollinearity
Multiple regression
Ordinary least squares (OLS)
Outcome variable
Perfect collinearity
Predicted Value
Predictor variable
Residual
Residual sum 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)

Smart Alex’s tasks

- **Task 1**: In Chapter 4 (Task 7) we looked at data based on findings that the number of cups of tea drunk was related to cognitive functioning (Feng et al., 2010). Using a linear 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? (see Chapter 8, Task 9) (**Tea Makes You Brainy 716.sav**) 🟩⬛⬛⬛

- **Task 2**: Estimate a linear model for the **pubs.sav** data in Jane Superbrain Box 9.1 predicting **mortality** from the number of **pubs**. Try repeating the analysis but bootstrapping the confidence intervals. 🟩🟩⬛⬛
- **Task 3**: In Jane Superbrain Box 2.1 we encountered data (**HonestyLab.sav**) relating to people's ratings of dishonest acts and the likeableness of the perpetrator. Run a linear model with bootstrapping to predict ratings of dishonesty from the likeableness of the perpetrator. 🟩🟩⬛⬛
- **Task 4**: A fashion student was interested in factors that predicted the salaries of catwalk models. She collected data from 231 models (**Supermodel.sav**). For each model she asked them their salary per day (**salary**), their age (**age**), their length of experience as a model (**years**), and their industry status as a model as their percentile position rated by a panel of experts (**beauty**). Use a linear model to see which variables predict a model's salary. How valid is the model? 🟩🟩⬛⬛
- **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. Analyse them with a linear model (**Child Aggression.sav**). 🟩🟩⬛⬛
- **Task 6**: Repeat the analysis in Labcoat Leni's Real Research 9.1 using bootstrapping for the confidence intervals. What are the confidence intervals for the regression parameters? 🟩🟩🟩⬛
- **Task 7**: Coldwell, Pike, & 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 measured dimensions of the child's perceived relationship with their mum: (1) warmth/enjoyment (**child_warmth**), and (2) anger/hostility (**child_anger**). Higher scores indicate more warmth/enjoyment and anger/hostility respectively. They measured the mum's perceived

relationship with her child, resulting in dimensions of positivity (**mum_pos)** and negativity (**mum_neg**). Household chaos (**chaos**) was assessed. The outcome variable was the child’s adjustment (**sdq**): the higher the score, the more problem behaviour the child was reported to display. Conduct a hierarchical linear model in three steps: (1) enter child age and gender; (2) add the variables measuring parent–child positivity, parent–child negativity, parent–child warmth, parent–child anger; (3) add chaos. Is household chaos predictive of children’s problem behaviour over and above parenting? (**Coldwell *et al.* (2006).sav**). 🟩⬛⬛⬛

Answers & additional resources are available on the book’s website at **https://edge.sagepub.com/field5e**